

**An Assessment of the Use of Routinely  
Recorded Data in the UK in a Randomised  
Controlled Trial**

**Graham Arnold Powell**

Thesis submitted in accordance with the requirements of the  
University of Liverpool for the degree of Doctor in Philosophy

April 2018



## **Abstract**

**Introduction and Objectives** Routinely recorded clinical data held in administrative healthcare databases have demonstrated utility in Randomised Controlled Trials (RCTs), such as informing recruitment feasibility assessments, assisting with recruitment and measuring certain study outcomes. Furthermore, routinely recorded non-clinical data, such as data regarding employment and taxation have the potential to inform the measurement of outcomes including health economic analyses. However, limitations with accuracy, access and feasibility have been identified.

This research systematically reviewed the use of routinely recorded data in the UK in RCTs and agreement of routinely recorded data compared to data collected using standard prospective research methods. Subsequently, the accessibility, quality, agreement and feasibility of using routinely recorded data compared to data collected using standard prospective methods in a UK RCT assessing antiepileptic drug treatments for individuals newly diagnosed with epilepsy were assessed.

**Methods** A systematic review was undertaken to assess the use of routinely recorded data in the UK in RCTs and to compare agreement between routinely recorded data and data collected using standard prospective methods. The quality of routinely recorded data was assessed according to a number of criteria, such as the degree of missing data. The agreement of routinely recorded data compared to data collected using standard methods in a UK RCT included calculation of Cohen's Kappa and construction of Bland Altman Plots for categorical and continuous variables, respectively. Assessment of the resources required for protocol development, application for data, accessibility, data preparation and attributes including data coverage and agreement were included in the assessment of feasibility.

**Results** The reviews identified few published studies using routinely recorded clinical data in the UK in RCTs. Mortality data were most commonly used and demonstrated the greatest agreement compared to data collected using standard methods. There was no evidence for the use of routinely recorded data from non-clinical sources, such as governmental employment and taxation data in clinical RCTs.

There was a significant degree of missing routinely recorded data identified for variables and outcome measures relevant to the case study RCT, compared to data collected using standard methods. Where paired data were available, agreement was not satisfactory for the majority of comparisons, including the assessments of seizure occurrence and adverse events. Agreement was satisfactory for two comparisons; date of prescription of antiepileptic drugs and healthcare resource use, assessed using the dates of RCT follow-up assessments. The poor accessibility, prolonged period of application and cost for data access together with the poor data quality and agreement resulted in the limited feasibility for use of routinely recorded data in this RCT assessing antiepileptic drug treatments for epilepsy.

**Conclusions** There is currently very limited experience of using routinely recorded data in UK RCTs for outcomes other than mortality and healthcare resource use for economic evaluations. In this research the accessibility and feasibility of use were limited and degree of missing data and agreement compared to data collected using standard methods unsatisfactory. The results of this research suggest routinely recorded data in the context of prospective clinical research could be an important source of additional data, for example to identify additional events such as seizures not recorded using standard methods. The results suggest that use of routinely recorded data as the primary data source or as a means of validating data collected using standard methods, would be limited. Recommendations include suggestions for improving the access to routinely recorded data for research, development of an integrated electronic health record for use in both clinical practice and research and further assessment of the attributes and 'optimal mix' of routinely recorded data compared to data collected using standard methods.

## Acknowledgements

I would like to take the opportunity to thank my supervisors, who have provided the guidance, support and encouragement to make completion of this thesis possible.

Professor Tony Marson has provided unwavering support, guidance and mentoring, both during this research and in the years leading up to it. For his invaluable academic, clinical and career support I am indebted. Dr Laura Bonnett has been extremely approachable and helpful, always available to answer the numerous technical queries and I am extremely grateful for the assistance and education she has provided. Professors Catrin Tudur-Smith, Paula Williamson and Dyfrig Hughes have provided direction and guidance throughout the research and finally Emily Holmes and Catrin Plumpton have always been available and helpful in responding to my numerous queries.

I would like to acknowledge the Medical Research Council Hubs for Trials Methodology Research (MRC HTMR) who have provided the funding and resources to complete this research, in addition to the enjoyable opportunities for student learning, networking and holistic support. Furthermore, NHS Digital and The Secure Anonymised Information Linkage Databank (SAIL) must also be acknowledged, without their collaboration this work could not have been completed. I must extend my thanks to the helpful individuals in the NHS Digital and SAIL Data Application and Information Governance Teams.

I would like to thank the patients who have taken the time to read the study documentation and provided consent to be included in this research.

I would like to thank my fellow PhD students Kirsty, Nish, Leanne and Rasheed, who have provided holistic support, advice, coffee and have made the process a thoroughly enjoyable experience.

My personal thanks extend to my parents and grandparents for their inspiration and support.

Finally, I thank wholeheartedly my wife Masi, for her understanding, compassion, encouragement, patience and together with Olanna, Mabel and Theodore, helping me keep everything in perspective.

*"I've done the work!"*

## **Authors Declaration**

This thesis is the result of my own work and the material contained therein has not been presented either wholly, or in part, for any other degree or qualification.

The work for this thesis was carried out at the Department of Molecular and Clinical Pharmacology, University of Liverpool, UK and The Walton Centre for Neurology and Neurosurgery, Liverpool, UK.

# Table of Contents

## Chapter One

### Introduction: Epilepsy, Research and Routinely Recorded Data

<b>1.1</b>	<b>Research Overview</b>	<b>1</b>
<b>1.2</b>	<b>An Introduction to Epilepsy</b>	<b>2</b>
1.2.1	Definitions	2
1.2.2	Classification	3
1.2.2.1	Seizures	3
1.2.2.2	Epilepsy and Epilepsy Syndromes	4
1.2.3	Pathogenesis and Epidemiology of Epilepsy	5
1.2.4	Investigations	6
1.2.5	Impact of Epilepsy	6
1.2.6	Treatment	7
1.2.6.1	Prognosis	7
1.2.6.2	Antiepileptic Drugs	7
1.2.6.3	Adverse Events	10
1.2.6.4	Failure of First-Line Treatment	10
1.2.6.4.1	Non-Pharmacological Treatment Approaches	11
<b>1.3</b>	<b>Randomised Controlled Trials in Epilepsy</b>	<b>12</b>
1.2.3.1	The Standard and New Antiepileptic Drugs II (SANAD II) Trial	14
<b>1.4</b>	<b>Routinely Recorded Data</b>	<b>17</b>
1.4.1	Introduction	17
1.4.2	Routinely Recorded Clinical Data	18
1.4.2.1	Routinely Recorded Clinical Data and Clinical Research	19
1.4.2.1.1	Limitations	21
1.4.2.2	Routinely Recorded Clinical Data Sources in the UK	23
1.4.3	Routinely Recorded Non-Clinical Data Sources	27
1.4.3.1	Routinely Recorded Non-Clinical Data Sources in the UK and Clinical Research	27
<b>1.5</b>	<b>Conclusions and Research Objectives</b>	<b>30</b>

## **Chapter Two**

### **The Use of Routinely Recorded Data in the UK to Assess Outcomes in Randomised Controlled Trials: A Review**

<b>2.1</b>	<b>Introduction</b>	<b>33</b>
<b>2.2</b>	<b>Objective</b>	<b>34</b>
<b>2.3</b>	<b>Methods</b>	<b>34</b>
2.3.1	Electronic Database Review	34
2.3.1.1	Registration	34
2.3.1.2	Inclusion Criteria	35
2.3.1.3	Search Strategy	37
2.3.1.4	Study Identification	37
2.3.1.5	Data Extraction	37
2.3.1.6	Data Analysis	38
2.3.2	Narrative Review	39
<b>2.4</b>	<b>Results</b>	<b>40</b>
2.4.1	Electronic Database Review	40
2.4.2	Narrative Review	45
2.4.2.1	NHS Digital	45
2.4.2.2	The Clinical Practice Research Datalink (CPRD)	45
2.4.2.3	ResearchOne	46
2.4.2.4	QResearch	46
2.4.2.5	The Health Improvement Network (THIN) Database	46
2.4.2.6	North West eHealth (NWEH)	46
2.4.2.7	HM Revenue and Customs (HMRC)	46
2.4.2.8	The Secure Anonymised Information Linkage (SAIL) Databank	46
2.4.2.9	The Administrative Data Research Network (ADRN)	47
2.4.2.10	Sources Lacking Data Release Registers	47
<b>2.5</b>	<b>Discussion</b>	<b>48</b>
<b>2.6</b>	<b>Limitations</b>	<b>51</b>
<b>2.7</b>	<b>Conclusions</b>	<b>52</b>

## **Chapter Three**

### **The Agreement of Routinely Recorded Data with Data Collected Using Standard Prospective Methods in UK Studies: A Systematic Review**

<b>3.1</b>	<b>Introduction</b>	<b>53</b>
<b>3.2</b>	<b>Objective</b>	<b>54</b>
<b>3.3</b>	<b>Methods</b>	<b>54</b>
3.3.1	Registration	54
3.3.2	Inclusion Criteria	54
3.3.2.1	Study Designs, Participants, Interventions and Outcome Measures	54
3.3.2.2	UK Routine Data Sources	54
3.3.3	Standard Prospective Methods	55
3.3.4	Assessment of Agreement	55
3.3.5	Search Strategy	55
3.3.5.1	Development of the Electronic Database Search Strategies	56
3.3.5.2	Identifying 'Health Economic' Studies	59
3.3.5.3	Manual Searches	59
3.3.6	Study Identification	60
3.3.7	Data Extraction	60
3.3.8	Data Analysis	61
<b>3.4</b>	<b>Results</b>	<b>62</b>
<b>3.5</b>	<b>Discussion</b>	<b>71</b>
<b>3.6</b>	<b>Limitations</b>	<b>74</b>
<b>3.7</b>	<b>Conclusions</b>	<b>75</b>

## **Chapter Four**

### **The Identification and Accessibility of UK Routinely Recorded Data Sources**

<b>4.1</b>	<b>Introduction</b>	<b>77</b>
<b>4.2</b>	<b>Routinely Recorded Data in the UK Relevant to SANAD II</b>	<b>77</b>
4.2.1	Introduction	77
4.2.2	Secondary Care Clinical Routine Data Sources	78
4.2.2.1	NHS Digital	78
4.2.2.2	NHS Wales Informatics Service	79
4.2.2.3	NHS National Services Scotland: Information Services Division (ISD)	80
4.2.3	Primary Care Clinical Routine Data Sources	81
4.2.3.1	The Clinical Practice Research Datalink (CPRD)	81
4.2.3.2	ResearchOne	82
4.2.3.3	QResearch	82
4.2.3.4	The Health Improvement Network (THIN) Database	83
4.2.3.5	North West eHealth (NWEH)	84
4.2.4	'Linked' Routine Data Sources	85
4.2.4.1	The Secure Anonymised Information Linkage (SAIL) Databank	85
4.2.4.2	The Administrative Data Research Network (ADRN)	86
4.2.5	Non-Clinical Routine Data Sources	87
4.2.5.1	The Office for National Statistics (ONS)	87
4.2.5.2	HM Revenue and Customs (HMRC)	88
4.2.5.3	The Department for Work and Pensions (DWP)	89
4.2.5.4	The Driver and Vehicle Licensing Authority (DVLA)	90
<b>4.3</b>	<b>The Accessibility of Routinely Recorded Data Relevant to SANAD II</b>	<b>91</b>
4.3.1	Accessibility	91
4.3.2	Accessible Routinely Recorded Data Sources	92
4.3.2.1	NHS Digital	92
4.3.2.2	The Secure Anonymised Information Linkage Databank	92
4.3.2.3	NHS National Services Scotland: Information Services Division	92
4.3.2.4	The Office for National Statistics	92
4.3.2.5	North West eHealth	92
4.3.3	Non-Accessible Routinely Recorded Data Sources	93
4.3.3.1	Primary Care Clinical Routine Data Sources	93
4.3.3.2	Non-Clinical Routine Data Sources	93
<b>4.4</b>	<b>Conclusions</b>	<b>94</b>

## **Chapter Five**

### **Methods: The Attributes of Routinely Recorded Data**

<b>5.1</b>	<b>Introduction</b>	<b>95</b>
<b>5.2</b>	<b>Research Objectives</b>	<b>95</b>
<b>5.3</b>	<b>Study Design</b>	<b>96</b>
<b>5.4</b>	<b>Participants</b>	<b>96</b>
<b>5.5</b>	<b>Routinely Recorded Data Sources</b>	<b>97</b>
<b>5.6</b>	<b>Ethical and Regulatory Approvals</b>	<b>97</b>
<b>5.7</b>	<b>Participants Recruitment</b>	<b>98</b>
<b>5.8</b>	<b>Routinely Recorded Data Source Applications</b>	<b>99</b>
<b>5.9</b>	<b>Data Management</b>	<b>101</b>
<b>5.10</b>	<b>Routinely Recorded Data and Data Coding Systems</b>	<b>102</b>
5.10.1	NHS Digital: Hospital Episode Statistics (HES)	102
5.10.2	NHS Wales Informatics Service (NWIS): The Secure Anonymised Information Linkage Databank (SAIL)	103
5.10.3	Primary Care Data	103
5.10.4	Clinical Coding	104
<b>5.11</b>	<b>The Routinely Recorded Data</b>	<b>105</b>
5.11.1	NHS Digital: Hospital Episode Statistics (HES)	107
5.11.2	The Secure Anonymised Information Linkage Databank (SAIL)	113
5.11.3	Primary Care Data, North West England	118
5.11.4	Study Participants	119
5.11.4.1	Gender	119
5.11.4.2	Age	120
<b>5.12</b>	<b>Data Processing and Analysis</b>	<b>121</b>
5.12.1	Overview	121
5.12.2	Data Preparation	121
5.12.3	Assessment of Routinely Recorded Data Quality	122
5.12.4	Assessment of Agreement	122
5.12.5	Assessment of Feasibility and Efficiency	124
<b>5.13</b>	<b>The Data Variables and Outcome Measures</b>	<b>125</b>
5.13.1	The SANAD II Data	125
5.13.1.1	The Identification of Eligible Individuals and Recruitment into SANAD II	125
5.13.1.2	The Follow-Up of Participants and Measurement of SANAD II Outcomes	126
5.13.2	The Routinely Recorded Data	129

5.13.3	The Identification of Seizure Occurrence in the Routinely Recorded Datasets	129
5.13.4	Data Variables Relevant to the Identification of Eligible Individuals and Recruitment into SANAD II	139
5.13.4.1	Seizure Occurrence: Baseline Variables	139
5.13.4.2	The Diagnosis and Classification of Epilepsy and Seizures in the Routinely Recorded Datasets	140
5.13.4.3	Diagnosis and Classification of Epilepsy and Seizures: Variables	143
5.13.4.4	The Assessment of Clinical Investigations in the Routinely Recorded Datasets	145
5.13.5	Data Variables and Outcomes Relevant to the Follow-Up of Participants in SANAD II	149
5.13.5.1	Seizure Occurrence: Follow-Up Variables	149
5.13.5.2	The Assessment of Antiepileptic Drugs in the Routinely Recorded Datasets	152
5.13.5.3	The Assessment of Adverse Events in the Routinely Recorded Datasets	154
5.13.5.4	The Assessment of Healthcare Resource Use in the Routinely Recorded Datasets	159
<b>5.14</b>	<b>Conclusions</b>	<b>162</b>

## **Chapter Six**

### **Results: The Assessment of Seizures and Data Variables Relevant to the Identification of Eligible Individuals and Recruitment into SANAD II**

<b>6.1</b>	<b>Introduction</b>	<b>163</b>
<b>6.2</b>	<b>The Identification of Seizure Occurrence in the Routinely Recorded Datasets</b>	<b>164</b>
<b>6.3</b>	<b>Date of First Seizure</b>	<b>166</b>
<b>6.4</b>	<b>Date of First Tonic-Clonic Seizure</b>	<b>170</b>
<b>6.5</b>	<b>Conclusions: Date of First Seizure and First Tonic-Clonic Seizure</b>	<b>174</b>
<b>6.6</b>	<b>The Identification of Diagnosis and Classification of Epilepsy and Seizures in the Routinely Recorded Datasets</b>	<b>176</b>
<b>6.7</b>	<b>Diagnosis of Epilepsy (Baseline)</b>	<b>178</b>
<b>6.8</b>	<b>Diagnosis of Epilepsy (All-Time)</b>	<b>182</b>
<b>6.9</b>	<b>Classification of Seizures (Baseline)</b>	<b>186</b>
<b>6.10</b>	<b>Classification of Seizures (All-Time)</b>	<b>188</b>
<b>6.11</b>	<b>Conclusions: Diagnosis and Classification of Epilepsy and Seizures</b>	<b>190</b>
<b>6.12</b>	<b>The Identification of Clinical Investigations in the Routinely Recorded Datasets</b>	<b>191</b>
<b>6.13</b>	<b>Magnetic Resonance Imaging (MRI)</b>	<b>192</b>
<b>6.14</b>	<b>Computed Tomography (CT)</b>	<b>193</b>
<b>6.15</b>	<b>Electroencephalogram (EEG)</b>	<b>194</b>
<b>6.16</b>	<b>Conclusions: Clinical Investigations</b>	<b>196</b>
<b>6.17</b>	<b>Conclusions</b>	<b>197</b>

## Chapter Seven

### Results: Data Variables and Outcomes Relevant to the Follow-Up in SANAD II, Healthcare

#### Resource Use and Primary Care Data

<b>7.1</b>	<b>Introduction</b>	<b>199</b>
<b>7.2</b>	<b>Date of First Follow-Up Seizure</b>	<b>200</b>
<b>7.3</b>	<b>Time to First Follow-Up Seizure</b>	<b>204</b>
<b>7.4</b>	<b>Date of First Follow-Up Tonic-Clonic Seizure</b>	<b>206</b>
<b>7.5</b>	<b>Conclusions: Date of First Follow-Up and First Tonic-Clonic Follow-Up Seizures</b>	<b>210</b>
<b>7.6</b>	<b>Date 12 Month Remission Achieved</b>	<b>212</b>
<b>7.7</b>	<b>Time to 12 Month Remission</b>	<b>215</b>
<b>7.8</b>	<b>Conclusions: Date and Time 12 Month Remission Achieved</b>	<b>216</b>
<b>7.9</b>	<b>Total Number of Follow-Up Seizures</b>	<b>217</b>
7.9.1	Total Number of Follow-Up Seizures (All Seizure Types)	217
7.9.2	Total Number of Follow-Up Tonic-Clonic Seizures	219
<b>7.10</b>	<b>Conclusions: Total Number of Follow-Up Seizures</b>	<b>221</b>
<b>7.11</b>	<b>The Assessment of Antiepileptic Drug Prescribing in Routine Datasets</b>	<b>222</b>
7.11.1	The Date of AED First Prescription	222
7.11.2	Compliance	224
<b>7.12</b>	<b>Conclusions: The Assessment of Antiepileptic Drugs</b>	<b>225</b>
<b>7.13</b>	<b>The Assessment of Adverse Events in the Routinely Recorded Datasets</b>	<b>226</b>
<b>7.14</b>	<b>Conclusions: The Assessment of Adverse Events</b>	<b>227</b>
<b>7.15</b>	<b>The Assessment of Healthcare Resource Use in Routine Datasets</b>	<b>229</b>
7.15.1	Planned Healthcare Attendances: The SANAD II Baseline Assessments	229
7.15.2	Planned Healthcare Attendances: The SANAD II Follow-Up Assessments	231
7.15.3	Conclusions: Planned Healthcare Attendances	234
7.15.4	Unplanned Healthcare Attendances	235
7.15.4.1	Emergency Department Attendances	235
7.15.4.2	Inpatient Admissions	235
7.15.4.3	Emergency Department Attendances: Agreement	237
7.15.4.4	Inpatient Admissions: Agreement	239
7.15.5	Conclusions: Unplanned Healthcare Attendances	240
<b>7.16</b>	<b>Primary Care Data, North West England</b>	<b>242</b>
<b>7.17</b>	<b>Conclusions</b>	<b>244</b>

## **Chapter Eight**

### **Results: The Feasibility and Efficiency of Accessing and Using Routinely Recorded Data**

<b>8.1</b>	<b>Introduction</b>	<b>247</b>
<b>8.2</b>	<b>Accessing and Using Routinely Recorded Data</b>	<b>247</b>
8.2.1	'Accessible' Routinely Recorded Data Sources	248
8.2.1.1	NHS Digital	248
8.2.1.2	The Secure Anonymised Information Linkage Databank	249
8.2.1.3	North West eHealth	249
8.2.1.4	General Practices', North West England	250
8.2.2	'Non-Accessible' Routinely Recorded Data Sources	250
8.2.2.1	Routinely Recorded Primary Care Clinical Data Sources	250
8.2.2.2	The Driver and Vehicle Licensing Agency	251
8.2.2.3	Department for Work and Pensions and HM Revenue and Customs	251
8.2.2.4	The Administrative Data Research Network	251
<b>8.3</b>	<b>The Feasibility and Efficiency of Accessing and Using Routinely Recorded Data</b>	<b>252</b>
8.3.1	Clinical Routine Data Sources	252
8.3.1.1	Secondary Care Data	252
8.3.1.2	Primary Care Data	253
8.3.2	Non-Clinical Routine Data Sources	255
<b>8.4</b>	<b>Discussion</b>	<b>256</b>
<b>8.5</b>	<b>Conclusions</b>	<b>259</b>
<b>8.6</b>	<b>Recommendations</b>	<b>259</b>
<b>8.7</b>	<b>General Conclusions</b>	<b>261</b>

## **Chapter Nine**

### **Discussion and Conclusions: An Assessment of the Use of Routinely Recorded Data in the UK in a Randomised Controlled Trial**

<b>9.1</b>	<b>Introduction</b>	<b>263</b>
<b>9.2</b>	<b>Discussion</b>	<b>264</b>
9.2.1	The Use of Routinely Recorded Data in the UK to Assess Outcomes in RCTs	264
9.2.2	The Agreement of Routinely Recorded Data with Data Collected Using Standard Prospective Methods in UK Studies	265
9.2.3	The Identification and Accessibility of UK Routinely Recorded Data Sources	266
9.2.4	The Attributes of Routinely Recorded Data Extracted from Electronic Medical Records Compared Against Data Collected Using Standard Prospective Methods in the RCT SANAD II	266
9.2.5	The Feasibility and Efficiency of Accessing and Using Data from Electronic Medical Records in the Randomised Controlled Trial SANAD II	269
9.2.6	Implications for Clinical Research and Practice	269
9.2.6.1	Implications for Clinical Research	270
9.2.6.2	Implications for Clinical Practice	272
<b>9.3</b>	<b>Recommendations and Further Research</b>	<b>273</b>
9.3.1	General Recommendations	273
9.3.2	Routinely Recorded Clinical Data	277
9.3.3	Routinely Recorded Non-Clinical Data	280
<b>9.4</b>	<b>Concluding Remarks</b>	<b>281</b>

<b><u>References</u></b>	<b>283</b>
--------------------------	------------

<b><u>Appendix A</u></b>	<b><i>Chapter Two: Search Strategies and Further Results</i></b>	<b>295</b>
--------------------------	--	------------

<b><u>Appendix B</u></b>	<b><i>Chapter Three: Search Strategies and Further Results</i></b>	<b>309</b>
--------------------------	--	------------

<b><u>Appendix C</u></b>	<b><i>Chapter Five: Information Leaflet and Consent Form</i></b>	<b>347</b>
--------------------------	--	------------

<b><u>Appendix D</u></b>	<b><i>Chapter Eight: Further Results and Publication</i></b>	<b>355</b>
--------------------------	--	------------

## **List of Abbreviations**

AED	Antiepileptic Drug
APC	Admitted Patient Care
ADR	Adverse Drug Reaction
ADRN	Administrative Data Research Network
A&E	Accident and Emergency
AED	Antiepileptic Drug
CC	Critical Care
CI	Confidence Interval
CT	Computed Tomography
CPRD	Clinical Practice Research Datalink
CRCA	Commissioners for Revenue and Customs Act
CRF	Case Report Form
CTRC	Clinical Trials Research Centre
DAAG	Data Access Advisory Group
DARS	Data Access Request Service
DIRUM	Database of Instruments for Resource Use Measurement
DVLA	Driver and Vehicle Licensing Authority
DWP	Department for Work and Pensions
EDDS	Emergency Department Dataset
EEG	Electroencephalogram
GDPR	General Data Protection Regulation
GP	General Practice / General Practitioner
HES	Hospital Episode Statistics
HMRC	Her Majesty's Revenue and Customs
HRA	Health Research Authority

HSCIC	Health and Social Care Information Centre
HTA	Health Technology Assessment
HTMR	Medical Research Council Hubs for Trials Methodology Research
ICD Problems	International Statistical Classification of Diseases and Related Health Problems
ICTMC	International Clinical Trials Methodology Conference
ID	Identification
IGARD	Independent Group Advising on the Release of Data
IGRP	Information Governance Review Panel
ILAE	International League Against Epilepsy
IP	Inpatient
ISAC	Independent Scientific Advisory Committee
ISD	Information Services Division
IT	Information Technology
LSOA	Lower-Layer Super Output Area
MHRA	Medicines and Healthcare Products Regulatory Agency
MRC	Medical Research Council
MRI	Magnetic Resonance Imaging
NGPSE	National General Practice Study of Epilepsy
NHS	National Health Service
NICE	National Institute of Health and Care Excellence
NIHR	National Institute for Health Research
NISCHR	National Institute of Social Care and Health Research
NWIS	NHS Wales Informatics Service
ONS	Office for National Statistics
OP	Outpatient
OPCS	Office of Population, Census and Surveys
PAS	Patient Administration Service

PEDW	Patient Episode Database for Wales
PPV	Positive Predictive Value
QOL	Quality of Life
RCT	Randomised Controlled Trial
REC	Research Ethics Committee
RR	Relative Risk
THIN	The Health Improvement Network Database
TRUD	Technology Reference Data Update Distribution
TTP	The Phoenix Partnership
UK	United Kingdom
UOL	University of Liverpool
SANAD	Standard and New Antiepileptic Drugs
SAIL	The Secure Anonymised Information Linkage Database
SFT	Secure File Transfer
SHIP	Scottish Health Informatics Programme
SIR	Salford Integrated Record
SNOMED CT	Systemised Nomenclature of Medicine – Clinical Terms
SUSAR	Suspected Unexpected Serious Adverse Reaction
SUS	Secondary Uses Service
WDS	Welsh Demographics Service

## **Chapter One**

### **Introduction: Epilepsy, Research and Routinely Recorded**

#### **Data**

##### **1.1 Research Overview**

In Chapter One, epilepsy and the current prospective research methods used to assess treatments for epilepsy will be discussed. The case study Randomised Controlled Trial (RCT), the Standard and New Antiepileptic Drugs II (SANAD II) RCT will be introduced. Routinely recorded data in the UK and the potential for use in clinical research will subsequently be introduced before finally the objectives of this research will be presented.

In Chapter Two the use of routinely recorded data in RCTs in the UK will be reviewed and in Chapter Three the agreement between UK routinely recorded data compared to data collected using standard methods in prospective studies will be assessed in a systematic review.

In Chapter Four, sources of routinely recorded data in the UK relevant to the outcomes of SANAD II will be reviewed and sources where routinely recorded data are accessible for individuals recruited into SANAD II will be identified.

In Chapter Five, the methods for the assessment of the attributes of routinely recorded data retrieved from electronic medical records compared to data collected using standard prospective methods in SANAD II will be presented. The assessment of seizure occurrence, diagnosis and classification of epilepsy in routinely recorded datasets will be assessed in Chapter Six and variables and outcome measures relevant to the follow-up of participants in SANAD II will be assessed in Chapter Seven. The feasibility and efficiency of accessing and using routinely recorded data for participants in SANAD II will be assessed in Chapter Eight.

In Chapter Nine, the results for the objectives of this research will be discussed and conclusions presented. Subsequently, recommendations for improving the use of routinely recorded data in research will be proposed and avenues for further research will be suggested.

## **1.2 An Introduction to Epilepsy**

### ***1.2.1 Definitions***

*Seizures* are the manifestation of paroxysmal, uncontrolled, abnormal electrical discharge of neurons within the cerebral hemispheres [1, 2].

*Epilepsy* is a common, chronic neurological condition with wide reaching medical and psychosocial implications characterised by recurrent and unprovoked seizures [1, 2].

The annual age-adjusted incidence of epilepsy is estimated to be between 40 and 70 per 100 000 persons [3]. The incidence rate is greatest at the extremes of life [3, 4].

The prevalence of epilepsy is 0.5-1%. In addition, 5% of the population, predominantly in the first year of life and those over 75 years of age, will experience a single unprovoked seizure or acute symptomatic seizure [3].

With no predilection for age or sex, epilepsy affects all members of society.

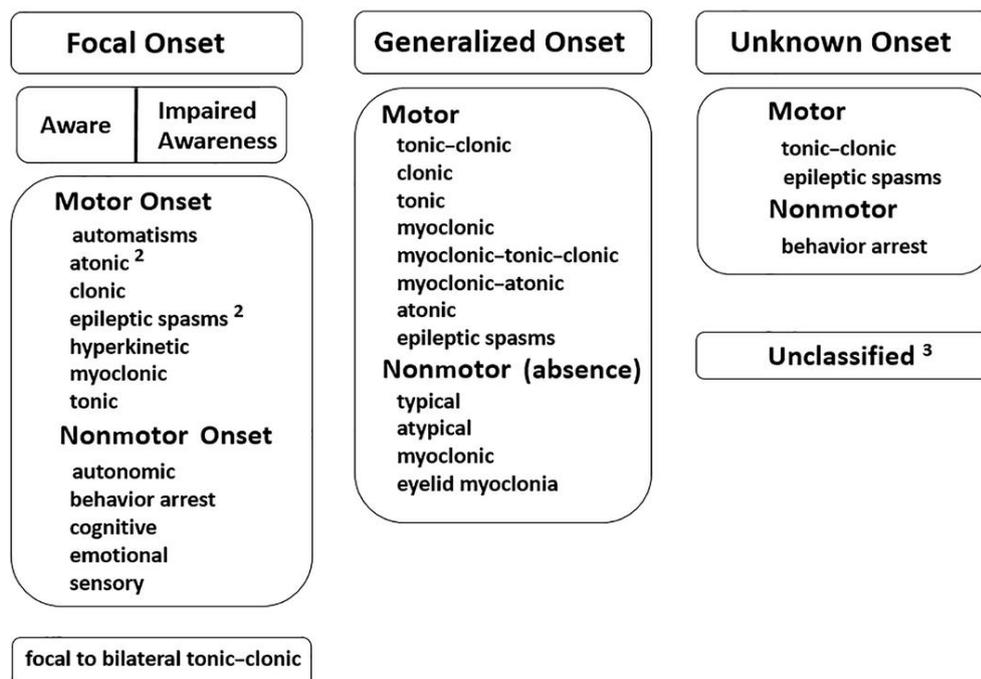
## 1.2.2 Classification

### 1.2.2.1 Seizures

The International League Against Epilepsy (ILAE) Classification (1981) proposed the first widely accepted criteria for classification of seizure type [5]. A multitude of seizure types were included, broadly classified as ‘partial’, originating from a focal, cortical onset in one hemisphere or ‘generalised’, originating in deeper midline structures and propagating to both hemispheres simultaneously. Where uncertainty remained, seizures were deemed ‘unclassified’. The ILAE classification has undergone numerous iterations and the most recent proposal is The ILAE Operational Classification of Seizure Types (2017) [6]. Much of the terminology has been updated, for example ‘partial’ has been replaced with ‘focal onset’ seizures, but the broad classification of seizure types is consistent. *Figure 1.1* summarises the most recent classification of seizure types.

*Figure 1.1: The ILAE Operational Classification of Seizure Types (2017) [6]*

### ILAE 2017 Classification of Seizure Types Expanded Version <sup>1</sup>

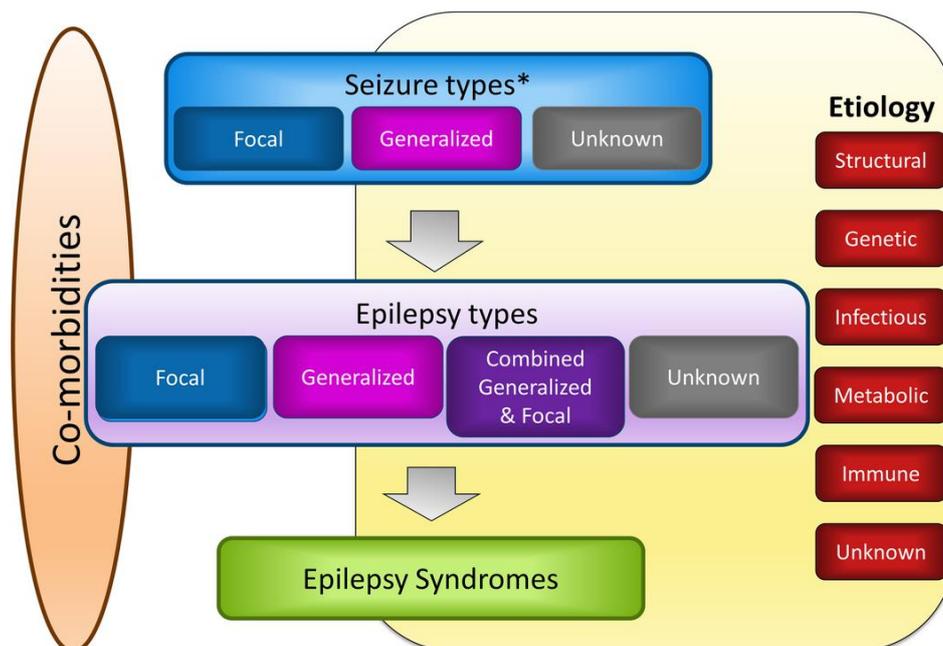


### 1.2.2.2 Epilepsy and Epilepsy Syndromes

Following the diagnosis of seizures, a diagnosis of epilepsy can be considered. The ILAE Practical Clinical Definition of Epilepsy (2014) [7] proposes the most recent criteria for diagnosis and updates the guidance published in 2005 [8]. The guidance proposes that at least two unprovoked seizures occurring greater than 24 hours apart are required for a diagnosis of epilepsy. Additionally, individuals experiencing a single seizure with a probability of subsequent seizures similar to the general recurrence risk following two unprovoked seizures may also qualify for a diagnosis. However, calculating such probabilities is troublesome in both clinical practice and the research environment.

The ILAE Revised Classification of the Epilepsies first included the classification of epilepsies and epilepsy syndromes in addition to seizures in 1989 [9]. The ILAE Classification of the Epilepsies (2017) is the most recently updated version [10] and the present three levels. Firstly, seizure type according to the 2017 ILAE Seizure Classification [6] is diagnosed. Subsequently epilepsy type is diagnosed and may include focal epilepsy, generalized epilepsy, combined generalized and focal epilepsy or unknown epilepsy. Finally, epilepsy syndrome is considered. The process is summarised in *Figure 1.2*. This recent classification considers aetiology at each stage as specific aetiology may influence specific treatment, previously discussed in *Section 1.2.2*.

*Figure 1.2: The ILAE Classification of the Epilepsies (2017) [10]*



### ***1.2.3 Pathogenesis and Epidemiology of Epilepsy***

The occurrence of seizures has been described throughout history but the recognition of uncontrolled discharge from cerebral neurones as the cause was not proposed until the 19th century. The electroencephalogram (EEG) was developed to assess this abnormal neuronal activity and it was rapidly appreciated that specific seizure types were characterised by specific EEG characteristics [1].

Despite recent advances in understanding the pathogenesis of seizures, in a large proportion of individuals with a diagnosis of epilepsy the aetiology cannot be defined, the majority of which experience generalised seizures. Aetiology in individuals with focal epilepsies can more readily be identified [2]. Of the individuals with newly diagnosed focal epilepsy enrolled in the National General Practice Study of Epilepsy (NGPSE) 32% has a clear aetiology and 9% a probable aetiology [11]. Cerebrovascular disease was the most common precipitant [11] with this aetiology also occurring in 11% of patients with adult onset epilepsy in the Rochester epidemiology studies [3]. Additional aetiologies included intracranial tumours, infection and acute or remote trauma [11]. More recently, understanding of the genetic architecture of specific epilepsies has resulted in a number of specific causes being identified. For example, Dravet's syndrome is associated with a mutation in the type 1 voltage gated sodium channel-encoding SCN1A gene and an abnormal glucose transport protein type 1 (GLUT1) and can result in seizures and developmental delay in infancy [12]. However, a complex polygenic inheritance is observed in the majority of idiopathic and cryptogenic generalised epilepsies, as demonstrated by the high concordance in monozygotic twins with generalised epilepsies, reported at 65% in some studies [13].

#### **1.2.4 Investigations**

Epilepsy is a clinical diagnosis and investigations are not mandated. However, certain investigations are frequently performed during the assessment of an individual with newly diagnosed epilepsy. Such investigations may inform the classification of seizures and epilepsy, assess for an underlying cause and resultantly, inform the therapeutic approach.

Magnetic Resonance Imaging (MRI) provides information on the anatomy of the brain and can identify structural abnormalities, such as space-occupying lesions or cortical abnormalities that may require or be amenable to surgery. Computed Tomography (CT) may also be useful in selected cases to exclude space-occupying lesions where there is a low pre-test probability or in cases where MRI is contraindicated.

Electroencephalography (EEG) provides information regarding the neuronal activity during the ictal (seizure) and inter-ictal (non-seizure) phases and can inform the diagnosis and classification as well as informing the pre-surgical evaluation, in appropriate patients.

#### **1.2.5 Impact of Epilepsy**

The World Health Organisation defines Quality of Life (QOL) as "an individual's perception of their position in life, in the context of the culture and value systems in which they live, and in relation to their goals, expectations, standards, and concerns. It is a broad ranging concept, effected in a complex way by the person's physical health, psychological state, level of independence, social relationships and their relationships to salient features of their environment" [14].

An issue of primary importance in the management of patients with epilepsy is the maximisation of QOL and this is the ultimate aim of treatment. Broadly, there are two major dimensions that impact upon patients QOL; clinical and psychosocial variables.

Clinical variables including adverse events of antiepileptic drugs and seizure frequency and severity have a strong negative association with subjective health status and perception of QOL of patients with epilepsy [15]. In the management of epilepsy, clinical issues need to be addressed to ensure QOL is maximised for the individual [16].

Psychosocial variables including psychological distress, adjustment and coping, loneliness and stigma perception contribute significantly to the variance in patients self-reported quality of life judgements [17]. Finally, psychological co-morbidity including anxiety and depression occur at a higher incidence in patients with epilepsy compared to the general population and are associated with a reduced QOL. Therefore, addressing both anxiety and depression can exert a positive influence on QOL [16]. Finally, a diagnosis of epilepsy may have an impact on an individuals' eligibility for specific roles in society, including roles in employment. For example, employment as a lifeguard would have to cease, although this may only be temporary. Furthermore, following a diagnosis of seizure or epilepsy in the UK, there are legal restrictions on driving licensing. Following an initial seizure, driving must cease and the Driver and Vehicle Licensing Agency (DVLA) must be informed. This applies for holders of both Group 1 (car and motorcycle drivers) and Group 2 (heavy goods vehicle drivers) licenses. The duration of suspension will depend on a number of factors including specific diagnosis, frequency of seizures, timing of seizures (daytime, nocturnal), treatment and type of license (Group 1, 2). In the majority of circumstances for Group 1 licenses, 12 months are required free from seizures before a license can be re-issued [18].

## **1.2.6 Treatment**

### *1.2.6.1 Prognosis*

The NGPSE identified that the prognosis of epilepsy is favourable, with 65-85% of patients entering long-term remission, more likely in patients with newly diagnosed epilepsy and receiving early treatment. Furthermore, good response to initial treatment and a prolonged period of remission from seizures are associated with an improved prognosis [19].

Mortality is highest in the initial period following diagnosis, often related to the underlying aetiology, although Sudden Unexpected Death in Epilepsy Patients (SUDEP) is recognised and raised mortality rate is observed throughout the course of epilepsy [19].

### *1.2.6.2 Antiepileptic Drugs*

Antiepileptic drugs (AEDs) remain the mainstay of treatment for patients with epilepsy. Seventy percent of patients with epilepsy will experience sustained remission from seizures following initiation of AED treatment [20, 21]. The number of AEDs has increased dramatically in recent years and there are now over 20 AEDs licensed and available [22].

Traditionally, carbamazepine and sodium valproate have been the recommended first line AEDs for focal and generalized epilepsy, respectively and monotherapy considered the appropriate initial approach. However, longer term RCTs assessing clinical and cost effectiveness such as the Standard and New Antiepileptic Drugs Trial (SANAD) provide evidence to support lamotrigine as a first line treatment for focal epilepsy, whilst also demonstrating that gabapentin and topiramate are poor first line treatments. Similarly topiramate and lamotrigine were shown to be inferior to valproate as a first line treatment for generalized or unclassified epilepsy [23, 24].

As a result of the number of available AEDs and complex evidence base, treatment guidelines have been published by national bodies such as the National Institute for Health and Care Excellence (NICE) [25] and medical associations such as the International League Against Epilepsy (ILAE) [26]. Guidelines aim to synthesise the available evidence and make appropriate recommendations, but the strength of recommendations is necessarily limited by the quality of evidence available. Due to the limited evidence available rather than recommend a single drug, NICE, in their most recent update of their epilepsy guidelines, recommend a number of AEDs that should be considered 'first-line' in the treatment of a range of seizure types and epilepsy syndromes. AED options by seizure type are summarised in *Box 1.1* [25]. An appropriate AED is selected following discussion between clinician and patient. Perhaps the most important choice being for women of child bearing age with genetic generalized epilepsy, where sodium valproate should be used with caution due to the risks of teratogenicity and greater risks of neurodevelopmental sequelae [27], where likely less effective but safer options such as levetiracetam and lamotrigine may be selected [28].

Box 1.1: NICE Guidance: Antiepileptic Drug Options by Seizure Type [25]

Seizure type	First line	Adjunctive	Others that may be considered on referral to tertiary care	Do not offer (may worsen seizures)
Generalised tonic-clonic	Carbamazepine, lamotrigine, oxcarbazepine*, sodium valproate	Clobazam*, lamotrigine, levetiracetam, sodium valproate, topiramate		If there are absence or myoclonic seizures, or if juvenile myoclonic epilepsy is suspected: carbamazepine, gabapentin, oxcarbazepine, phenytoin, pregabalin, tiagabine, vigabatrin
Tonic or atonic	Sodium valproate	Lamotrigine*	Rufinamide*, topiramate*	Carbamazepine, gabapentin, oxcarbazepine, pregabalin, tiagabine, vigabatrin
Absence	Ethosuximide, lamotrigine*, sodium valproate	Ethosuximide, lamotrigine*, sodium valproate	Clobazam*, clonazepam, levetiracetam*, topiramate*, zonisamide*	Carbamazepine, gabapentin, oxcarbazepine, phenytoin, pregabalin, tiagabine, vigabatrin
Myoclonic	Levetiracetam*, sodium valproate, topiramate*	Levetiracetam, sodium valproate, topiramate*	Clobazam*, clonazepam, piracetam, zonisamide*	Carbamazepine, gabapentin, oxcarbazepine, phenytoin, pregabalin, tiagabine, vigabatrin
Focal	Carbamazepine, lamotrigine, levetiracetam, oxcarbazepine, sodium valproate	Carbamazepine, clobazam*, gabapentin*, lamotrigine, levetiracetam, oxcarbazepine, sodium valproate, topiramate	Eslicarbazepine acetate*, lacosamide, phenobarbital, phenytoin, pregabalin*, tiagabine, vigabatrin, zonisamide*	
Prolonged or repeated seizures and convulsive status epilepticus in the community	Buccal midazolam, rectal diazepam†, intravenous lorazepam			
Convulsive status epilepticus in hospital	Intravenous lorazepam Intravenous diazepam, buccal midazolam	Intravenous phenobarbital Phenytoin		
Refractory convulsive status epilepticus	Intravenous midazolam†, propofol† (not in children), thiopental sodium†			

\*At the time of publication of the main NICE guidance (January 2012), this drug did not have UK marketing authorisation for this indication and/or population. Informed consent should be obtained and documented.

†At the time of publication of the main NICE guidance (January 2012), this drug did not have UK marketing authorisation for this indication and/or population. Informed consent should be obtained and documented in line with normal standards in emergency care.

### *1.2.6.3 Adverse Events*

An 'adverse event' can be defined as the occurrence of an undesirable symptom or event during treatment with an AED that may or may not be caused by the AED [29]. Similar terms such as adverse effect and adverse drug reaction imply a more direct cause and effect relationship. AEDs are associated with four types of adverse events:

- Acute Dose Related Toxicity
- Acute Idiosyncratic Toxicity
- Chronic Toxicity
- Teratogenicity

Adverse events associated with AEDs are troublesome in up to 69% of adult patients with epilepsy [30] and 30% of patients experience treatment failure during initial monotherapy, often due to the occurrence of adverse events [11]. Furthermore, adverse events resulting from AED therapy significantly impact on psychological and social functioning, and the overall patient perceived impact of epilepsy [31].

### *1.2.6.4 Failure of First-Line Treatment*

The first line AED will fail in a significant proportion of patients due to lack of efficacy, intolerability or a combination of both, regardless of AED [32]. Clinical factors including seizure type and EEG result may be associated with treatment failure although notably, choice of AED is not a significant predictor of treatment failure [33]. Following a first treatment failure, the probability of 12 month remission remains high; overall 70% of patients will achieve a 12 month remission, 80% with a first treatment failure due to adverse events and 65% with treatment failure due to inadequate seizure control [34]. Inevitably, the more AEDs that a patient fails due to lack of seizure control, the less likely that subsequent treatment trials will be successful [21]. The ILAE has defined pharmacoresistance as 'failure of adequate trials of two tolerated, appropriately chosen and used anti-epileptic drug schedules whether as monotherapy or in combination to achieve sustained seizure freedom' [35]. A proportion of patients will achieve seizure control following further AED changes [34] and therefore in patients experiencing treatment failure an alternative AED monotherapy or rational AED polytherapy may be considered.

#### *1.2.6.4.1 Non-Pharmacological Treatment Approaches*

Non-pharmacological treatment approaches may be considered in individuals with pharmaco-resistant epilepsy and in some circumstances as first-line treatment.

Neurosurgical approaches include resection of structural lesions such as space-occupying lesions. 'Epilepsy surgery' may be indicated and involves resection of a non-eloquent localised area of the brain, known to be the epileptogenic focus. In such cases thorough pre-surgical evaluation is required including but not limited to imaging, EEG and psychological assessment. Novel surgical approaches may also be employed and include vagal nerve or deep brain stimulation. In paediatric patients such as those with an abnormal glucose transport protein type 1 (GLUT1) the ketogenic diet may be effective.

### **1.3 Randomised Controlled Trials in Epilepsy**

The National Institute for Health and Care Excellence (NICE) defines the Randomised Controlled Trial (RCT) as ‘a study in which a number of similar people are randomly assigned to two (or more) groups to test a specific drug, treatment or other intervention. One group (the experimental group) has the intervention being tested; the other (the comparison or control group) has an alternative intervention, a dummy intervention (placebo) or no intervention. Outcomes are measured at specific intervals and the difference in response between groups assessed statistically’ [36].

The Randomised Controlled Trial (RCT) is a rigorous research technique, minimising allocation bias and remaining the standard for regulatory approval of new treatments in healthcare. Consequently, the majority of RCTs assessing the effectiveness of AEDs in the treatment of epilepsy are undertaken to meet the requirements of regulatory authorities such as the Food and Drug Administration (FDA) in the United States or the European Medicines Agency (EMA). RCTs assessing new AEDs are usually undertaken first in adults with refractory focal epilepsy who have failed multiple previous AEDs and who experience regular seizures. The initial comparison is typically with placebo in order to demonstrate statistically significant superiority for efficacy. The initial regulatory licensing is frequently for use as an add-on treatment for refractory focal epilepsy. RCTs may then be undertaken in other populations including children, patients with refractory generalized seizures and specific epilepsy syndromes and finally as monotherapy treatment for epilepsy. However, such RCTs sponsored by the pharmaceutical industry primarily aim to fulfil regulatory requirements for licensing rather than inform the clinical treatment of epilepsy and have limited external validity. Typically, an eight week baseline is followed by a 16 week study duration, too short to provide evidence about longer term clinical and cost effectiveness that patients and health services require. There are few RCTs comparing different AED add on regimens with the majority of the RCT evidence provided through regulatory placebo controlled add-on RCTs. Furthermore, the outcome reported is commonly a measure of change in seizure frequency; the FDA preferring median reduction in seizure frequency and the EMA preferring the proportion of patients with a 50% or greater reduction in seizure frequency. Neither outcome is particularly meaningful to patients. Placebo controlled studies are often followed by open label extension studies which may provide longer term, but uncontrolled, data about AEDs. Whilst such designs may provide additional safety data, they are uninterpretable from the point of view of efficacy due to selection bias, heterogeneity, and their uncontrolled design [37].

Publically funded, pragmatic RCTs assessing long term clinical and cost effectiveness provide data more informative and directly relevant to routine clinical practice. However, pragmatic RCTs are expensive, time-consuming and resource intensive.

Techniques such as network meta-analysis can usefully summarise the available data. This approach uses data from RCTs (placebo controlled) although it is important to emphasize that any comparison is indirect and not randomized. Network meta-analyses make a number of assumptions, in particular that the population of patients recruited to RCTs is similar and consistent. This assumption is most likely violated in refractory epilepsy as the typical patient recruited to trials in the late 1980s and early 1990s is systematically different to a current typical patient. Current patients are likely more refractory given the range of treatments now available to try before considering joining a trial, and trials are conducted over many more centres world-wide than previously. The escalating placebo response rate is likely due to multiple causes, however may provide evidence for this change in case-mix [38].

It is important to highlight the limitations associated with RCTs. Important safety data will arise from designs other than RCTs such as observational record-linkage studies. As such, coordinated post marketing surveillance is required to minimize risk to patients, particularly important for rare but life threatening events such as felbamate associated liver failure, long term events such as vigabatrin retinopathy [39] and retigabine pigmentation and retinopathy [40], and teratogenic effects [41]. As a result of selection bias, RCTs may lack external validity. Significant ethical concerns have been raised such as using placebo controlled RCTs during the initial development of antiretroviral drugs. RCTs may be less appropriate or feasible in certain sectors of medicine, including RCTs assessing surgical or psychological interventions [42]. A major limitation associated with RCTs, is cost. The infrastructure required has resulted in a burgeoning corporate enterprise, where a phase III RCT may cost up to \$30 million [43]. Contract Research Organisations have now become an industry worth \$25 billion [44] and such high pharmaceutical development costs may be cited as justification for inflated prescription drug costs. As a result of high costs, expectation of positive results has increased and evidence suggests that industry funded RCTs are more likely to achieve positive results than publically funded RCTs [45].

Considering the clinical treatment of epilepsy with antiepileptic drugs, publically funded, pragmatic RCTs assessing the longer term clinical and cost effectiveness provide the most informative data. Published in 2007, The Standard and New Antiepileptic Drug Trial (SANAD) is a notable example, providing evidence to support lamotrigine as a first line treatment for focal epilepsy [24]. The results subsequently influencing national guidelines [25]. However, pragmatic RCTs are expensive, time-consuming and resource intensive. Routinely recorded data in administrative healthcare databases have the potential to address these limitations and provide an alternative, accessible and informative data source for clinical research [46-48].

### ***1.3.1 The Standard and New Antiepileptic Drugs II (SANAD II) Trial***

The Standard and New Antiepileptic Drugs II (SANAD II (*EudraCT No: 2012-001884-64, ISRCTN30294119*)) RCT is the successor to the SANAD RCTs published in 2007 [23, 24]. SANAD II is a RCT assessing the clinical and cost effectiveness of lamotrigine, levetiracetam and zonisamide as first line treatments for patients' with newly diagnosed epilepsy and is funded by the National Institute of Health Research (NIHR) Health Technology Assessment (HTA) programme.

#### ***Population:***

A total of 1510 individuals with newly diagnosed epilepsy and previously untreated with antiepileptic drugs were recruited between April 2013 and May 2017. Similarly to the first SANAD RCT, participants were recruited into one of two arms; focal onset seizures (990 individuals) or generalised/unclassified seizures (520 individuals). The study centres included UK National Health Service (NHS) outpatient epilepsy, general neurology and paediatric clinics. Follow-up duration ranged from a minimum of two years to maximum of 5.5 years.

*Inclusion Criteria:*

- Aged five years or older
- Two or more spontaneous seizures that require antiepileptic drug treatment
- Not currently or previously treated with antiepileptic drugs
- Antiepileptic drug monotherapy considered the most appropriate option
- Willing to provide consent (patients parent/legal representative willing to give consent where the patient is aged under 16 years of age)

*Exclusion Criteria:*

- Provoked seizures (e.g. alcohol)
- Acute symptomatic seizures (e.g. acute brain haemorrhage or brain injury)
- Currently treated with antiepileptic drugs
- Progressive neurological disease (e.g. known brain tumour)

*Intervention:*

Participants diagnosed with focal onset seizures were randomised to levetiracetam, zonisamide or the study control lamotrigine.

Participants diagnosed with generalised or unclassifiable seizures were randomised to levetiracetam or the study control sodium valproate.

*Objectives:*

- *Primary Objective:*
  - Time to 12 month remission from seizures
- *Secondary Objectives:*
  - Time to treatment failure
    - Time to treatment failure due to inadequate seizure control
    - Time to treatment failure due to adverse events
  - Time to first seizure
  - Time to 24 month remission
  - Adverse events
  - Quality of Life
  - Health Economic Outcomes

*Methods:*

SANAD II is a pragmatic RCT and data were recorded using standard prospective methods.

A *baseline assessment* was completed at the time of recruitment. Data were recorded by a member of the study team on a Case Report Form (CRF) including seizure history, history of neurological insult, febrile seizures, family history of epilepsy, EEG and imaging (CT or MRI) results. Eligibility was then confirmed, written consent signed and the participant was randomised to a study AED. *Follow-up assessments* were then completed at three, six, twelve months and annually thereafter, integrated into routine clinical practice. Data were recorded on CRFs during each follow-up including medication history, adverse events, healthcare resource use and seizure occurrence. Details of seizures were recorded for the time period between each follow-up. The 'first' and 'last' seizure occurrences within the specified time period were recorded, together with the total number of seizures. The dates of all seizure occurrences were not recorded as this was not practicable. Additionally, during the follow-up period a number of self-completed questionnaires were requested recording details regarding adverse events, healthcare resources use and further information to inform the quality of life and cost-effectiveness analyses.

SANAD II opened in 2013, recruitment was completed in May 2017 and report is expected in 2019. Data are managed by the Clinical Trials Research Centre (CTRC), University of Liverpool.

## **1.4 Routinely Recorded Data**

### ***1.4.1 Introduction***

Routinely recorded data can be defined as data that are routinely recorded for specific primary purposes, other than audit or research [49].

In the UK, there is a plethora of individual-level data routinely recorded by a number of organisations. Broadly, data may be held for national, legal and governmental requirements, private industry purposes and for the benefit of the individual. Certain data are recorded with individual consent, for example the majority of private industry data. However, governmental data regarding taxation and clinical data, as examples, do not require individual consent. The Data Protection Act 1998 [50], to be replaced in the UK in May 2018 with The General Data Protection Regulation (GDPR) [51] and The Freedom of Information Act 2000 [52] provide the framework for data security, confidentiality and disclosure in the UK. Furthermore, recently the UK Department of Health has published a response to the National Data Guardian for Health and Care's review of healthcare data security, consent and opt-outs and the Care Quality Commission's review 'Safe Data, Safe Care' [53]. The document commits to improvements in healthcare data security, including sanctions following misuse and individual choice regarding data sharing. Legislation states data should be recorded to fulfil specific, defined purposes and the recording of excessive or unnecessary data should be avoided. Furthermore, legislation is in place to restrict access and to prevent inappropriate use of data. Where consent is required data must only be used for the defined purposes explicitly stated during the consent procedure. However, where consent is not required for certain routinely recorded data, use for secondary purposes may be permitted if there is a clear demonstrated 'secondary benefit'. Secondary benefit may broadly include benefits to population and society. Such secondary use has restrictions including strict ethical and data management requirements. As an example, data regarding clinical care is recorded without the need for consent in electronic medical records in the UK to assist with healthcare delivery and the remuneration for use of healthcare services [48, 54]. Such data are formatted and accessible for clinical research through National Health Service (NHS) Digital and Hospital Episode Statistics (HES) if secondary benefit can be demonstrated and the application requirements are fulfilled [55].

### **1.4.2 Routinely Recorded Clinical Data**

Data regarding clinical care are routinely recorded in electronic medical records and stored in administrative healthcare databases to assist with the delivery of healthcare and the remuneration for use of healthcare services [48, 54].

Electronic medical records are in use in developed countries worldwide and in different healthcare settings, such as primary and secondary care. There is variation in the recorded data variables which depend on for example healthcare setting, with different data recorded for emergency care compared to inpatient care and healthcare system, where different variables may be recorded in a private compared to public healthcare system. Examples of recorded data variables include diagnoses, investigations, treatments and procedures, age, gender, and area of residence [48]. Data is structured by individual patient, with each record including data variables common to all records in the database [47]. Data is frequently coded and codes may differ depending on country, healthcare setting and individual database. For example, electronic medical records within an individual hospital Patient Administration System (PAS) in the UK may use a different coding system to data recorded in a national database, such as the Hospital Episode Statistics (HES) database recorded by NHS Digital [55]. However, standardised diagnostic coding systems exist, the most commonly used in secondary care being the World Health Organisation's International Classification of Diseases system (currently in its 10th revision: ICD-10) [56]. In primary care in the UK, the NHS READ coding system is used in addition to ICD 10 [57]. Data are entered into the electronic medical records and administrative databases by clinicians directly or transcribed from medical records by trained clinical coders and administrative staff members.

#### *1.4.2.1 Routinely Recorded Clinical Data and Clinical Research*

Clinical research methodologies are heterogeneous but involve the common processes of identifying eligible individuals and recording data to assess study outcomes. Clinical research can be expensive, time-consuming and resource intensive. Routinely recorded data in administrative healthcare databases have the potential to provide an alternative, accessible and informative data source for clinical research [46-48]. Furthermore, there are potential advantages in resource and time efficiency, impacting on the feasibility and funding, scope and time to study completion. Such potential advantages are applicable to a number of study designs, possibly resulting in a more efficient research process and therefore earlier results and translation into clinical practice. The potential of routinely recorded data to inform clinical research has been recognised for some time, together with the limitations of accuracy, confidentiality, ownership and access [58].

Routinely recorded clinical data have been used in retrospective research studies including observational, record linkage population studies where administrative healthcare databases have proved an efficient means for providing study data [59]. In addition to the discussed advantages and relevant to retrospective studies, individual-level data may be available for a number of previous years resulting in the immediate availability of a study dataset for retrospective research. Finally, individual anonymity may be maintained more effectively using administrative healthcare databases where data is frequently de-identified, resulting in a less intrusive research process [48].

Routinely recorded data may also be of value in prospective research studies, including cohort studies and RCTs [60]. Routinely recorded clinical data regarding primary and secondary care have demonstrated utility in informing recruitment feasibility assessments for RCTs. For example, the TrialViz initiative developed by Dataline provides a means to perform a recruitment feasibility assessment using data provided by the Clinical Practice Research Datalink (CPRD) [61, 62]. The identification of specific individuals in administrative datasets eligible for a RCT may be possible, assisting with RCT recruitment [63]. Routinely recorded data also have the potential to measure prospective study outcomes [64]. For example, death certification data from the Office of National Statistics may be accessed to measure mortality [65] and HES data may be accessed to inform the health economic analyses within RCTs [66]. In selected cases it may be feasible to conduct a pragmatic RCT using administrative healthcare databases, including the stages of recruitment, randomisation, administration of intervention and follow-up assessments. Although there is limited evidence of the use of routinely recorded data for this purpose and pragmatic RCTs with simple interventions are most appropriate, cluster RCTs have been conducted entirely within CPRD, including patient recruitment, randomisation, administration of intervention and trial assessments [67]. *The Salford Lung Study* is an example of a large on-going RCT where both primary and secondary care data are accessed to measure the study outcomes [68, 69].

The potential advantages of routinely recorded data in the context of clinical research have resulted in a political drive to increase the use of administrative healthcare databases, as detailed in *The Plan for Growth* [70]. Consequently, the NHS constitution presents research as a 'core' activity of the NHS making the link between the provision of NHS services and research explicit [71].

#### *1.4.2.1.1 Limitations*

The limitations of routinely recorded data must be discussed when considering the use of data from administrative healthcare databases in clinical research. Routinely recorded data is by definition recorded to fulfil a specific, pre-defined primary purpose other than research or audit and only the data required to meet the specific purpose is recorded [49]. The accuracy of routinely recorded data for an alternative purpose, including clinical research where there may be a higher standard of scientific rigor, is therefore questionable [47, 72]. For example, in the UK the accuracy of routinely recorded clinical data is dependent upon the clinical information recorded by the clinician in the medical records and the accuracy of transcription by non-clinical administrative staff during the coding process [47]. However, using the diagnosis of epilepsy as an example, there are studies worldwide assessing the accuracy of algorithms using routinely recorded data to identify individuals with a diagnosis of epilepsy, compared to medical records. ICD-10 codes consistent with a diagnosis of epilepsy together with  $\geq 1$  antiepileptic drug (AED) recorded in the Australian National Hospital Morbidity Database resulted in a Positive Predictive Value (PPV) of 81.4% for diagnosis of epilepsy [73]. A similar study using multiple linked Canadian administrative healthcare databases found a PPV of 91.9% [74] and there are further studies with equivalent findings indicating algorithmic approaches using routinely recorded data are sensitive for identifying individuals with diagnoses of epilepsy [75-78].

A single routinely recorded dataset may not provide sufficient data to meet the study objective, requiring data for study participants from multiple datasets to be retrieved. A number of countries have integrated healthcare systems allowing for national administrative healthcare databases, such as the Swedish Hospital Discharge Register, the Danish National Hospital Register and the Canadian Chronic Disease Surveillance System. In these examples it is possible to retrieve routinely recorded data from electronic medical records for individuals across hospital inpatient admissions and emergency care, outpatient clinic and primary care attendances. In the UK, healthcare administrative databases are not universal and to obtain all of the data required to measure a study objective, a number of administrative healthcare databases may need to be accessed. Individuals can frequently be 'linked' between databases to address this limitation, usually using a 'unique identifier' such as National Health Service (NHS) Number. However, the detail of the identifying information between sources and the different identifying data variables recorded may result in the 'link' being of reduced accuracy or in some cases not possible.

There are also logistical limitations when using routinely recorded data, amplified when accessing data from multiple administrative databases. For example, there is potentially a significant time interval between the time of occurrence of an event and the availability of the record of the event in an administrative database. This presents less of a concern to retrospective studies however, when considering prospective studies where prompt reporting is both clinically important and may be a regulatory requirement, the delay in availability of data may be prohibitive. Furthermore, the costs required to retrieve routinely recorded data can be significant, despite the majority of data holders operating on a not-for-profit, cost recovery basis.

Finally, access to individual, identifiable routinely recorded data is an ethical concern [79] and there is the prevailing discomfort felt by professionals and the public of the use of routinely recorded data for secondary purposes. The belief that the 'right to privacy' is being finely balanced against the pursuit of data access for secondary purposes [80] is likely contributing to the difficulties reported in implementation of administrative databases in research due to the contradictions in the process perceived by professionals and the public [81]. Involving patients as important stakeholders and re-gaining their trust will be an essential factor in realising the individual and population healthcare benefits of analysing routinely recorded data [82, 83].

#### *1.4.2.2 Routinely Recorded Clinical Data Sources in the UK*

##### *Secondary Care*

National public sector organisations provide information technology and data systems for commissioners, analysts and clinicians in health and social care. Data is recorded to inform patient care, provide the data for remuneration for Hospital Trusts and is subsequently used to monitor and improve clinical services through research. *Table 1.1* introduces the administrative healthcare databases where access to individual-level, secondary care routinely recorded data for research is possible.

##### *Primary Care*

Data in electronic medical records are recorded routinely by the General Practitioner to inform patient care and remuneration, but are not currently available for research on a national basis. A number of organisations represent collaborations between governmental bodies or academic institutions and providers of primary care information technology systems. *Table 1.2* introduces the administrative healthcare databases where access to individual-level, primary care routinely recorded data for research is possible on a regional basis.

Table 1.1: Sources of Routinely Recorded Secondary Care Data

<p><b><u>NHS Digital [55]</u></b></p> <p><b>Data Access for Clinical Research:</b>  <i>The Data Access Request Service</i> provides a method of access to a number of routinely recorded datasets for England. <i>Hospital Episode Statistics (HES)</i> provides clinical, health and socio economic data for all secondary care attendances in England. Datasets include <i>Accident and Emergency, Admitted Patient, Outpatient, Adult Critical Care, Maternity</i> and selected <i>Patient Reported Outcome Measures</i>.</p> <p><b>Previous Experience in Clinical Research:</b>  HES data have been accessed for retrospective linkage studies [84] and to provide data for prospective studies, for example estimation of healthcare resource use or measuring outcomes such as long term mortality [85].</p>
<p><b><u>The NHS Wales Informatics Service (NWIS) [86]</u></b></p> <p><b>Data Access for Clinical Research:</b>  Data access can be facilitated through The NWIS Bespoke Analysis Service. The <i>Patient Episode Database for Wales (PEDW)</i> provides clinical, health and socio economic data for all secondary care attendances in Wales and is broadly comparable to the Admitted Patient HES dataset, with data regarding elective and emergency admissions and maternity care recorded. Additional datasets of relevance to this study include the <i>Emergency Department</i> and <i>Outpatient Datasets</i>.</p> <p><b>Previous Experience in Clinical Research:</b>  PEDW data have been accessed for retrospective analyses, for example analysis of the incidence of obstetric complication rates [87].</p>
<p><b><u>The NHS National Services Scotland; Information Services Division (ISD) [88]</u></b></p> <p><b>Data Access for Clinical Research:</b>  The <i>electronic Data Research and Innovation Service (eDRIS)</i> provides a method of access to ISD datasets including <i>Outpatient, General Acute / Inpatient, Emergency Department, Unscheduled Care, GP Out of Hours</i> and <i>The Prescribing Information System</i>. Clinical, health and socio economic data are recorded and datasets are largely comparable to HES.</p> <p><b>Previous Experience in Clinical Research:</b>  ISD data have been accessed for retrospective linkage studies, for example analysis of the incidence of gastrointestinal bleeding and complications including mortality [89].</p>

Table 1.2: Sources of Routinely Recorded Primary Care Data

<p><b><u>The Clinical Practice Research Datalink (CPRD) [90]</u></b></p> <p><b>Data Access for Clinical Research:</b> CPRD is a governmental research service jointly funded by the NHS National Institute for Health Research and the Medicines and Healthcare products Regulatory Agency. Following approval by the <i>Independent Scientific Advisory Committee</i>, CPRD provides access to de-identified primary care clinical, health and socioeconomic data for a geographically representative 13 million patients in England for healthcare research.</p> <p><b>Previous Experience in Clinical Research:</b> CPRD data have been used in retrospective studies for estimating healthcare resource use, prescription medicines and clinical outcomes [84]. Gulliford conducted two cluster-randomised trials using CPRD: one aimed to reduce inappropriate antibiotic prescribing for acute respiratory infection; the other aimed to increase physician adherence with secondary prevention interventions after first stroke [67].</p>
<p><b><u>ResearchOne [91]</u></b></p> <p><b>Data Access for Clinical Research:</b> ResearchOne is a collaboration between the University of Leeds and The Phoenix Partnership (TTP), developers of the SystmOne clinical database and IT system. De-identified clinical, health and socioeconomic data are available from primary, secondary and out-of-hours care settings for approximately 26 million patients in the UK.</p> <p><b>Previous Experience in Clinical Research:</b> ResearchOne data have been used in public health surveillance studies, retrospective studies [91] and currently in combination with CPRD data to measure the outcomes of a cluster RCT [92].</p>
<p><b><u>QResearch [93]</u></b></p> <p><b>Data Access for Clinical Research:</b> QResearch is a collaboration between the University of Nottingham and the developers of the EMIS IT systems. De-identified clinical, health and socioeconomic data are available for approximately 18 million patents in the UK.</p> <p><b>Previous Experience in Clinical Research:</b> QResearch data have been used to measure clinical outcomes in case-control and cohort studies [94].</p>
<p><b><u>The Health Improvement Network (THIN) Database [95]</u></b></p> <p><b>Data Access for Clinical Research:</b> THIN is a collaboration between IMS Health and In Practice Systems, developers of the IT software Vision. De-identified clinical, health and socioeconomic data are available for approximately 11.1 million patients in the UK.</p> <p><b>Previous Experience in Clinical Research:</b> THIN data have been accessed to measure clinical outcomes in cohort and case-control studies [96].</p>
<p><b><u>North West eHealth (NWEH) [97]</u></b></p> <p><b>Data Access for Clinical Research:</b> NWEH is a collaboration between The University of Manchester, Salford Royal Foundation Trust and Salford Clinical Commissioning Group. NWEH has developed the methodology and governance framework to implement the <i>Salford Integrated Record</i>, an integrated primary and secondary care electronic medical record, into research as part of the Salford Lung Study [69]. The infrastructure permits access to secondary care electronic medical records accessed through the NHS Digital <i>Secondary Uses Service</i>. With participant and GP practice enrolment and consent, the Apollo [98] and Graphnet [99] data extraction tools are employed to extract participant primary care electronic medical records that can then be linked to data regarding secondary care. NorthWest eHealth is unique in that data is not de-identified and therefore participant consent is required. Furthermore, GP practice enrolment and consent is required to permit the installation of third party software on their systems and subsequent extraction of data.</p> <p><b>Previous Experience in Clinical Research:</b> NWEH offer a number of primary care research tools including a RCT recruitment feasibility assessment, but do not currently routinely provide a bespoke primary care data extraction service for research. However, the methodology for this process has been demonstrated [69].</p>

### *'Linked' Routine Data Sources*

In order to provide a 'complete' dataset including all information required to meet research objectives, data from multiple administrative databases, both clinical and non-clinical, may need to be accessed and linked either on an individual or aggregate level. This is typically accomplished using unique identifiers such as name, date of birth, National Insurance Number or NHS Number. In response to the growing recognition of the potential of routinely recorded data, initiatives have been established to assist with the provision of linked, de-identified data between data sources:

- ***The Secure Anonymised Information Linkage (SAIL) Databank*** is an initiative developed by Swansea University and funded by the Welsh Government. SAIL provides a method of access to individual-level, routinely-recorded, de-identified electronic data for patients across Wales to support research [100]. Access to clinical datasets provided by NWIS is complemented with numerous non-clinical administrative datasets including births, deaths and demographic data. Following the scoping process a formal application is submitted to the *Information Governance Review Panel* before access to data is granted. SAIL data has been accessed to measure clinical outcomes in retrospective research [101].
- ***The Administrative Data Research Network (ADRN)*** is a UK-wide partnership between universities, government departments, national statistics authorities, funders and researchers, funded by the *Economic and Social Research Council*. ADRN provides a method of access to a number of non-clinical administrative routine datasets including employment, socioeconomic, crime and education data [102] in addition to clinical datasets detailed previously such as those recorded by NHS Digital. Following development of a project proposal a formal application is reviewed by the *Approvals Panel*.

### ***1.4.3 Routinely Recorded Non-Clinical Data***

'Non-clinical data' includes data recorded for national, legal and governmental requirements, private industry purposes and for the benefit of the individual. A plethora of 'non-clinical data' are routinely recorded in the UK. As previously discussed in *Section 1.4.1*, individual consent is required for certain data, for example data recorded for private industry purposes. Such data must only be used for the explicit purposes included in the consent procedure. However, there are non-clinical data recorded without the requirement for consent and in such circumstances there is potential for secondary use, including use for clinical research. For example in the UK, HM Revenue and Customs (HMRC) record individual-level data regarding employment status, salary, taxation and national insurance contributions [103]. Such routinely recorded data have the potential to inform clinical research.

#### ***1.4.3.1 Routinely Recorded Non-Clinical Data Sources in the UK and Clinical Research***

Routinely recorded non-clinical data have the potential to contribute to clinical research, including measuring clinical study outcomes. When considering the potential implementation of non-clinical data in clinical research, the specific use is dependent on the nature of the data. Example sources of UK non-clinical routinely recorded data relevant to clinical research are discussed.

The Driver and Vehicle Licensing Agency (DVLA) is responsible for the licensing of drivers and vehicles in the UK and issuing, reviewing and maintaining guidance regarding driver license status in the context of medical diagnoses [18]. The DVLA therefore record personal data relevant to driving and there is a legal requirement for individuals to inform the DVLA of relevant medical diagnoses. Using epilepsy as an example, the legal requirement for driving license holders to inform the DVLA of the occurrence of seizures and subsequently to regain normal driving privileges after a specified period of seizure freedom, raises the possibility of DVLA providing an accurate data source to measure clinical outcomes in epilepsy research. The DVLA publish limited de-identified, aggregate datasets for research, usually assessing driving restrictions. However, there was no evidence of individual-level DVLA data being accessed for clinical research in a scoping search performed in MEDLINE via OVID.

HM Revenue and Customs (HMRC) [103] is responsible for taxation including National Insurance and Student Loan repayments and the administration of tax credits, child benefit and statutory sick and maternity pay. The Department for Work and Pensions (DWP) [104] is responsible for the provision of state pensions and welfare benefits. Relevant data is therefore routinely recorded and includes details regarding employment status, salary, tax and National Insurance contributions and the receipt of benefits. Such routinely recorded data has the potential to inform health economic analyses in clinical research, including an assessment of the broader societal impacts of healthcare interventions. As a result of the primary purpose of data collection, data routinely recorded by HMRC and DWP are likely to be accurate and complete compared to standard methods used in prospective clinical research studies. For example, data to inform health economic analyses may be obtained through completion of self-report questionnaires in clinical research and such data are frequently poorly recorded [105]. Both HMRC and DWP are involved in external research and following approval permit access to de-identified aggregate datasets. However, there are restrictions, for example the *HMRC Datalab* requires research to include a 'listed function' of the HMRC of which clinical research is not [103]. There is no precedent for individual-level HMRC [106] or DWP data being accessed for clinical research.

The Office for National Statistics (ONS) [107] is an administrative database routinely recording a multitude of data including individual-level birth and mortality data and aggregate economic and societal statistics. The smallest reported level is the Lower Layer Super Output Area (LSOA) consisting of a population of 1000-3000. Routinely recorded data retrieved from ONS have the potential to measure clinical study outcomes. Mortality data can be requested through application to the NHS Digital DARS Service and there are examples of mortality being measured in retrospective and prospective studies [85]. Additionally, aggregate data can be accessed via services provided by ONS such as NOMIS [108] and Data for Neighbourhoods and Regeneration [109] and have the potential to contribute to health economic analyses in addition to data collected using standard methods.

Routinely recorded non-clinical data share many of the potential advantages of routinely recorded clinical data when considering use in clinical research, including the potential for improved feasibility and efficiency. However, in addition, the 'non-clinical' data discussed are likely to be of improved accuracy compared to similar data recorded using standard methods in clinical research. For example, data to inform an assessment of the broader societal impacts of healthcare interventions such as employment status retrieved from HMRC, are likely to be of greater accuracy compared to standard research methods such as completion of self-report questionnaires, known to be poorly recorded [105]. Similarly, many of the discussed limitations are applicable. However, in the context of non-clinical sources the ethical concerns regarding routinely recorded data access for research [79] are likely to be of greater significance. Concerns have been raised regarding the 'right to privacy' when routinely recorded clinical data is accessed for clinical research. Such concerns are likely to be amplified when the access to non-clinical data for the purpose of clinical research is proposed. As is the case with clinical data, the public are an important stakeholder and their involvement will be essential in realising the individual and population healthcare benefits of routinely recorded non-clinical data [82, 83].

## **1.5 Conclusions and Research Objectives**

Publically funded, pragmatic RCTs assessing the longer term clinical and cost effectiveness of antiepileptic drugs in the treatment of newly diagnosed epilepsy provide the data most informative to the clinical treatment of epilepsy. However, pragmatic RCTs are expensive, time-consuming and resource intensive. Routinely recorded data in administrative healthcare databases have the potential to address these limitations and provide an alternative, accessible and informative data source for clinical research [46-48]. However, although the 'accuracy' of the diagnosis of epilepsy using routinely recorded data compared to medical records has been assessed [73], there is minimal evidence of the assessment of 'agreement' to standard methods of data collection employed in prospective research. Acknowledging the rapidly increasing use of routinely recorded data in prospective research including RCTs and the status of the RCT in remaining the standard for approval of novel treatments in healthcare [110], an assessment of the feasibility and agreement of routinely recorded data compared to data collected using standard prospective methods is pressing.

This thesis will review the use of routinely recorded data in prospective research and agreement compared to standard prospective data collection methods, review accessible sources of routinely recorded data in the UK and assess the agreement and feasibility of using routinely recorded data compared to data collected in a RCT assessing treatments for epilepsy using standard prospective methods.

The objectives of this thesis are as follows:

- 1. Review the Use of Routinely Recorded Data in the UK to Assess Outcomes in Randomised Controlled Trials (Chapter Two)***

This review summarises the use of individual-level routinely recorded data from specified data sources in the UK to inform the assessment of outcomes of RCTs. An electronic database search using MEDLINE via OVID and a narrative review of additional relevant resources was completed.

- 2. Review the Agreement of Routinely Recorded Data with Data Collected Using Standard Prospective Methods in UK Studies (Chapter Three)***

This systematic review summarises the assessments of agreement between routinely recorded data in the UK and data collected using standard prospective methods to measure the outcomes in prospective clinical studies, including RCTs.

**3. *Identify and Assess the Accessibility of UK Routinely Recorded Data Sources (Chapter Four)***

Relevant sources of routinely recorded data in the UK are presented, followed by an assessment of the 'accessibility' for the purposes of retrieving data to measure the outcomes in a RCT.

**4. *Compare the Attributes of Data Extracted from Electronic Medical Records Against Data Collected Using Standard Methods, in the Randomised Controlled Trial (RCT), SANAD II: (Chapters Five, Six, Seven)***

*a. Assess the Quality of Data Extracted from Electronic Medical Records*

The 'quality' of routinely recorded data is assessed, including an assessment of the 'comparability' and 'completeness' compared to data collected using standard prospective methods in SANAD II

*b. Assess the Agreement between Data Extracted from Electronic Medical Records and Data Collected Using Standard Prospective Methods*

Agreement between routinely recorded data and data collected using standard prospective methods in SANAD II is assessed for a number of variables and outcome measures.

**5. *Assess the Feasibility and Efficiency of Accessing and Using Routinely Recorded Data from Electronic Medical Records (Chapter Eight)***

The feasibility and efficiency of the process of retrieving routinely recorded data are assessed before recommendations for future improvement are proposed.



## **Chapter Two**

### **The Use of Routinely Recorded Data in the UK to Assess**

### **Outcomes in Randomised Controlled Trials: A Review**

#### **2.1 Introduction**

In Chapter One, epilepsy and the current prospective research methods used to assess treatments for epilepsy were discussed. The case study RCT, the Standard and New Antiepileptic Drugs II (SANAD II) RCT was introduced. Routinely recorded data in the UK and the potential for use in clinical research were introduced before finally the objectives of this thesis were presented. In this chapter the use of routinely recorded data in RCTs in the UK will be reviewed.

There is potential for use of routinely recorded data in clinical research and Health Technology Assessment [46] and access for 'secondary purposes' including clinical research is permitted providing there is demonstrable secondary benefit.

There are numerous examples of retrospective observational, record linkage population studies where routine sources have proved to be a valid and efficient method for providing data for clinical research [59]. In the context of prospective clinical research such as RCTs routinely recorded data have been used to inform judgements about the feasibility of sample size and recruitment targets [62] and measuring participant outcomes [46, 64]. Pragmatic cluster RCTs have been coordinated through routine data sources including participant recruitment, randomisation, administration of intervention and trial assessments, such as through The Clinical Practice Research Datalink (CPRD) [67].

The majority of RCTs incur costs as clinicians assess participants, record outcomes and complete Case Report Forms - hence using routinely recorded data may provide an efficient alternative method for data collection in addition to reducing the burden on investigators and participants. Furthermore, data from non-clinical routine sources may inform outcomes beyond the standard RCT assessments of clinical efficacy and effectiveness. For example, cost data (such as use of healthcare resources) and socio-economic data (such as employment and means-tested benefits data) may inform health economic analyses and the assessment of the broader societal impact of healthcare interventions. However, limitations with accuracy of coding, confidentiality, ownership and access have been identified as significant barriers to using routinely recorded data for research [58].

## **2.2 Objective**

This review aims to summarise published reports of the use of individual-level routinely recorded data from specified data sources in the UK to inform the assessment of outcomes of RCTs.

## **2.3 Methods**

This review includes a search of the electronic database, MEDLINE via OVID and a narrative review of additional relevant resources.

### ***2.3.1 Electronic Database Review***

#### ***2.3.1.1 Registration***

A protocol for this electronic database review has been prospectively developed. However, with the focus concerning data sources and methodological approaches to clinical research, the review was not eligible for registration in the PROSPERO Database. This report has been structured according to the PRISMA Checklist where relevant, included in *Appendix A, Table A.3*.

### 2.3.1.2 Inclusion Criteria

#### *Study Designs, Participants, Interventions and Outcome Measures*

Clinical RCTs were included that accessed individual-level data relevant to clinical research from UK routine data sources. This methodological review included studies meeting this criterion regardless of aim, clinical diagnosis, participants, interventions or outcome measures. It was therefore expected that included studies would be heterogeneous. As a result of limitations in the resources required for translation and the nature of the review, only English language studies were included.

#### *Routine Data Sources*

Electronic medical records of patients' use of secondary care services in the UK are routinely managed on a national basis. Each country has a governmental body that is the national provider of information, data and IT systems for commissioners, analysts and clinicians in health and social care. Data are recorded to inform patient care, provide the data for remuneration and subsequently used to monitor and improve clinical services through clinical research. Electronic medical records of patients' use of primary care services in the UK are recorded routinely by the General Practitioner to inform patient care and for remuneration, but are not currently available for clinical research on a national basis. A number of organisations represent collaborations between governmental bodies or academic institutions and private providers of primary care IT systems. Included clinical routine data sources are presented in *Box 2.1* and *Tables 1.1* and *1.2*.

Non-clinical, individual-level data are recorded routinely by UK governmental bodies for specific indications. Sources routinely recording data potentially informative to prospective clinical research and included in this review are presented in *Box 2.1* and *Table 1.3*.

In response to the growing recognition of the potential of routinely recorded data, initiatives have been established to assist with the provision of linked, de-identified, aggregate data between data sources. *Box 2.1* and *Section 1.1.2.2* present sources included in this review.

There are a number of smaller, disease specific routinely recorded data sources, for example individual disease registers. In this review, we have purposively included the larger sources with regional or national coverage that hold clinical and non-clinical information potentially applicable to any disease area, in order to ensure the results are generalizable. RCTs using data accessed from individual disease registries and other smaller routinely recorded data sources have therefore not been included.

*Box 2.1: Included Sources of Routinely Recorded Data in the UK*

<p><b><u>Routinely Recorded Clinical Data</u></b></p> <ul style="list-style-type: none"><li>- NHS Digital [55]</li><li>- The NHS Wales Informatics Service (NWIS) [86]</li><li>- The NHS National Services Scotland; Information Services Division (ISD) [88]</li><li>- The Office for National Statistics (ONS) [107]</li><li>- The Clinical Practice Research Datalink (CPRD) [90]</li><li>- The General Practice Research Database (GPRD) [90]</li><li>- ResearchOne [91]</li><li>- QResearch [93]</li><li>- The Health Improvement Network (THIN) Database [95]</li><li>- North West eHealth (NWEH) [97]</li></ul> <p><b><u>Routinely Recorded Non-Clinical Data</u></b></p> <ul style="list-style-type: none"><li>- HM Revenue and Customs (HMRC) [103]</li><li>- The Department for Work and Pensions (DWP) [104]</li><li>- The Driver and Vehicle Licensing Agency (DVLA) [18]</li></ul> <p><b><u>Routinely Recorded Linked Data</u></b></p> <ul style="list-style-type: none"><li>- The Secure Anonymised Information Linkage (SAIL) Databank [100]</li><li>- The Administrative Data Research Network [102]</li></ul>
--

### *2.3.1.3 Search Strategy*

A search strategy was developed for the electronic database MEDLINE via OVID. The search was developed using an iterative process using Index and MeSH terms, subheadings and free text terms. The final search strategy is included in *Appendix A, Table A.4*. The search has been developed to ensure maximal sensitivity, with no clinical diagnoses, interventions or outcome measures specified. The included routine data sources have been purposively included together with selected generic terms and abbreviations combined using the Boolean operator AND with the National Institute of Health and Care Excellence (NICE) UK filter and Cochrane RCT Highly Sensitive Search Strategy.

### *2.3.1.4 Study Identification*

The study title and abstract of all studies identified in the search were reviewed. The full text was retrieved for studies meeting the inclusion criteria, studies possibly meeting the inclusion criteria and studies where insufficient detail could be obtained from the screening procedure. Studies identified in multiple publications were included under the same study name, the source providing the richest data included in the analysis.

### *2.3.1.5 Data Extraction*

Data were extracted from all included studies onto a standardised electronic data extraction template. *Box 2.2* presents the extracted data items.

### Box 2.2: Extracted Data Items

- Study ID
- Reference
- Date of publication
- RCT design
- Routine data source
- Clinical focus
  - o Clinical speciality
  - o Diagnoses
- Implementation of routine data
  - o Assessment of recruitment feasibility
  - o Recruitment to RCT
  - o Dataset for primary analysis
  - o Dataset for secondary analysis
- Outcome measures
  - o Clinical
    - Efficacy
    - Harm
  - o Mortality
  - o Health Economic
- Appraisal
  - o Study reported limitations of implementing routinely recorded data
  - o Study reported advantages of implementing routinely recorded data

#### 2.3.1.6 Data Analysis

The objective was to review reports of the use of routinely recorded data from UK routine data sources in RCTs rather than appraise the individual outcomes of the RCT. We planned to perform a narrative assessment of the risk of bias for all studies and formal assessment where routinely recorded data had been used in the study and there was the potential for the introduction of bias, for example when routinely recorded data was used to address missing RCT data collected through standard methods, the risk of bias was assessed using the Cochrane Risk of Bias Tool. However, there was no planned routine formal risk of bias assessment for RCTs, assessment of heterogeneity, reporting bias, sensitivity analyses or meta-analyses.

The results were analysed using simple descriptive statistics and a narrative appraisal. Further statistical manipulations including meta-analysis were not appropriate due to the heterogeneity of aims, interventions, outcomes measures and routinely recorded data sources accessed.

### **2.3.2 Narrative Review**

A narrative review was completed to complement the electronic database search.

Resources that may include eligible RCTs were identified. As a result of the nature of the review, manual searching of journals and conference abstracts and the contact of researchers in the field was not feasible. This review aimed to assess methodology and specifically sources used to provide data for RCTs. In completed systematic reviews with a focus on methodology, relevant details were poorly indexed in electronic databases [111]. This limitation results in the requirement to potentially review 'all' UK RCTs if a rigorous systematic review is to be completed. Focussing solely on the review of RCTs included in specific conference proceedings or journals would not represent a valid approach. It was not feasible to review 'all' UK RCTs during this review.

An alternative more specific approach was taken to narratively review the publically available data release information published by the included routinely recorded data sources, where available. This narrative review of published data release registers was completed for the included routinely recorded data sources detailed in *Box 2.1*.

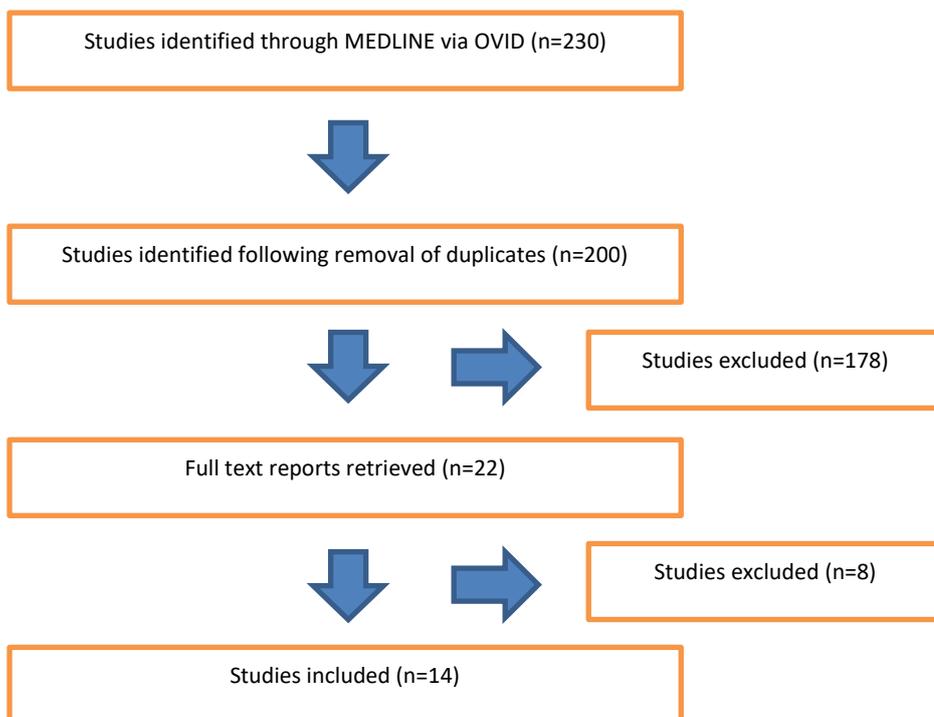
## 2.4 Results

### 2.4.1 Electronic Database Review

The titles and abstracts of the 200 studies identified in the search, completed 03/07/2016, were reviewed. The full text was retrieved and reviewed for 22 studies possibly meeting the inclusion criteria or where insufficient detail could be obtained from the abstract. Eight studies were excluded from the final review. Two studies were published protocols for included subsequent final study reports [112, 113]. Two studies accessed non-identifiable data from routine sources, one to provide the general mortality rate [114] and one to provide general data relevant to health economic analyses [115]. One study was a review article involving two studies included in this review [67]. One study accessed individual level data from a UK routine data source but in the context of socio economic research with no clinical outcomes [116]. Finally, one study accessed individual level data from a routine data source in the Netherlands to recruit and measure the outcomes in a RCT [117].

The process for the identification of studies is summarised in *Figure 2.1*.

*Figure 2.1: The Identification of Studies*



Fourteen studies were included in the review. A summary of the study characteristics including the implementation of routinely recorded data is presented in *Table 2.1*. Greater detail is provided in *Appendix A, Tables A.1 and A.2*. The Office of National Statistics (ONS) was the source of routinely recorded data most frequently accessed in the sample. Nine studies involved the access of ONS Mortality data to measure mortality in the short and long-term. The maximum duration of follow-up using ONS data was 20 years [118]. In most cases, follow up is reported as 98-99% complete for the sample, but the methods of determining participants remaining alive was only detailed in one study, and involved the ONS failing to identify 0.6% of participants, for example as a result of assumed emigration or non-specified loss to follow up [118]. Five studies included review of the International Classification of Diseases (ICD) information recorded on the death certificate to determine the cause of death. Two studies also included access to a subset of individuals' clinical documentation and reported 'good agreement' between cause of death recorded on the death certificate with cause of death identified on review of clinical documentation [118, 119]. The limitations of ONS data access are poorly recorded. One study identified the delay in death data being recorded in ONS [120], another identified 'limitations associated with ONS data' but with no further explanation [121]. One study protocol involved access to NHS Digital (The Health and Social Care Information Centre), Hospital Episode Statistics in addition to ONS data. During the RCT, clinical details including diagnoses and cause of death will be requested from NHS Digital. Participants will be prospectively flagged and the completion of 10 year follow up is due in 2017 [85].

The Secure Anonymised Information Linkage Databank (SAIL) was accessed in two RCT recruitment feasibility assessments [62, 122]. Pragmatic RCT inclusion criteria were applied to the SAIL dataset in order to identify individuals meeting the inclusion criteria and their location by GP practice. However, SAIL records only de-identified data which results in an inability to re-identify such eligible individuals. Therefore individual GPs in practices with eligible individuals would need to participate in the recruitment process if an RCT were to be conducted using datasets accessed through SAIL. The Clinical Practice Research Datalink (CPRD) was accessed during two cluster RCTs [123, 124] and access is planned in one further cluster RCT, due to report in 2017 [125]. In all cases, the RCT was coordinated through CPRD and the consented GP practices. GPs were recruited and simple, pragmatic interventions were included in the RCT. For example, introduction of decision support tools to prompt GPs of the guidelines for secondary prevention measures for stroke [123].

Participants' outcomes were recorded through CPRD and involved simple clinical measures such as assessment of blood pressure. The analysis of data in such studies was anonymous and patient consent was not required. There is a reported three month delay before data is available in CPRD, but the importance of this limitation is negated by the nature of the research question.

There were no RCTs identified with access to data from non-clinical sources including HM Revenue and Customs (HMRC), The Department of Work and Pensions (DWP) and The Driver and Vehicle Licensing Agency (DVLA).

Table 2.1: Included Studies: Summary of the Use of UK Routinely Recorded Data in RCTs

Study Reference	Trial Summary	Outcome Measures
<b>Office for National Statistics (ONS) (General Register Office (GRO), General Statistics Office of Ireland (GSOI))</b>		
Ashton et al: 2002 [119]	67,800 participants randomised to ultrasound abdominal screening or no intervention. Clinical details were recorded through standard prospective methods. Mortality was measured by accessing data from ONS.	Primary Outcome: Aneurysm related mortality Secondary Outcomes: All-cause mortality, frequency of ruptured aneurysm, quality of life
Bale et al: 2008 [121]	375 patients with moderate/severe chronic obstructive pulmonary disease were randomised to fluticasone or placebo for 3 years. Clinical details were recorded through standard prospective methods. Mortality was measured by accessing data from ONS.	Primary Outcome: All-cause mortality
Brown et al: 2012 [120]	1252 participants with abdominal aortic aneurysm (AAA) were randomised to EndoVascular Aneurysm Repair (EVAR) or laparotomy. 404 participants with AAA were randomised to EVAR or conservative management. Clinical details were recorded through standard prospective methods. Mortality was measured by accessing data from ONS.	Primary Outcome: Mortality (all-cause, operative, aneurysm-related)
Henderson et al: 2015 [126]	1810 patients with non-ST-segment elevation acute coronary syndrome were randomised to an early invasive strategy (coronary arteriography and myocardial revascularization) or a selective invasive strategy (coronary arteriography for recurrent ischemia only). 10 year mortality was measured in this study by accessing data from ONS for England and the GRO for Scotland.	Primary Outcome: All-cause mortality Secondary Outcomes: Mortality (cardiovascular or non-cardiovascular)
Molyneux et: 2015 [127]	1624 participants with ruptured cerebral aneurysm were randomised to neurosurgical clipping or endovascular coiling. 18 year mortality was measured by accessing data from ONS.	Primary Outcome: All-cause mortality Secondary Outcomes: Functional status, dependency
Perera et al: 2012 [128]	301 patients with severe ventricular impairment and coronary artery disease were randomised to Intra-Aortic Balloon Pump during Percutaneous Coronary Intervention (PCI), or PCI alone. Clinical details were recorded through standard prospective methods. Mortality was measured at a median 51 months by accessing data from ONS and GRO.	Primary Outcome: All-cause mortality
Scholefield: 2012 [118]	152 850 individuals by household were randomised to biennial Faecal Occult Blood screening vs no intervention. Clinical details were recorded through standard prospective methods. Mortality was measured at a median 19.5 years by accessing data from ONS.	Primary Outcomes: Mortality (all-cause, colorectal cancer related)
Simmons: 2012 [129]	20 185 participants in 33 GP's were randomised, by GP to Type 2 Diabetes Mellitus screening followed by intensive treatment, screening plus routine care or no-screening. Mortality was measured at a median 9.6 years by accessing data from ONS, GRO and CSOI.	Primary Outcome: All-cause mortality Secondary Outcomes: Death from CV disease, cancer, DM related death
<b>NHS Digital and Office of National Statistics (ONS)</b>		
Turner et al: 2014 [85]	785 GP's randomised to prostate specific antigen screening vs standard care in a cluster RCT. This reference reports the design and recruitment results. Clinical details including diagnoses and cause of death will be obtained through access to NHS Digital HES / ONS data. Follow up at 10 years due 2017.	Primary Outcome: 10 year 'definite' or 'probable' prostate cancer mortality Secondary Outcomes: All-cause mortality (10 and 15 years), cost effectiveness

<b><i>The Secure Anonymised Information Linkage Databank (SAIL)</i></b>		
Brooks et al: 2009 [122]	SAIL Databank was used as the data source to perform a recruitment feasibility assessment for two fictitious RCTs involving patients with diabetes mellitus and pragmatic inclusion criteria. Of 250,086 individuals in SAIL, 284 were eligible for the first RCT and 711 for the second.	N/A
McGregor et al: 2010 [62]	SAIL Databank was used as the data source to perform a recruitment feasibility assessment for an existing RCT assessing folate use in patients with depression. 867 potential participants were identified.	N/A
<b><i>The Clinical Practice Research Datalink (CPRD)</i></b>		
Dregan et al: 2014 [123]	106 participating GP's contributing data to CPRD were allocated to the intervention; installation of IT decision support tools to improve adherence to secondary care stroke prevention measures during the patient consultation or control; standard practice. RCT duration 12 months. Pragmatic clinical details including blood pressure and blood tests were recorded through CPRD to measure the study outcomes.	Primary Outcome: Systolic blood pressure  Secondary Outcomes: Diastolic blood pressure, total cholesterol, prescription of cardiovascular drugs
Guilliford et al: 2014 [124]	104 participating GP's contributing data to CPRD were allocated to the intervention; installation of decision support tools to improve adherence to antibiotic prescribing guidelines during the patient consultation for respiratory tract infection or control; standard practice. RCT duration 12 months. Pragmatic clinical details including prescription of antibiotics and record of respiratory diagnoses were recorded through CPRD to measure the study outcomes.	Primary Outcome: Proportion of consultations for respiratory tract infection with antibiotics prescribed  Secondary Outcomes: Proportion of antibiotics prescribed in other respiratory infective diagnoses
Horspool et al: 2013 [125]	Protocol for a cluster RCT involving 140 GP's contributing data to CPRD. GP's will be allocated to the intervention; a letter informing parents of the importance of adherence to their child's asthma treatment throughout the summer holidays or control; standard practice. The pragmatic clinical details will be recorded through CPRD to measure the study outcomes.	Primary Outcome: Unscheduled medical contact in September  Secondary Outcomes: Unscheduled medical contacts at other time points associated with prescriptions, respiratory diagnoses
<b><i>The Clinical Practice Research Datalink (CPRD) and ResearchOne</i></b>		
Herrett et al: 2014 [92]	Protocol for a cluster RCT involving GP's contributing data to CPRD or ResearchOne. GP's will be randomised to the intervention; a text messaging campaign to increase uptake of the flu vaccine or control; standard practice. The pragmatic clinical outcome of flu vaccine administration will be recorded through CPRD and ResearchOne to measure the study outcome.	Primary Outcome: Flu vaccine administration

## **2.4.2 Narrative Review**

A narrative review of the online electronic resources for each included routinely recorded data source was completed on 22/09/2016. Data release registers and bibliographies were reviewed where available, results are presented:

### *2.4.2.1 NHS Digital*

NHS Digital publishes data release registers usually on a three-monthly basis and includes details regarding data releases in England for HES and ONS mortality data [130]. The most recent publication at the time of review (April-August 2016) included >1000 individual releases of identifiable data. Searching the 'objective for processing' field with the term 'RCT' identified seven RCTs involving HES data with three including ONS mortality data. One RCT was also identified in the electronic database review [85]. Examples include a RCT assessing the treatment of coronary vascular disease, with the outcome of major coronary events, a RCT assessing the chemotherapy treatment of colon cancer with the outcome of cancer progression and a RCT assessing the treatment of abdominal aortic aneurysms with the outcome of overall and cause-specific mortality.

The data release registers prior to and including January -March 2016 provide less detailed information regarding the 'objective for processing' resulting in difficulty in accurately identifying data releases for use in RCTs. In the January-March 2016 release, 809 individual data releases are presented. Searching for the term 'RCT' in the 'purpose' field identified seven data releases that are related to RCTs, two of which were also included in the April-August 2016 release.

NHS Digital data release registers have been published on a three-monthly basis from April 2013 and all contain limited evidence of HES or ONS access to provide data for RCTs. Prior to April 2013, the study titles are presented, but there is no information presented regarding the 'objective for processing' or 'purpose' of the data releases. It was therefore not possible to identify relevant data releases prior to April 2013.

### *2.4.2.2 The Clinical Practice Research Datalink (CPRD)*

Data releases for studies approved by the CPRD Independent Scientific Advisory Committee are presented in an online resource. This resource includes all studies approved from July 2015 [131]. The studies have been searched using the terms 'RCT' and 'trial'. No approved studies involving data access for use in a RCT were identified.

#### *2.4.2.3 ResearchOne*

A summary of 'current projects' is published on the ResearchOne online electronic resource [132]. It is not clear if this summary includes previous data releases or completed projects, but this is unlikely to represent a formal data release register. One RCT is identified involving the release of ResearchOne data, also identified in the electronic database review [92].

#### *2.4.2.4 QResearch*

A summary of completed studies involving the provision of QResearch data is published on the online electronic resource [133]. Ongoing studies are also presented on a linked webpage. There was no evidence of QResearch data access for use in a RCT.

#### *2.4.2.5 The Health Improvement Network (THIN) Database*

A bibliography including the references of all completed studies involving access to THIN data is published on the online electronic resource [134]. The bibliography was searched using the terms 'RCT', 'trial' and 'randomised'. No data releases for a RCT were identified.

#### *2.4.2.6 North West eHealth (NWEH)*

A data release register for NWEH could not be identified following review of the available online electronic resources. However, one RCT coordinated by NWEH, the Salford Lung Study, being conducted using the 'Linked Database System' [135] is detailed.

#### *2.4.2.7 HM Revenue and Customs (HMRC)*

A bibliography of studies approved for the access to HMRC data is available on the online electronic resource [106]. No data releases for RCTs were identified and further, no data releases for clinical research of any methodology were identified.

#### *2.4.2.8 The Secure Anonymised Information Linkage (SAIL) Databank*

A bibliography of studies approved for access to SAIL data is available on the online electronic resource [136]. SAIL data access is planned in one RCT protocol for long term, anonymised follow up [137]. This RCT protocol was not identified in the electronic database review. There was one further RCT protocol identified, but on full review of the publication the role of SAIL was unclear. This study may be requesting use of the services regarding linkage of data provided by SAIL.

#### *2.4.2.9 The Administrative Data Research Network (ADRN)*

A bibliography of studies approved for access to ADRN data is available on the online electronic resource [138]. No data releases for use in a RCT were identified.

#### *2.4.2.10 Sources Lacking Data Release Registers*

Data release registers could not be identified for the following data sources:

- *The NHS Wales Informatics Service (NWIS) / Public Health Wales Observatory*
- *The NHS National Services Scotland; Information Services Division (ISD)*
- *The Department for Work and Pensions (DWP)*
- *The Driver and Vehicle Licensing Agency (DVLA)*

## **2.5 Discussion**

Routinely recorded data have demonstrated potential in prospective research including measuring the outcomes of RCTs [64] and providing additional benefits such as a method to address missing RCT data. Academic, political [70] and health service [71] interest in UK sources of routinely recorded data has resulted in expansion and improvements, notably in the access to linked datasets. However, in this review evidence of the use of routinely recorded data to assess the outcomes in UK RCTs was variable dependant on the data source and nature of routinely recorded data. Routinely recorded mortality and secondary care clinical data were most commonly accessed in RCTs. There was very limited use of routinely recorded primary care clinical data and no evidence of the use of data from non-clinical sources in RCTs.

In the electronic database review, ONS was the most frequently accessed data source, providing death notification and certification data. The legal requirements with regards to death certification and registration result in a largely complete dataset and identified studies reporting follow up for 98-99% of participants. However, deriving cause of death, based on medical certification diagnoses or clinical coding is likely to be less accurate than ascertaining death status, although two studies reported 'good agreement' between information derived from the ONS and clinical documentation. In one of these studies, there was some discrepancy between 'certified' and 'verified' causes of death, verified using case note review, but the overall result and significance of the primary outcome was unchanged [119]. The narrative review provided further evidence of the numerous RCTs including access to ONS mortality data. There was some discrepancy, with RCTs identified in the data release registers that were not included in the electronic database review. However, this discrepancy was more obvious with English NHS Digital HES data access, with only one RCT being identified in the electronic database review and numerous being identified in the data release registers. Similar data release registers could not be identified for the Welsh NWIS or Scottish ISD.

There was limited evidence of the use of routinely recorded primary care data. In the electronic database review, two recruitment feasibility assessments were performed using SAIL data and four cluster RCTs were conducted using CPRD data, with data from ResearchOne also involved in one RCT. In the narrative review, one additional RCT conducted through NWEH was identified. Primary care sources are limited by selective coverage based on General Practitioner (GP) consent and GP Patient Administration System and restrictions on the identification of individuals within the datasets. However, cluster RCTs have been conducted although simple pragmatic interventions and GP consent and participation to deliver the study intervention were required and notably, patient consent was not required. The NHS Digital General Practice Extraction Service is the only national record of primary care data and potentially represents the most informative data source. However access is currently limited to Department of Health initiatives such as national screening programmes [139].

Despite the potential of non-clinical routinely recorded data to measure outcomes beyond the standard RCT assessments of clinical efficacy and effectiveness, such as an assessment of the broader societal impacts of interventions, there was no evidence of use of non-clinical data.

Although poorly specified in the studies included in this review, it is likely that there are a number of limitations when accessing routinely recorded data in the context of prospective research. The process of quality assurance is unclear and the level of agreement of routinely recorded data with data recorded through standard RCT methods remains uncertain and was not reported in the majority of studies, particularly when measuring clinical outcomes. Furthermore, the time delay before routinely recorded data becomes available may have implications for RCTs where prompt reporting is both clinically important and in some instances a regulatory requirement, such as during pharmaceutical trials.

The discrepancy between the results of the electronic database review and narrative review is notable. RCTs including the use of ONS mortality data were identified from the electronic database review with a minority identified from the narrative review of the data release registers, but not included in the electronic database results. However, only one RCT using HES data was identified from the electronic database review compared to evidence of numerous RCTs in the review of NHS Digital data release registers, despite a search strategy developed to be sensitive. Such discrepancies highlight the poor indexing of methodological information in electronic databases. It is likely that ONS mortality data was used to assess RCT primary outcomes and therefore may be more likely to be included in the abstract and electronic database indexing. Conversely, HES data may be accessed as an additional dataset to assess secondary outcomes, such as healthcare resource use and this may explain the poor indexing.

## **2.6 Limitations**

The nature of the objective introduces limitations into the review. The review does not focus on clinical diagnosis, intervention or outcome but rather on methodology. The alternative focus on data sources accessed resulted in the electronic database search strategy being inclusive of all clinical RCTs but with a purposive search of specified data sources and relevant generic terms filtered for the UK. There is therefore a risk that studies poorly indexed or poorly documenting the data source were not identified. This limitation was most notable for the clinical data sources such as NHS Digital HES and was addressed by narrative review of the registers of data releases, where available.

It was not feasible to summarise the details of all RCTs involving the release of HES or ONS data from the NHS Digital data release registers. This was the result of resource and time limitations of the researcher during the review and the less detailed data published in the data release registers from earlier years, which would have necessitated reviewing 'all' data releases. However, the objective of the review was to summarise the use of routinely recorded data in RCTs and the identification of numerous RCTs in the data release registers informs this objective. Furthermore, the heterogeneity of the included studies did not permit any further manipulation of data and a narrative discussion is presented.

## **2.7 Conclusions**

In this chapter, the use of routinely recorded data in UK RCTs has been reviewed. Routinely recorded data may present benefits to prospective research including RCTs, but the overall experience of accessing data for this purpose remains limited. Registry mortality and secondary care routinely recorded data were most commonly accessed for RCTs. Primary care routinely recorded data were infrequently accessed but there was evidence of the feasibility for completing pragmatic cluster RCTs. Despite the potential for non-clinical data to measure outcomes beyond the standard clinical assessments, there was no evidence of use of data for this purpose in a RCT. Furthermore, a data release register could only be identified for HMRC and on review there was no evidence of data use for clinical research of any methodology.

The search of the electronic database was limited by the likely poor indexing of methodological details in electronic databases and therefore future reviews with a focus on methodology should be complemented with the manual review of additional relevant sources.

To further improve the use of routinely recorded data in RCTs, research is required to assess the accessibility, feasibility and agreement of routinely recorded data compared to data recorded through standard RCT methods. In the following chapter, the agreement between routinely recorded data and data collected using standard methods will be assessed in a systematic review.

## **Chapter Three**

### **The Agreement of Routinely Recorded Data with Data**

### **Collected Using Standard Prospective Methods in UK Studies:**

### **A Systematic Review**

#### **3.1 Introduction**

In Chapter One, epilepsy and the current prospective research methods used to assess treatments for epilepsy were discussed. The case study RCT, the Standard and New Antiepileptic Drugs II (SANAD II) RCT was introduced. Routinely recorded data in the UK and the potential for use in clinical research were discussed. In Chapter Two the use of routinely recorded data in RCTs in the UK was reviewed. Mortality data and secondary care data were most commonly accessed, although there was relatively limited use in general. In this chapter the agreement between UK routinely recorded data compared to data collected using standard methods in prospective studies will be assessed in a systematic review.

There is potential for use of routinely recorded data in prospective clinical research, including RCTs [46]. There are examples of studies assessing the 'validation' of routinely recorded data. Such studies compare the agreement of data within a routine source to the 'source' data that is usually also 'routinely recorded'. For example assessment of the agreement of diagnoses retrieved from the General Practice Research Datalink (GPRD) to the directly accessed primary care medical records [140]. When considering the increasing use of routinely recorded data in prospective research including RCTs, of greater relevance is an assessment of the agreement between routinely recorded data and data collected through 'standard prospective methods', such as self-reported questionnaires or completion of Case Report Forms (CRFs).

## **3.2 Objective**

This systematic review aims to assess the agreement of routinely recorded data in the UK to data collected using standard prospective methods to inform the assessment of outcomes in prospective clinical studies including RCTs.

## **3.3 Methods**

### ***3.3.1 Registration***

As the focus of this systematic review was on methodological approaches and the agreement of data from alternative sources, the protocol was not eligible for registration in the PROSPERO Database. This systematic review report has been structured according to the PRISMA Checklist where relevant, included in *Appendix B, Table B.4*.

### ***3.3.2 Inclusion Criteria***

#### ***3.3.2.1 Study Designs, Participants, Interventions and Outcome Measures***

The method of data collection rather than the study design was of primary interest in this review. However, following the development of the search strategy, detailed below, there was justification for including only prospective study designs such as RCTs and non-randomised studies (non-randomised controlled trials, prospective cohort studies).

This methodological review included studies regardless of aim, diagnosis, participants, intervention or outcome measures and therefore heterogeneity was expected. As a result of resource limitations for translation and the objective of the review, only English language studies were included.

#### ***3.3.2.2 UK Routine Data Sources***

Relevant clinical and non-clinical sources of routinely recorded data in the UK with the potential to inform prospective clinical research were identified and purposively included, previously presented in *Box 2.1*. Additionally, to ensure identification and inclusion of other relevant sources of routine data, generic, descriptive terms such as 'administrative data', 'medical records', 'routine data' and 'electronic data' were included in the search strategy.

### **3.3.3 Standard Prospective Methods**

An inclusive approach was taken for the definition of 'standard prospective methods'. This included data that were collected as part of any prospective study (such as RCT, non-randomised controlled trial, prospective cohort) and included study follow up assessments and the completion of study documents such as CRFs as well as the completion of self, family or carer reported questionnaires. 'Standard' methods may where prospectively planned within the study protocol involve review of local medical records. For example, prospective research nurse review of individual hospital Patient Administration Systems to measure patient contact with secondary care services and transfer of this data to a CRF would reasonably represent a 'standard prospective method' and a comparison to routinely recorded data such as Hospital Episode Statistics (HES) data provided by NHS Digital was informative to the objective of this systematic review.

### **3.3.4 Assessment of Agreement**

To ensure an inclusive approach, all comparisons between UK routinely recorded data and data collected using standard prospective methods were included. Matched comparisons on an individual level as well as a cohort level were included. Eligible methods included simple descriptive comparisons, assessment of statistically significant differences and formal statistical assessments of agreement, such as calculation of Cohens Kappa, the Intraclass Correlation Coefficient or construction of Bland Altman Plots. In order to maximise the quality of the review, studies involving only narrative comparisons, without the inclusion of numerical data were excluded.

### **3.3.5 Search Strategy**

To ensure identification of all relevant studies, regardless of publication status (published, unpublished, in press, ongoing) search strategies were developed for electronic databases and a manual search of relevant resources was completed.

Limitations in the specificity of the electronic database searches was expected, informed by the experience of published systematic reviews with a primary focus on methodology [111] and the previous electronic database search in *Chapter Two*. Acknowledging these expected limitations and to ensure identification of all relevant studies, search strategies were developed for electronic databases and additional approaches were taken to identify studies with a primary focus on health economic analyses, including assessment of healthcare resource use.

### 3.3.5.1 Development of the Electronic Database Search Strategies

A search strategy for electronic databases was developed using an iterative process, initially in MEDLINE via OVID using Index and MeSH terms, subheadings and free text terms. The strategy was subsequently adapted to SCOPUS (including EMBASE) and the Cochrane Methodology Register. As a result of the primary focus on methodology, in this section the development of the search strategy is detailed including the total number of studies and total number of eligible studies identified in each iteration, summarised in *Table 3.1*. The final search strategies are presented in *Appendix B, Tables B.5-B.7*.

To inform the development of the search strategy, a scoping search was completed using MEDLINE via OVID, SCOPUS and a manual review. Nine studies eligible for inclusion were identified [66, 141-148]. Generic terms to describe routinely recorded data were combined using the Boolean operator AND with the National Institute of Health and Care Excellence (NICE) filter for UK studies. Together with purposively included UK routine data sources, the initial search, informed by the nine studies identified in the scoping searches, included a number of 'comparator' terms. Text words such as 'compare', 'agree', or 'valid' adjacent to text words such as 'self-report', 'questionnaire' or 'survey' were included. Truncation and wildcards were used to account for variations in spelling and identify different derivations of search terms. This initial search, not limited by study design identified 207 studies. Six of the nine studies identified in the scoping searches were included, confirming the strategy was identifying relevant studies, although sensitivity was limited.

In the next iteration, in an effort to increase the sensitivity, the 'comparator terms' were removed. Not limited by study design, 4423 studies were identified including seven of the nine eligible studies identified in the scoping searches. To improve the specificity of this strategy, the search was limited to prospective studies and 592 studies were identified, including seven of the nine eligible studies. To ensure the search for prospective studies was not omitting eligible studies, the 538 studies published in 2015-2016 from the search not limited by study design were reviewed. One additional eligible study was identified that was not identified in the search limited to prospective studies.

The final strategy included generic terms to describe routinely recorded data filtered for the UK and purposively included UK routinely recorded data sources, with results limited to studies with prospective designs. The search was identifying relevant studies, including seven out of the nine studies identified in the scoping searches. The two studies not identified in the MEDLINE search during the development of the search strategy were not included in the MEDLINE database and were identified in SCOPUS [66, 142].

The search strategy was adapted to SCOPUS (including EMBASE) identifying 1911 studies, including eight of the nine eligible studies. Subsequently the search was adapted to the Cochrane Methodology Register, the most relevant database of the Cochrane Library. Forty eight studies were identified when limited to those with prospective designs. The final search, not limited by study design identified 89 studies. Four of the nine eligible studies were included.

The final search strategy applied to MEDLINE, SCOPUS and the Cochrane Methodology Register, included the nine studies identified in the scoping review in addition to 17 eligible studies. A further 10 studies were also identified but excluded on subsequent review of the full report.

Table 3.1: The Development of the Electronic Database Search Strategies

Search Strategy	Total Studies	Scoping Studies Included	Scoping Studies Eligible	Additional Studies Included	Additional Studies Excluded
<b>Search Strategies: Development</b>					
MEDLINE "All Study Designs with Comparator Terms" 29.07.16	207	[141, 143, 144, 146-148]	9	N/A	N/A
MEDLINE "All Study Designs" 1.08.16	4423	[141, 143-148]	9	N/A	N/A
MEDLINE "All Study Designs 2015 -2016" 1.08.16	538	Nil	2	[149]	N/A
<b>Search Strategies: Final Iterations</b>					
MEDLINE "Prospective Studies" 1.10.16	592	[141, 143-148]	9	[150-160]	[117, 161-165]
SCOPUS "Prospective Studies" 19.10.16	1911	[66, 142-148]	9	[64, 65, 152-154, 157, 158, 160, 166-168]	[117, 165, 169-171]
Cochrane Methods Register "All Study Designs" 1.10.16	89	[143, 144, 146, 147]	9	[46, 64, 154, 172]	[117]

### 3.3.5.2 Identifying 'Health Economic' Studies

An additional search was completed to maximise the inclusion of studies with a primary focus on health economic measures, including the assessment of healthcare resource use. The Database of Instruments for Resource Use Measurement (DIRUM) is an open access database of resource use questionnaires [173]. The questionnaires and relevant methodological research papers are included. In the development of the database, an electronic search strategy was developed and MEDLINE, EMBASE and PsycINFO were searched (05/2012). The methodological studies included in the DIRUM database were reviewed in this review and in addition, the search strategy was updated in MEDLINE via OVID and adapted for use in SCOPUS (including EMBASE) to identify studies published during the period 2012-2016. The search of the DIRUM database and subsequent updates are detailed in *Table 3.2*. The updated search strategies are presented in *Appendix B, Tables B.8-B.9*.

*Table 3.2: The Existing and Updated DIRUM Searches*

<b>Data Source</b>	<b>Total Studies</b>	<b>Eligible Studies</b>
Database of Instruments for Resource Use Measurement (DIRUM) 1.10.16	94	[158-160, 167, 174-177]
DIRUM MEDLINE 2012-2016 21.10.16	327	[174, 178-180]
DIRUM SCOPUS 2012-2016 21.10.16	331	[180]

### 3.3.5.3 Manual Searches

To ensure a sensitive approach a manual search was completed and is summarised in *Table 3.3*. The following resources were reviewed:

- Health Technology Assessment (HTA) Journal
- International Clinical Trials Methodology (ICTMC) Conference Proceedings 2015
- Scottish Health Informatics Programme (SHIP) Conference Proceedings 2013

*Table 3.3: Results of the Manual Searches*

<b>Data Source</b>	<b>Eligible Published Studies</b>	<b>Eligible Unpublished Studies</b>
Health Technology Assessment (HTA) Journal 15.06.16	[46, 64, 159, 177]	N/A
International Clinical Trials Methodology (ICTMC) Conference Proceedings 2015 16.11.15	N/A	[181-183]
Scottish Health Informatics Programme (SHIP) Conference Proceedings 2013 28.08.13	N/A	[184]

### **3.3.6 Study Identification**

The study title and abstract of all studies identified in the searches were reviewed. The full text was retrieved for studies meeting the inclusion criteria, studies possibly meeting the inclusion criteria and studies where insufficient detail could be obtained from the screening procedure. Studies identified in multiple publications were included under the same study name, the source providing the richest data included in the analysis.

### **3.3.7 Data Extraction**

Data were extracted from all included studies onto a standardised electronic data extraction template. *Box 3.1* presents the extracted data items.

### *Box 3.1: Extracted Data Items*

- Study ID
- Reference
- Date of publication
- Study design
  - o For non-randomised studies, specific study design features (Higgins and Green 2010)
- Routinely recorded data source
- Standard prospective method
- Clinical focus
  - o Clinical speciality
  - o Diagnoses
- Outcome measures
  - o Clinical
    - Efficacy
    - Harm
  - o Mortality
  - o Health Economic
- Assessment of agreement
- Appraisal
  - o Limitations
    - Study reported limitations of implementing routinely recorded data
    - Appraisal, bias
  - o Benefits
    - Study reported benefits of implementing routinely recorded data
    - Appraisal, bias

### **3.3.8 Data Analysis**

We planned to perform a narrative assessment of risk of bias for all studies and formal assessment of the risk of bias for included studies when relevant to the assessment of agreement. When relevant, for RCTs the Cochrane Risk of Bias Tool and for non-Randomised Studies (NRS) the Cochrane Risk Of Bias In Non-randomized Studies - of Interventions (ROBINS-I) Tool would be completed. For NRS where completion of the ROBINS-I was not appropriate a narrative assessment of factors associated with bias and methodological quality would be performed. There was no planned routine formal assessment of reporting bias or sensitivity analyses as the aim of this review was to assess the agreement of data rather than appraise the outcomes of the included studies.

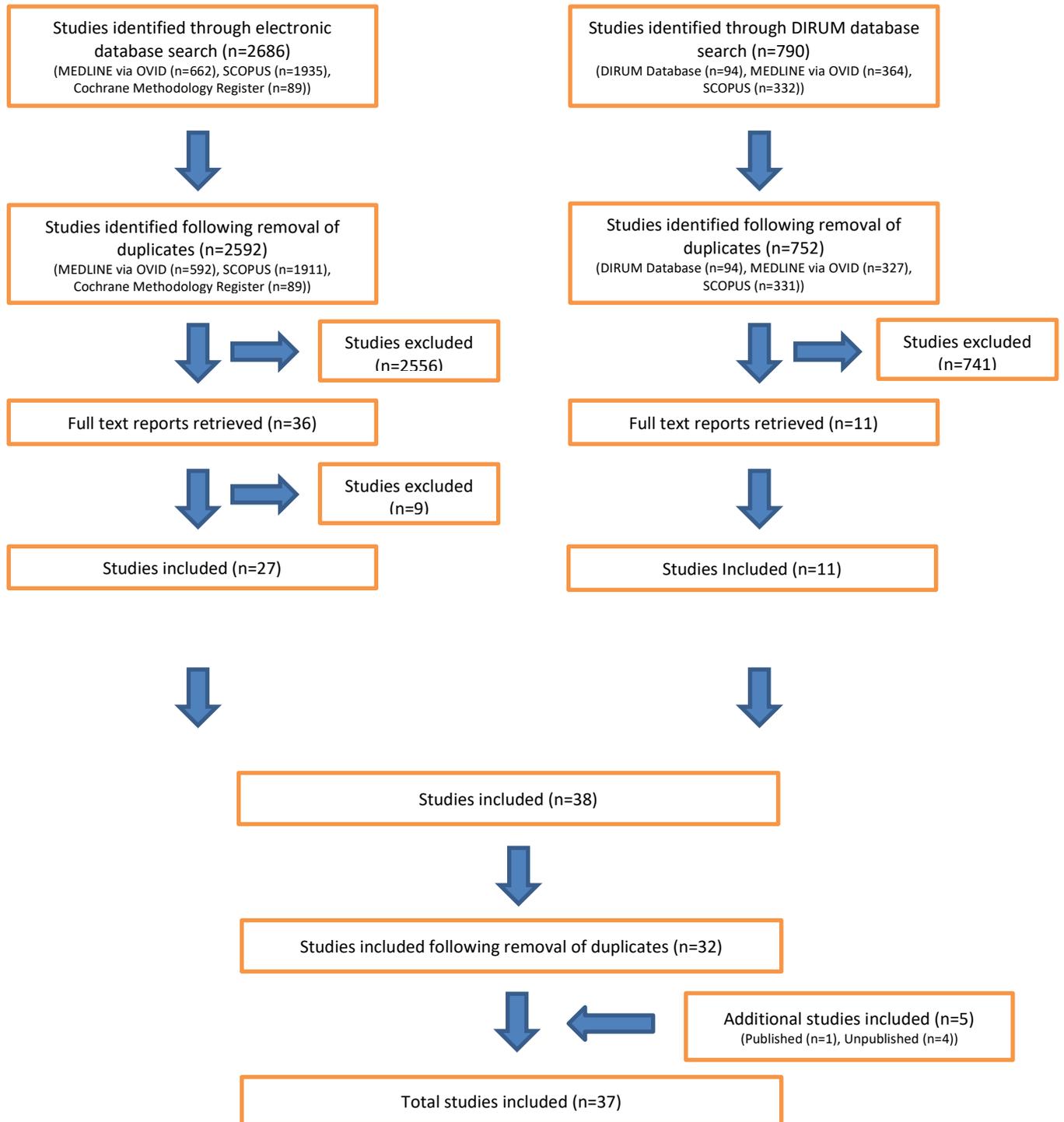
The results were analysed using simple descriptive statistics and a narrative appraisal. Further statistical manipulations including meta-analyses were not appropriate due to the heterogeneity of aims, interventions, outcomes measures and routinely recorded data sources accessed.

### **3.4 Results**

The titles and abstracts of 2592 studies identified in the electronic database searches were reviewed. The full text was reviewed for 36 studies and 27 were eligible for inclusion. Additionally, the title and abstracts of 752 studies identified in the review of the DIRUM database and subsequent update were reviewed. The full text was reviewed for 11 studies and all were eligible for inclusion. The combined results identified 38 studies. Four exact duplicates were removed [158-160, 167]. Two further studies presented similar data to other included studies and were also removed [143, 146]. The addition of four eligible studies identified in the manual search and one study identified during the development of the search strategy [149] resulted in a total of 37 eligible, included studies. The process for the identification of studies is summarised in *Figure 3.1*.

In addition to the two studies presenting similar data to other included studies [143, 146], a further nine studies were excluded from the final review, with reasoning detailed in *Table 3.4*. Five studies did not present an assessment of agreement of routinely recorded data compared to data recorded through standard prospective methods [161-164, 171], three studies did not include UK routinely recorded data [117, 165, 169] and the full text for one study was not retrievable during the review [170].

Figure 3.1: The Identification of Eligible Studies



*Table 3.4: Excluded Studies*

<b>Study Reference</b>	<b>Reason for Exclusion</b>
Andersen et al: 2015 [165]	The study presents the results of source data verification of three large multinational RCTs, only one involving a small number of patients in the UK. Excluded as no clear comparison to routinely recorded data in the UK.
Barbara et al: 2012 [169]	This study presents the agreement between self-reported diagnoses and those recorded in Canadian primary care records.
Cornish et al: 2015 [164]	No comparison of routinely recorded data to data collected through standard prospective methods. Additionally, a non-clinical data source is accessed (National Pupil Database).
Delaney et al: 2008 [161]	No comparison of routinely recorded data to data collected through standard prospective methods.
Dobbie et al: 2015 [162]	The study involves analyses of questionnaire and biochemical test data that have been prospectively recorded. The accuracy of questionnaire data is being assessed by comparison to biochemical test data, although both data have been prospectively recorded within the same data source. There is no comparison of such prospectively recorded data to alternative routinely recorded data.
Ishihara-Paul et al: 2008 [163]	No comparison of routinely recorded data to data collected through standard prospective methods.
Lewsey et al: 1999 [143]	This study is an eligible reference, however the same data is presented in an included comprehensive HTA report [46] and therefore this study has been excluded.
Mosis et al: 2006 [117]	This study involves a 'Randomised Database Study' using Dutch primary care routinely recorded data.
Mukergee: 1999 [170]	This study may be eligible but there is limited detail in the abstract and the full report was not retrievable through the available resources during this review.
O'Brien et al: 1992 [171]	No comparison of routinely recorded data to data collected through standard prospective methods.
Tannen et al: 2009 [146]	This study is an eligible reference, however the same data is presented in previously published, more comprehensive reports [144, 145, 147, 148].

Included studies were widely heterogeneous. *Tables 3.5* and *3.6* present a narrative summary of the studies including a subjective interpretation of the statistical and clinical significance of agreement between routinely recorded data and data collected through standard prospective methods. This interpretation considers if the evidence presented indicates an acceptable level of agreement to recommend the use of the assessed routinely recorded data to measure outcomes in future prospective studies such as RCTs. Extracted data is presented in greater detail in *Appendix B, Tables B.1-B.3*.

Table 3.5: Included Studies: Clinical Data - Assessment of Agreement

Reference	Data Sources	Assessment of Agreement	Acceptable Agreement?
Barry et al 2013 [65]	Standard: RCT clinical follow up Routine: Scottish Morbidity Record Secondary care medical records	Data recorded in a RCT was compared to mortality and secondary care medical records. The primary outcomes were broadly comparable. Cardiovascular death or myocardial infarction in the placebo compared to pravastatin group was 212 vs 147 (P<0.001, RR 32 (16,45)) in the RCT dataset and 195 vs 121 (P<0.001, RR 39 (24,51)) in the routine dataset.	Yes:  Cardiovascular mortality Myocardial infarction
Britton et al 2012 [178]	Standard: Questionnaire Routine: Hospital Episode Statistics Inpatient Dataset (HES IP) Primary and secondary care medical records	Patient-reported episodes of 'stroke' were compared to episodes recorded in HES, primary and secondary care medical records. 7.5% of self-reported strokes were not recorded as 'stroke' in routine sources. 62.3% of self-reported strokes were validated in HES data. 15.1% were validated by hospital records alone and were not recorded in HES. 10.4% were recorded in HES alone. 11.3% were validated by GP only.	No
Bryant et al 2015 [149]	Standard: Cohort study follow up Routine: Personal Child Health Record	There were no statistically significant differences between research measured infant height and weight and routinely recorded height and weight in the Personal Child Health Record.	Yes:  Infant height / weight
Cleland et al 2007 [150]	Standard: Questionnaire Routine: Primary care medical records	The difference between 'intervention' and 'control' groups in a RCT assessing asthma control are narratively compared for standard data and routine data. Both methods of data collection resulted in no significant difference between groups but the measures are not adequately comparable.	No
Doshi et al 2007 [152]	Standard: Questionnaire Routine: Secondary care medical records	Post-operative bleeding reported through self-report questionnaire was compared to secondary care medical records. A greater number of events were self-reported. It is possible events not included would be recorded in primary care medical records.	No
Embleton et al 2015 [183]	Standard: RCT clinical follow up Routine: Secondary care medical records	Data recorded in a RCT were compared to medical records. Of 253 'cancer progressions' recorded in the RCT, 2 additional progressions were identified in medical records. The HR of the primary outcome and its significance was unchanged following the inclusion of the 2 additional progressions.	Yes:  'Cancer progression'
Herrington et al 2015 [182]	Standard: RCT clinical follow up Routine: 'UK Mortality Registers'	Data recorded in a RCT were compared to UK mortality registers. 2778 of 2835 deaths could be verified. The RR of the primary outcome, overall mortality was in agreement. Individual causes of death were in agreement (Kappa 0.78-0.86) with the exception of ischaemic stroke (Kappa 0.19).	Yes:  All-cause mortality Cardiovascular mortality
Hutchings et al 2005 [154]	Standard: Questionnaire Routine: Primary and secondary care medical records	Assessment of medical records to substitute commonly used self-report questionnaires. Roughly two thirds of items in the SF-36 could be substituted in medical records, but only one third of relevant codes were ever used.	No
Iyer et al 2013 [155]	Standard: Questionnaire Routine: Secondary care medical records	Post-operative complications were identified through self report questionnaires and secondary care records. For serious complications, all were recorded, for minor complications, half were recorded. Minor events may be alternatively recorded in primary care records.	No
Kingston et al 2010 [153]	Standard: Questionnaire Routine: ONS Omnibus Surveys	The incidence of 'unrecorded' mammograms in the control arm of a RCT was assessed and compared to a similar cohort of women responding to the ONS Omnibus Survey. The populations and rates were comparable.	Yes:  Mammography

Lewsey et al 2000 [46]	Standard: RCT clinical follow up Routine: Scottish Morbidity Record Secondary care medical records	A RCT assessing cardiovascular treatments was replicated as far as possible using a similar cohort of patients from routine data. The RR for the primary outcome was comparable.	Yes:  Cardiovascular mortality Myocardial infarction
Mitchell et al 2016 [166]	Standard: Questionnaire Routine: Secondary care medical records	Self-reported episodes of self harm were compared with secondary care medical records. Patients under-reported both the occurrence and absence of previous episodes of self harm. Cohen Kappa demonstrated poor agreement (Kappa=0.353, CI 0.287–0.419).	No
Pastorino et al 2015 [180]	Standard: Questionnaire Routine: Primary care medical records	Self-reported diagnosis of diabetes was compared to primary care medical records. There is good agreement (94.9%) for diagnosis but patients significantly over estimated the duration of their disease by a mean 0.6 years.	Yes:  Diabetes mellitus
Smith et al 2015 [181]	Standard: Research nurse review Routine: NHS Safety Thermometer	Research nurse review of pressure ulcers compared to routinely recorded incidence. Reported 'low accuracy' with a calculated sensitivity of 48%.	No  Pressure ulcers
Steward et al 1993 [156]	Standard: RCT clinical follow up Routine: Secondary care medical records	Data recorded on CRFs in a RCT was compared to medical records. Data discrepancies occurred in 3-7.5%. 20% of data recorded in the RCT could not be verified in the medical records.	No
Tannen et al 2006 [144]	Standard: RCT clinical follow up Routine: General Practice Research Database (GPRD)	A RCT assessing treatments for hypertension was replicated as far as possible using a similar cohort of patients from routine data. Baseline characteristics were comparable. Of the 12 clinical outcomes, statistically significant differences between standard and routine data were observed in 2 outcomes.	No
Tannen et al 2007 [147]	Standard: RCT clinical follow up Routine: General Practice Research Database (GPRD)	A RCT assessing hormone replacement therapy was replicated as far as possible using a similar cohort of patients from routine data. Notable differences were observed in baseline characteristics. Of the 10 clinical outcomes, statistically significant differences between standard and routine data were observed in 4 outcomes.	No
Tannen et al 2007 [148]	Standard: RCT clinical follow up Routine: General Practice Research Database (GPRD)	A RCT assessing hormone replacement therapy was replicated as far as possible using a similar cohort of patients from routine data. Notable differences were observed in baseline characteristics. Of the 10 clinical outcomes, statistically significant differences between standard and routine data were observed in 5 outcomes.	No
Tannen et al 2008 [145]	Standard: RCT clinical follow up Routine: General Practice Research Database (GPRD)	Two RCTs assessing treatments for hypertension were replicated as far as possible using a similar cohort of patients from routine data. Notable differences were observed in baseline characteristics. Of the 10 clinical outcomes, statistically significant differences between standard and routine data were observed in 9 outcomes.	No
Tudur-Smith et al 2012 [141]	Standard: RCT clinical follow up Routine: Secondary care medical records Office for National Statistics	Data recorded in an RCT assessing chemotherapy for cancer were compared to medical records and ONS data. The primary outcome of mortality was comparable. Minor data discrepancies did not alter the result or significance of RCT outcomes.	Yes:  All-cause mortality
Weiner et al 2008 [157]	Standard: RCT clinical follow up Routine: General Practice Research Database (GPRD)	A RCT assessing treatment for hypercholesterolemia was replicated as far as possible using a similar cohort of patients from routine data. Notable differences were observed in baseline characteristics. Of the 3 clinical outcomes, a statistically significant difference between standard and routine data was observed in 1 primary outcome.	No

Williams et al 2003 [64]	Standard: Questionnaire RCT clinical follow up Routine: Primary and secondary care medical records	RCT 1: 9/13 outcomes on the health status questionnaire SF-36 could be substituted using coded terms from medical records. RCT 2, 3, 4: RCTs in different disease areas were replicated as far as possible using a similar cohort of patients from routine data. In RCT 2, out of 5 clinical outcomes, 4 could not be assessed. In RCT 3: out of 25 clinical measures, 16 could not be assessed. In RCT 4: out of 11 outcomes, 2 could not be assessed and of the remaining 9 outcomes, 2 had discrepant results.	RCT 1: No RCT 2: No RCT 3: No RCT 4: No
--------------------------	--	--	--

*Table 3.6: Included Studies: Health Economic Data - Assessment of Agreement*

Reference	Data Sources	Assessment of Agreement	Acceptable Agreement?
Breeman et al 2011 [172]	Standard: Questionnaire Routine: Medical records and 'routine data sources'	Data recorded through questionnaires for healthcare resource use were compared to medical records. Limited data were presented in the report. 15% of self-reported admissions could not be verified in medical records.	No
Byford et al 2007 [158]	Standard: Questionnaire Routine: Primary care medical records	Data recorded through questionnaires for healthcare resource use were compared to medical records. At 12 months, a mean 1.88 fewer GP appointments were self-reported. There were greater numbers of self-reported appointments for services not administered in primary care.	No
Chishti et al 2013 [174]	Standard: Questionnaire Routine: Primary care medical records	Data recorded through questionnaires for healthcare resource use were compared to medical records. At 12 months, there were significantly fewer GP appointments self-reported (mean difference 1.6, (95% CI 0.5–2.7), P = 0.004).	No
Dixon et al 2009 [151]	Standard: Questionnaire Routine: Secondary care and ambulance medical records	Data recorded through questionnaire for healthcare resource use at 28 days were compared to medical records. Costs derived for the intervention and control groups are presented. Using routinely recorded data the costs were; £3966 and £4166. For the questionnaire reported attendances were fewer; £2102 and £2641.	No
Ford et al 2007 [175]	Standard: Questionnaire Routine: Secondary care medical records	Data recorded through questionnaire for healthcare resource use were compared to medical records. At 2 years, fewer appointments were self-reported (non-significant). For appointments at the primary centre, there was good agreement (ICC = 0.77, 95% CI: 0.67–0.85).	Yes: Healthcare resource use
Hussain et al 2012 [179]	Standard: Questionnaire: Routine: Secure Anonymised Information Linkage (SAIL) Databank (Inpatient, Outpatient, Primary Care, Emergency Care)	Data recorded through questionnaires for healthcare resource use were compared to medical records accessed through SAIL. Primary care 'visits' were underreported compared to the 'events' recorded in primary care records and inpatient admissions were underreported compared to secondary care records. Outpatient visits were over-reported, the largest discrepancy seen in patients with high disease severity, reporting 2.55 vs 1.51 visits.	No

Kennedy et al 2002 [160]	Standard: Questionnaire Routine: Secondary care medical records	Data recorded through questionnaire for healthcare resource use were compared to medical records. At 12 months a mean 5.6 attendances were self-reported compared to 4.3 recorded in medical records (P=0.006). The ICC of 0.54 indicates moderate agreement.	No
Mistry et al 2005 [176]	Standard: Questionnaire Routine: Primary care medical records	Data recorded through questionnaire for healthcare resource use were compared to medical records. At 12 months, for all healthcare contacts, a mean 17.20 attendances were self-reported compared to 12.64 recorded in medical records (P=0.083). For GP appointments, there was moderate agreement (Kappa: 0.370).	No
Morrell et al 2000 [177]	Standard: Questionnaire Routine: Primary care medical records	Data recorded through questionnaire for healthcare resource use were compared to medical records. At 6 weeks, self-reported contacts were over-reported by a mean 0.5 of a contact (95% CI, 0.2, 0.7). At 6 months, self-reported contacts were underreported by a mean – 0.1 contacts (95% CI, –0.7, 0.5).	Yes:  Healthcare resource use
Petrou et al 2002 [159]	Standard: Questionnaire: Routine: Primary and secondary care medical records	Data recorded through questionnaire for healthcare resource use were compared to medical records. At 4 months 29.4% of patients recorded primary care visits in agreement with medical records with 56.9% underreporting (At 12 months; 28% and 58%) (P<0.001). There were no significant differences in secondary care attendances.	Yes:  Secondary Care: Healthcare resource use
Richards et al 2003 [167]	Standard: Questionnaire: Routine: Primary, secondary and community medical records	Data recorded through questionnaire for healthcare resource use were compared to medical records. For the majority of variables, there was significant under reporting of self-reported attendances. However, for certain measures there was good agreement: 'Hospital Readmission' (Kappa: 0.68) and GP surgery visits (Kappa: 0.60).	No
Shaw et al 1998 [168]	Standard: Clinician report Routine: Secondary care medical records	There was agreement between clinician report and medical records in 118/140 (84.3%) cases for grade of doctor seen and in 105/139 (76.7%) cases for the management decision following outpatient appointments.	No
Thorn et al 2016 [142]	Standard: RCT clinical follow up, primary care medical records Routine: Hospital Episode Statistics Outpatient Dataset (HES OP)	Data recorded prospectively during a RCT were compared with HES OP data. 4088 of the total 4922 appointments recorded in the RCT were identified in HES OP (83.1 %). 215/370 men (58.1 %) had at least one appointment that was unmatched in HES OP.	No
Thorn et al 2016 [66]	Standard: RCT clinical follow up, primary care medical records Routine: Hospital Episode Statistics Inpatient Dataset (HES)	Healthcare costs derived from data recorded prospectively during a RCT were compared with costs derived using HES data. Costs associated with HES data were 8% lower (P=0.3). 11 men (3.8%) for whom events were recorded in HES had all these events missing from RCT data and 7 men (2.4%) with no events according to HES had events identified in RCT data.	Yes:  Healthcare costs
Wright-Hughes et al 2013 [184]	Standard: RCT clinical follow up, secondary care medical records Routine: Hospital Episode Statistics Inpatient Dataset (HES)	Data recorded prospectively during a RCT were compared to HES data. Limited data presented in report. Narrative discussion reports comprehensive and timely outcome data obtained from NHS Digital, but ambiguity in the agreement of the outcome measures.	No

### *Clinical Data*

Twenty two studies assessed the agreement of clinical routinely recorded data with clinical data collected using standard prospective methods. Comparisons included individual –level, paired data and matched cohort level data. The assessment of agreement was heterogeneous, with methods including calculation of statistically significant differences between outcomes and formal statistical assessments of agreement. Comparisons ranged from study outcomes measured using routinely recorded data and data collected using standard prospective methods to comparisons of individual data items.

Mortality data from UK mortality registers resulted in the most rigorous evidence for an acceptable level of agreement compared to data collected using standard prospective methods. In Tudur-Smith et al [141] the outcome ‘overall mortality’ calculated using RCT data (HR: 1.18 (95% CI: 0.99 to 1.42), secondary care medical records (1.18 (0.99 to 1.41) and ONS data (1.18 (0.99 to 1.40) were almost identical. Herrington et al [182] (RCT data (RR: 0.83, 95% CI: 0.75-0.91), Registry data (RR: 0.81, 95% CI: 0.74-0.90)) and Barry et al [65] (RCT data (RR 32 95% CI: 16-45), Registry data (RR 39 95% CI: 24-51)) also found comparable outcomes for mortality derived from RCT data and UK mortality registers.

Four studies accessed primary care medical records for recruited individuals. Pastorino et al [180] found an acceptable level of agreement between self-reported diagnosis of diabetes and diagnosis recorded in primary care medical records, with agreement in 94.9% of cases. However, for a similar ‘common’ diagnosis, Britton et al [178] found that 11.3% of self-reported episodes of stroke had no record in either primary or secondary care medical records. In Cleland et al [150], there was no significant difference found in ‘asthma severity’ between intervention and control groups in a RCT, with severity calculated using self-reported questionnaire data or prescribing data from primary care medical records, although the measures are not directly comparable. Hutchings et al [154] assessed the potential for primary care medical records to substitute commonly used self-reported health status questionnaires. Although relevant codes could be identified for the majority of the items within the questionnaire, in practice only one third of eligible codes were ever used in primary care medical records. A further five studies accessed primary care data through the GPRD. ‘Replicated RCTs’ were performed with a cohort of patients matched as far as possible to the RCT cohort. Although agreement was noted in a small number of clinical measures and outcomes, considering all of the data, none of these studies produced evidence for an acceptable level of agreement.

For example, in Tannen et al [145], statistically significant differences were observed in nine of the total 10 clinical outcomes for the two replicated RCTs. However, the level of agreement improved in all such replicated RCTs involving GPRD data following statistical manipulation to address unmeasured confounding, such as prior exposure to the intervention treatment under study.

Nine studies assessed agreement with data recorded in secondary care medical records, three of which found acceptable agreement. Embleton et al [183] recorded 253 'cancer progressions' using standard prospective methods. Two additional progressions were recorded in secondary care medical records. The HR of 0.57 was unchanged and the CI change was minor; 0.45-0.74 to 0.44-0.73. Barry et al [65] recorded cardiovascular death or myocardial infarction between placebo and intervention groups identifying 212 vs 147 events ( $P < 0.001$ , RR 32 (16,45)). The outcome calculated using secondary care medical records was comparable ( $P < 0.001$ , RR 39 (24,51)). Lewsey et al [46] recorded myocardial infarction or cardiac death in a meta-analysis of eight RCTs for two treatment interventions. The Relative Risk was 1.03 (95% CI 0.84 to 1.27) and was comparable to the RR calculated using routinely recorded data from secondary care and mortality records (1.15 (95% CI 0.90 to 1.48)). Two studies [155, 156] assessed the agreement of self-reported post-operative complication rates to data recorded in secondary care medical records. In both studies a significantly greater number of events were self-reported and it is likely that patients either did not seek medical advice or contacted their General Practitioner when experiencing minor complications. In Mitchell et al [166] the opposite was found, with patients tending to under-report previous episodes of self-harm. Finally, of the additional routine data sources accessed, acceptable agreement was reported in Bryant et al [149] for infant height and weight recorded in the Personal Child Health Record, compared to measurements completed during research.

#### *Health Economic Data*

Healthcare resource use was self-reported in 15 studies using validated questionnaires and compared to primary or secondary care medical records with acceptable agreement found in just four studies. Ford et al [175] assessed parent-reported healthcare use for 87 children attending secondary care mental health services. The mean parent-reported number of total appointments was 5.6 compared to 7.0 retrieved from medical records ( $P = 0.1$ ) with acceptable agreement (ICC = 0.77, 95% CI: 0.67–0.85).

Morrel et al [177] assessed self-reported healthcare use in 623 patients compared to primary care medical records. At six months self-reported contacts were marginally under-reported, the mean difference was -0.1 contacts (95% CI, -0.7, 0.5). Petrou et al [159] identified '90-100% agreement' between self-reported healthcare attendances and attendances identified in secondary care medical records for women attending postpartum services. Two studies assessed agreement of the HES Inpatient Dataset with data collected during a RCT. Thorn et al [66] calculated healthcare resource use costs using data collected during an RCT (£11 122 (95% CI £9083 to £13 161)) and HES data (£10 223 (95% CI £8880 to £11 565)). Costs calculated with HES data were 8% lower but the difference was not significant (P=0.3). Wright-Hughes et al [184] did not present data, but altered their follow up method to include HES data to identify healthcare resource use in an on-going RCT, indicating their confidence in the level of agreement. For the studies without acceptable agreement, self-reported attendances were found to be both under and over-reported compared to medical records, with no distinguishing pattern.

### **3.5 Discussion**

Routinely recorded data may provide an efficient alternative method for data collection in prospective research and inform outcomes beyond the standard RCT assessments of clinical efficacy and effectiveness. However, limitations with the accuracy of coding, confidentiality, ownership and access have been identified as significant barriers to accessing routinely recorded data for prospective research [58]. This systematic review assessed the agreement of routinely recorded data in the UK compared to data collected using standard prospective methods. A lack of evidence was identified, with 'all-cause mortality' being the only variable with a level of agreement sufficient to recommend use in outcome measurement in RCTs.

Studies were heterogeneous in terms of the routinely recorded data sources accessed, outcomes assessed and methods used to assess agreement. Resultantly, further statistical manipulations of data including meta-analyses were inappropriate. Alternatively, a narrative approach was taken to summarise the assessments of agreement and propose an interpretation to determine if the evidence presented indicated a level of agreement that was acceptable and sufficient to endorse substituting data collected using standard prospective methods, with routinely recorded data in future prospective research.

For clinical outcomes, mortality data from UK Mortality Registers, such as the ONS provided the most rigorous evidence for an acceptable level of agreement. The occurrence of death, or 'all-cause mortality' was in agreement with data collected using standard prospective methods. However, cause of death, which relies on coded data recorded on the medical death certificate, was found to be in poor agreement in some studies. This is consistent with the general pattern of poor agreement for clinical outcomes from primary and secondary care medical records, with few studies providing evidence for acceptable agreement. In one example where rigorous agreement was found [183], the outcome of 'cancer progression' was broad and it is likely that a host of investigations, specialist reviews and treatments had occurred in a secondary care setting, providing ample opportunity for events indicating 'cancer progression' to be recorded in the routine medical records.

The poor agreement of data in either primary or secondary care in studies included in this review may be explained by the clinical event occurring in an alternative care setting. For example, post-operative minor complication rates were self-reported in greater numbers than recorded in secondary care medical records [155, 156], the likely explanation is that patients were attending their GP rather than re-attending the hospital for treatment. However, although self-reported diagnosis of diabetes was in good agreement with primary care medical records [180], self-reported diagnosis of stroke was in poor agreement with either primary or secondary care medical records [178], indicating that even for common or chronic diagnoses where evidence in medical records would be expected, poor agreement has been observed. Alternative explanations for the poor pattern of agreement noted for clinical outcomes include the possible limited accuracy of coded data [181] and the potential underutilisation of clinical codes within routinely recorded data sources [154].

Further to the discussed assessments of agreement for individual-level, paired data, the GPRD was used to perform a number of 'replicated RCTs'. Using unadjusted data, the replicated RCTs performed poorly in the assessment of agreement to the comparator RCT. The studies discuss the composite influence of unmeasured confounding as the likely explanation. For example, patients in GPRD fulfilling the inclusion criteria may have been exposed to the treatment under investigation before the commencement of the study period. A statistical technique to address unmeasured confounding was developed and the agreement between outcomes improved, although did not reach 'acceptable' levels [145].

Healthcare resource use also has a general pattern of poor agreement. Of the 15 included studies comparing routinely recorded data with self-reported healthcare resource use, four had an acceptable level of agreement compared to primary and secondary care medical records and the HES inpatient dataset. Patients both over and under-reported healthcare attendances with no distinguishing pattern evident. Recall bias is a possible explanatory factor. However, Ford et al [175] reported agreement with self-reported attendances over two years, whereas Dixon et al [151] concluded the opposite, with significantly different healthcare costs calculated from self-reported attendances recalled over 28 days compared to medical records.

Healthcare attendances are perhaps more likely to be accurately recorded as a result of the nature of the method of remuneration in the UK National Health Service. In this context, the self-reported attendances are likely to be of reduced accuracy and have resulted in the general pattern of poor agreement. In contrast, although healthcare attendances are likely to be accurately recorded in medical records, the coded clinical details are likely to be less accurate [185] compared to standard prospective methods, such as research nurse appointments, patient recall of clinical details and research completion of CRFs. Although the issue of 'accuracy' of routinely recorded data remains limited, this review specifically assessed the agreement compared to data collected using standard prospective methods and similarly, there remains limited evidence.

### **3.6 Limitations**

The nature of the objective introduces limitations into the systematic review. The review does not focus on clinical diagnosis, intervention or outcome but rather on methodology. The alternative focus on routinely recorded data resulted in the search strategy being inclusive of all UK prospective studies but with a purposive search for specified data sources and relevant generic terms filtered for the UK. There is therefore a risk that studies poorly indexed or poorly documenting the data source or methodology may not have been identified. Additionally, although the majority of included studies are assessing 'agreement' as their primary objective, there remains a possibility that studies including relevant information have been omitted. For example, a RCT report may include relevant information regarding the agreement between data sources, but this may not be the primary focus of the report. To address this limitation as far as feasible and possible, a search of other relevant resources was completed. This included a review of the DIRUM database and an adaptation and extension of the electronic database searches in addition to a manual search of relevant resources.

Finally, the heterogeneity of the included studies did not permit any further synthesis of data and results have been narratively presented. Such a narrative presentation and interpretation is associated with a degree of subjectivity, particularly in the interpretation of 'acceptable' agreement.

### **3.7 Conclusions**

In this chapter, the agreement between routinely recorded data and data collected using standard prospective methods was assessed in a systematic review. Routinely recorded data has a generally poor pattern of agreement for both clinical data and healthcare economic data. In general, the level of agreement identified in this review is not currently sufficient to recommend use in place of data collected using standard methods in prospective studies including RCTs, with the exception of 'all-cause mortality'. However, there are notable limitations with this systematic review and the lack of evidence available, for example no outcomes relevant to the treatment of epilepsy were identified and further research for outcomes relevant to this condition amongst others is needed.

In the following chapters, the accessibility, feasibility and agreement between routinely recorded data compared to data collected using standard prospective methods for participants enrolled in SANAD II will be assessed.



## **Chapter Four**

### **The Identification and Accessibility of UK Routinely Recorded**

#### **Data Sources**

##### **4.1 Introduction**

In Chapter One, routinely recorded data and data sources in the UK were introduced, including the potential of routinely recorded data for use in clinical research. Epilepsy was introduced and the SANAD II RCT presented. In Chapter Two the use of routinely recorded data in randomised controlled trials in the UK was reviewed and in Chapter Three the agreement of UK routinely recorded data compared to data collected using standard methods in prospective studies was assessed in a systematic review.

In this chapter, sources in the UK routinely recording data relevant to the outcomes of SANAD II will be reviewed. Subsequently, sources where routinely recorded data are accessible for individuals recruited into SANAD II are presented. In the following chapter, the methods for the comparison of the attributes of routinely recorded data retrieved from the accessible sources to data collected using standard prospective methods in SANAD II are presented.

##### **4.2 Routinely Recorded Data in the UK Relevant to SANAD II**

###### ***4.2.1 Introduction***

Routinely recorded data that are potentially relevant to SANAD II are recorded by a number of organisations in the UK. Sources routinely recording data that could potentially be used to directly measure the outcomes of SANAD II or contribute additional relevant data have previously been introduced in *Chapter One, Section 1.4* and *Tables 1.1 and 1.2*, including a discussion of use in clinical research. In this section, the sources, data and datasets recorded within each source together with the procedures for requesting access to data for research, where they exist, will be reviewed. This review, together with the assessment of feasibility and efficiency presented in Chapter Eight has been published [186].

## **4.2.2 Secondary Care Clinical Routine Data Sources**

### **4.2.2.1 NHS Digital [55]**

NHS Digital is an executive non-departmental public body sponsored by The Department of Health and the national provider of information, data and IT systems for commissioners, analysts and clinicians in health and social care in England. NHS Digital records, analyses and presents English health and social care data [55].

NHS Digital Hospital Episode Statistics (HES) record details of all inpatient admissions (1989- ), outpatient appointments (2003- ) and A&E attendances (2007- ) at NHS hospitals in England. Data is routinely recorded in Patient Administration Systems (PAS) in all NHS Trusts in England, submitted to the Secondary Uses Service (SUS) for the primary purpose of re-imburement and subsequently re-purposed as HES. HES includes a number of relevant datasets:

- Accident and Emergency Dataset
- Admitted Patient Care Dataset
- Outpatient Dataset
- Adult Critical Care Dataset
- Maternity Care
- Patient Reported Outcome Measures

HES data is published in anonymised aggregate reports annually. Bespoke individual-level datasets for use in clinical research are available through application to the Data Access and Request Service (DARS). Following approval by the Research Ethics Service and Health Research Authority, an application is submitted accompanied with supporting documentation (study protocol, patient information leaflets, approval documentation). A key criterion for approval is the demonstration that a study can directly (or indirectly) contribute to the improvement of the health and social care system in England.

Furthermore, for individual-level identifiable data, there must be a valid legal basis for data release in place, such as participant consent. The Data Access Advisory Group (DAAG) assesses the application and will make a recommendation regarding approval. Following approval, payment will be required, a Data Sharing Agreement will need to be signed on a study level and a Data Sharing Framework Agreement will need to be signed on an institutional level before data is processed and securely transferred to the institution.

Notably, during the course of this research, the application and approval processes for access to data have evolved, together with the name of the organisation (formerly The Health and Social Care Information Centre). An online application portal now exists and the approval process has been revised with studies now reviewed by the Independent Group Advising on the Release of Data (IGARD).

#### *4.2.2.2 NHS Wales Informatics Service [86]*

The NHS Wales Informatics Service (NWIS) is an executive non-departmental public body sponsored by The National Assembly for Wales and is the national provider of digital services for commissioners, analysts and clinicians in health and social care in Wales [86].

The NWIS Information Services Division is responsible for the collection, management, and analysis of data held in a number of national databases, and the production and distribution of information derived from these databases [86]. NWIS record details of all inpatient admissions (1991- ), outpatient appointments (2003- ) and A&E attendances (2009- ) at NHS hospitals in Wales. Data is routinely recorded in Patient Administration Systems (PAS) in all NHS Trusts in Wales, submitted to NWIS for the primary purpose of reimbursement and subsequently re-purposed in a number of relevant datasets:

- Emergency Department Dataset
- Patient Episode Database for Wales
- Outpatient Dataset
- Critical Care Dataset

NWIS data is published in anonymised aggregate reports annually. Bespoke individual-level datasets for use in clinical research are available through application to the NWIS Bespoke Analysis Service or Public Health Wales Observatory. Following approval by the Research Ethics Service and Health Research Authority, an application is submitted accompanied with supporting documentation (study protocol, patient information leaflets, approval documentation). For individual-level identifiable data, there must be a valid legal basis for data release such as participant consent.

#### 4.2.2.3 NHS National Services Scotland: Information Services Division (ISD) [88]

The Information Services Division (ISD) is a division of National Services Scotland, part of NHS Scotland. ISD provides health information, health intelligence, statistical services and advice that support the NHS in progressing quality improvement in health and care and facilitates robust planning and decision making [88].

The Information Services Division is responsible for the collection, management, and analysis of data held in a number of administrative databases. ISD records details of all inpatient admissions, outpatient appointments and A&E attendances at NHS hospitals in Scotland. Data is routinely recorded in Patient Administration Systems (PAS) in all NHS Trusts in Scotland, submitted to ISD and re-purposed. A variety of relevant data is recorded:

- Emergency Care
- Inpatient Care
- Outpatient Care
- Critical Care
- Primary Care Out-Of-Hours / Unscheduled Care
- Prescribing Data, The Prescribing Information System

Bespoke individual-level datasets for use in clinical research are available through application to the electronic Data Research and Innovation Service (eDRIS), which operates the Information Request Service. Following approval by the Research Ethics Service and Health Research Authority, an application is submitted accompanied with supporting documentation (study protocol, patient information leaflets, approval documentation). For individual-level identifiable data, there must be a valid legal basis for data release such as participant consent.

### **4.2.3 Primary Care Clinical Routine Data Sources**

#### *4.2.3.1 The Clinical Practice Research Datalink (CPRD) [90]*

CPRD is a governmental research service jointly funded by the NHS National Institute for Health Research (NIHR) and the Medicines and Healthcare products Regulatory Agency (MHRA) aiming to provide anonymised individual-level data to inform clinical research. The primary care records of 8.5% (12.6 million) of the UK population are included, distributed geographically [90].

In addition to the primary care data, CPRD aims to provide linked individual-level data across a number of sources including primary and secondary care databases, disease registries, demographic and socioeconomic datasets. Linked data is provided through the Trusted Third Party, NHS Digital.

- Primary Care Data
  - o General Practice's enrolled with CPRD and using a compatible IT system
- Secondary Care Data
  - o NHS Digital HES
- Registry Data
  - o Death data, Office for National Statistics (ONS)
  - o Disease registries
- Socioeconomic Data:
  - o Lower Layer Super Output Area (LSOA) Level, ONS

Bespoke individual-level datasets for use in clinical research are available through application to the Independent Scientific Advisory Committee (ISAC). Following approval by the Research Ethics Service and Health Research Authority, an application is submitted accompanied with supporting documentation (study protocol, patient information leaflets, approval documentation).

#### *4.2.3.2 ResearchOne [91]*

ResearchOne is a collaboration between the University of Leeds and The Phoenix Partnership (TTP), developers of the SystmOne clinical database and IT system. ResearchOne provides de-identified individual-level data to inform clinical research. The primary care records of 26 million patients in the UK are included retrieved from General Practices' enrolled with ResearchOne and using the SystmOne clinical database. In addition, health and socioeconomic data may in some circumstances be linked such as data retrieved from palliative care settings and secondary care settings including emergency and inpatient care [91]. Linked data is provided through the Trusted Third Party, NHS Digital.

Bespoke individual-level datasets for use in clinical research can be considered on application. Following approval by the Research Ethics Service and Health Research Authority, an Expression of Interest Form is completed. If feasible, an application would subsequently include the supporting documentation (study protocol, patient information leaflets, approval documentation) and be considered by the ResearchOne Project Committee.

#### *4.2.3.3 QResearch [93]*

QResearch is a collaboration between the University of Nottingham and the developers of the EMIS IT systems. QResearch provides de-identified individual-level data to inform clinical research. The primary care records of 24 million patients in the UK are included retrieved from General Practices' enrolled with QResearch and using the EMIS IT system [93]. In addition, aggregate socioeconomic data at LSOA level can be provided.

Bespoke individual-level datasets for use in clinical research can be considered on application. Following approval by the Research Ethics Service and Health Research Authority, discussion with QResearch is required to determine feasibility. A QResearch Application Form is then completed and considered by the QResearch Scientific Committee, together with the supporting documentation (study protocol, patient information leaflets, approval documentation).

#### *4.2.3.4 The Health Improvement Network (THIN) Database [95]*

THIN is a collaboration between IMS Health and In Practice Systems, developers of the IT software Vision. THIN provides de-identified individual-level data to inform clinical research. The primary care records of 11 million patients in the UK are included retrieved from General Practices' enrolled with THIN and using the Vision IT system [95]. In addition, aggregate socioeconomic data at LSOA level can be provided. Data are validated by the Trusted Third Party, CSD Medical Research UK.

Bespoke individual-level datasets for use in clinical research can be considered on application. Following approval by the Research Ethics Service and Health Research Authority, discussion with THIN is required to determine feasibility. The study protocol together with supporting documentation (study protocol, patient information leaflets, approval documentation) is then submitted for consideration by the THIN Scientific Review Committee.

#### 4.2.3.5 North West eHealth (NWEH) [97]

NWEH is a not-for-profit collaboration between The University of Manchester, Salford Royal Foundation Trust and Salford Clinical Commissioning Group, aiming to develop research in health informatics and improve the links between academic institutions and the NHS [69, 97]. NWEH has an established infrastructure for accessing linked primary and secondary care data and prescribing data. The *Salford Integrated Record (SIR)* is an integrated electronic medical record recording all primary and secondary care attendances in Salford. NWEH have developed the methodology and governance framework to implement the SIR in research to provide primary and secondary care data and additional data such as details regarding pharmacy dispensing. The Salford Lung Study is an ongoing example of the implementation of this system in research [69]. This *Linked Database System* involves access to secondary care data recorded by the Secondary Uses Service (SUS) and access to primary care data for consenting participants and General Practices' using the Apollo [98] and Graphnet [99] data extraction tools, both used routinely for clinical audit and Quality and Outcome Framework (QOF) assessment. For the wider research community, NWEH provides services including the COCPIT and FARSITE initiatives to provide the facility to monitor individual patients and perform RCT recruitment feasibility assessments, respectively. NWEH do not currently routinely provide a bespoke data extraction service for research. However, the methodology and governance framework are in place and proof-of concept has been demonstrated [69].

#### **4.2.4 'Linked' Routine Data Sources**

##### *4.2.4.1 The Secure Anonymised Information Linkage (SAIL) Databank [100]*

The Secure Anonymised Information Linkage Databank is an initiative developed by the College of Medicine, Swansea University and receives core-funding from the National Institute of Social Care and Health Research (NISCHR) of the Welsh Government. SAIL aims to provide and improve access to electronic routinely-collected, anonymised individual-level data to support clinical research [100].

The SAIL Primary Care GP Dataset provides primary care data for individuals registered to enrolled General Practices'. However, SAIL also provides the infrastructure to provide de-identified linked individual-level data from a number of other clinical and socioeconomic datasets. Linked data is provided through the Trusted Third Party, NWIS. Relevant datasets include:

- Primary Care GP Dataset (SAIL)
- Patient Episode Database for Wales (PEDW, NWIS)
- Outpatient Dataset (NWIS)
- Emergency Department Dataset (EDDS, NWIS)
- Critical Care Dataset (NWIS)
- Annual District Death Extract (NWIS)
- Welsh Demographics Service (NWIS)

Bespoke individual-level datasets for use in clinical research are available through application to the Information Governance Review Panel (IGRP). Following approval by the Research Ethics Service and Health Research Authority, an application is discussed with SAIL and a Scoping Document completed. Once the feasibility of the project, including funding has been confirmed the application is reviewed by the IGRP, together with supporting documentation (study protocol, patient information leaflets, approval documentation). Access to data is usually remotely, through the SAIL Gateway Platform.

#### 4.2.4.2 The Administrative Data Research Network (ADRN) [102]

The Administrative Data Research Network is a UK-wide partnership between universities, government departments, national statistics authorities, funders and researchers, funded by the *Economic and Social Research Council*. ADRN provides a method of access to a number of non-clinical administrative routine datasets including employment, socioeconomic, crime and education data [102] in addition to clinical datasets detailed previously such as those recorded by NHS Digital.

Notably, the ADRN does not record data, but rather provides a bespoke service for researchers to negotiate for data access with each relevant data source. The aim is to provide a linked de-identified dataset to the researcher. The linking service is provided through a Trusted Third Party, such as NHS Digital. An important caveat is that at least a proportion of the data to be sought is not available through another route of access.

Examples of potentially relevant data include:

- Primary and Secondary Care Clinical Datasets (NHS Digital, NWIS, ISD)
- Economic Datasets
- Employment Datasets
- Population Datasets, Deaths

Bespoke individual-level datasets for use in clinical research are available through application to the ADRN Approvals Panel. Following approval by the Research Ethics Service and Health Research Authority, an application is discussed with ADRN and feasibility confirmed. The application is reviewed by the Approvals Panel, together with supporting documentation (study protocol, patient information leaflets, approval documentation). Access to data is usually through a dedicated Administrative Data Research Centre, for England this is based at University College London.

#### **4.2.5 Non-Clinical Routine Data Sources**

##### *4.2.5.1 The Office for National Statistics (ONS) [107]*

The Office for National Statistics is the recognised national statistics institute and the largest independent producer of official statistics. ONS are responsible for recording and publishing statistics related to the economy, population and society at national, regional and local levels, in addition to conducting the census in England and Wales [107].

ONS records a limited amount of identifiable data, such as births and deaths data. However, the majority of statistics produced by ONS are non-identifiable, aggregate statistics recorded through a number of channels such as the Census and General Household Survey. Examples of potentially relevant data include:

- Population, demography and migration statistics
- Labour market Statistics (employment, unemployment and earnings)
- Vital events statistics (births, marriages and deaths)
- Social statistics (regarding neighbourhoods, families, crime)
- Economic, societal and personal well-being

Aggregate statistics are in the public domain and are available on a population level or presented in a Super Output Area (SOA). Clinical sources of routine data such as NHS Digital and NWIS record LSOA reference numbers allowing individual-level clinical data to be linked to LSOA level socio-economic data. Aggregate statistics can be accessed via services provided by ONS such as NOMIS [108] and Data for Neighbourhoods and Regeneration [109]. Notably, such data is non-identifiable and is in the public domain. Specific requests for data not published in official statistical outputs are permitted and contact via email is suggested in the first instance. Finally, individual data regarding deaths can be requested and projects are reviewed by the Microdata Release Panel. Such projects must be in the public interest and the researcher must attain 'Approved Researcher Status'.

#### 4.2.5.2 HM Revenue and Customs (HMRC) [103]

HM Revenue and Customs is the national tax authority and was commissioned and established in 2005 by the Commissioners for Revenue and Customs Act, replacing the Inland Revenue and Customs and Excise. HMRC is a non-ministerial department responsible for tax policy maintenance and implementation, strategic tax policy and policy development [103].

HMRC is responsible for taxation including income tax, national insurance and student loan repayments and the administration of tax credits, child benefit and statutory sick and maternity pay. HMRC records individual-level data and examples of potentially relevant data include:

- Employment Data
  - o Income
  - o Tax contributions
- Statutory Benefit Data
  - o Tax credits
  - o Statutory sick and maternity pay

Data recorded by HMRC are likely to be accurate, complete and informative to health and socioeconomic analyses, including an assessment of the broader, societal impacts of treatments.

HMRC are actively involved in external research with academic institutions. The 'HMRC Datalab' provides access to de-identified HMRC data for approved research. Projects must benefit both 'HMRC and the wider academic community'. The Datalab is governed by the Commissioners for Revenue and Customs Act (CRCA) 2005 and research must serve one of HMRC's functions under the CRCA 2005. Such 'functions' include procedures or responsibilities of the HMRC. A Project Proposal Form is completed and applications are reviewed quarterly. Approved Researcher Status must be obtained before access to data is granted.

#### 4.2.5.3 *The Department for Work and Pensions (DWP) [104]*

The Department for Work and Pensions is a ministerial department responsible for welfare, pensions and child maintenance and provides its services through a number of outlets including Job Centre Plus, The Pension Service, Child Support Service and The Child Maintenance Service [104].

DWP is responsible for the provision of the state pension and provision of benefits; individually assessed non-taxable monetary credits provided to support reasonable living costs. Benefits are numerous, frequently assessed and changes in circumstances result in frequent alterations to individuals' eligibility. DWP records individual-level data and examples of potentially relevant data include:

- Employment Data
  - o Income
  - o Tax contributions
- Benefit Data
  - o Attendance Allowance
  - o Carer's Allowance
  - o Child Benefit
  - o Disability Living Allowance
  - o Jobseekers Allowance
  - o Universal Credit
  - o State Pension

Data recorded by DWP are likely to be accurate, complete and informative to health and socioeconomic analyses, including an assessment of the broader, societal impacts of treatments.

The DWP are actively involved in external research with academic institutions. De-identified, aggregate data at LSOA level are available via services provided by DWP.

#### 4.2.5.4 The Driver and Vehicle Licensing Authority (DVLA) [18]

The DVLA is an executive governmental agency sponsored by the Department for Transport and responsible for maintaining an active register of drivers and vehicles in the UK, improving road safety, supporting vehicle tax, reducing vehicle related crime, and supporting environmental initiatives [18].

The DVLA is responsible for the licensing of drivers and vehicles in the UK and issuing, reviewing and maintaining guidance regarding driver license status in the context of medical diagnoses. There is a legal requirement for the individual to inform the DVLA if they have been diagnosed with specific diseases, including seizures and epilepsy. DVLA records individual-level data and examples of potentially relevant data include:

- Driving License Status
  - o Dates issued, withdrawn, re-issued
- Diagnoses
  - o Diagnosis
  - o Date
  - o Medical data supplied by clinician

The DVLA have been involved in external research with academic institutions. De-identified, aggregate data are available on application and initial discussion with the DVLA is advised in the first instance.

### **4.3 The Accessibility of Routinely Recorded Data Relevant to SANAD II**

#### ***4.3.1 Accessibility***

An objective of this research is to assess the attributes of routinely recorded data compared to data collected using standard prospective methods for individuals recruited into SANAD II. During the development of the study protocol it was necessary to identify data sources where routinely recorded data could be retrieved for the specific individuals recruited into SANAD II. This identification of accessible sources of routinely recorded data was required to inform the protocol and consent documentation. As a consequence of strict ethical and governance regulations, both nationally and within each source of routinely recorded data, a 'blanket' consent procedure requesting access to all potentially relevant routinely recorded data was not appropriate or permissible. Consent documentation needed to meet the requirements of the Research Ethics Service, Health Research Authority and each accessible source of routinely recorded data and therefore included highly specific information relevant to each organisation.

For the purposes of this research the assessment of accessibility included the following criteria:

- Access to routinely recorded data for research, including clinical research is possible
- Individual-level data for specific individuals recruited into SANAD II can be retrieved

Following a prolonged period of scoping discussions and development of the protocol and consent documentation, sources of routinely recorded data in the UK accessible for this study were identified.

### **4.3.2 Accessible Routinely Recorded Data Sources**

#### *4.3.2.1 NHS Digital [55]*

The Hospital Episode Statistics datasets provide emergency, inpatient, outpatient and critical care clinical and socioeconomic data for participants of SANAD II resident in England.

#### *4.3.2.2 The Secure Anonymised Information Linkage Databank [100]*

Routinely recorded primary care, emergency, inpatient, outpatient and critical care clinical and socioeconomic datasets provided through SAIL for participants of SANAD II resident in Wales.

#### *4.3.2.3 NHS National Services Scotland: Information Services Division [88]*

Routinely recorded emergency, inpatient, outpatient and critical care clinical and socioeconomic data provided through ISD for participants of SANAD II resident in Scotland.

#### *4.3.2.4 The Office For National Statistics [107]*

Routinely recorded mortality data is accessible for individuals recruited into SANAD II.

#### *4.3.2.5 North West eHealth [97]*

Routinely recorded primary and secondary care clinical, prescribing and socioeconomic data provided through NWEH for participants of SANAD II resident in the North West of England.

### ***4.3.3 Non-Accessible Routinely Recorded Data Sources***

A notable proportion of the available sources of routinely recorded data were non-accessible for individuals recruited into SANAD II. Non-accessible sources are listed in this chapter and discussed further in the assessment of feasibility, Chapter Eight.

#### ***4.3.3.1 Primary Care Clinical Routine Data Sources***

With the exception of North West eHealth, the remaining primary care sources were not accessible for individuals recruited into SANAD II. The Clinical Practice Research Datalink, ResearchOne, QResearch and The Health Improvement Network all operate on a de-identified basis, resulting in an inability to re-identify specific individuals. Therefore, participants recruited into SANAD II cannot be re-identified within each database.

#### ***4.3.3.2 Non-Clinical Routine Data Sources***

HM Revenue and Customs, The Department for Work and Pensions and The Driver and Vehicle Licensing Authority were not accessible. Enquiries for HMRC and DWP were directed to the ADRN, whose negotiations for data access were not successful. The DVLA refused permission for data access, citing 'insufficient resources' together with 'security measures in excess of the NHS or University' as the explanation. This outcome correlates with the findings in Chapters Two and Three, where no clinical studies accessing individual-level data from these data sources were identified.

#### **4.4 Conclusions**

In this chapter the accessibility of UK routinely recorded data for participants of SANAD II was assessed. Routinely recorded secondary care data and mortality data were accessible. Primary care data from the majority of sources were not accessible as a result of the de-identified record of the data, resulting in an inability to identify the specific individuals recruited into SANAD II. Non-clinical data sources were not accessible, the ADRN being unsuccessful in negotiating access to HMRC and DWP data and the DVLA citing insufficient resources and stringent security measures. Streamlining the ADRN data application and access procedures and including the DVLA on the list of ADRN data sources to alleviate resource demands on this institution may result in improved access. Furthermore and relevant to the DVLA, utilising the ADRN who retrieve data on a de-identified basis before providing to the researcher, could be proposed as a method of improving data security.

There are limitations associated with this assessment of 'accessibility', which involved the identification of relevant data sources and detailed scoping discussions to determine if access to data was possible. It is possible that relevant sources were not identified and therefore not contacted. Furthermore, the focus of the research was on large regional or national sources of routinely recorded data to ensure generalisability of the results, relevant to the UK-wide SANAD II RCT. Smaller sources, such as disease specific registers are likely to have different properties regarding accessibility, as well as the subsequent assessment of quality and agreement compared to data collected using standard prospective methods.

In the following Chapter Five, the methods for the assessment of quality and agreement of routinely recorded data compared to data collected using standard methods in SANAD II will be presented.

## **Chapter Five**

### **Methods: The Attributes of Routinely Recorded Data**

#### **5.1 Introduction**

In Chapter Two the use of routinely recorded data in randomised controlled trials in the UK was reviewed and in Chapter Three a systematic review of studies that have compared the agreement of UK routinely recorded data to data collected using standard methods in prospective studies was conducted. Routinely recorded data sources in the UK were reviewed in Chapter Four and those accessible for the purposes of this study were identified.

In this chapter, the methods for the assessment of quality and agreement of routinely recorded data compared to data collected using standard methods in SANAD II will be presented. The objectives and aspects of the study design are initially presented. Subsequently, the methods for interrogation of the routinely recorded data, extraction of relevant data using an algorithmic approach and assessment of the quality and agreement for defined variables and outcome measures are presented.

#### **5.2 Research Objectives**

- 1. Compare the attributes of data extracted from electronic medical records against data collected using standard methods, in the randomised controlled trial (RCT), SANAD II:**
  - a. Assessment of the quality of data extracted from electronic medical records
  - b. Assessment of the agreement between data extracted from electronic medical records and data collected using standard prospective methods
  
- 2. Assess the Feasibility and Efficiency of Accessing and Using Routinely Recorded Data from Electronic Medical Records**

### **5.3 Study Design**

This study was a retrospective, observational study assessing the attributes of data routinely recorded in electronic medical records and accessed through administrative healthcare databases, or 'routine sources', relevant to SANAD II. The 'quality' and agreement were assessed using descriptive statistics and quantitative measures of agreement for the data variables and outcome measures, where possible. Subsequently, the feasibility and efficiency of accessing routinely recorded data were assessed.

### **5.4 Participants**

The inclusion criteria were as follows:

- Individuals aged 16 years or over
- Individuals recruited into SANAD II with a minimum of 12 months follow-up
- Individuals with capacity to consent to the retrieval of routinely recorded data

SANAD II remains in progress and participants newly recruited will have a limited follow-up period. Therefore, to ensure adequate follow-up duration, the inclusion of all eligible participants with a minimum 12 months follow-up period was specified. This pragmatic approach ensured sufficient data was recorded to assess the data variables and outcome measures relevant to SANAD II. Four hundred and seventy participants in SANAD II fulfilled the inclusion criteria, with 98 providing consent to participate in this study. A sample size calculation *a priori* was not meaningful or informative as a result of the limited number of eligible individuals; rather the impact of sample size was considered during the analyses.

## **5.5 Routinely Recorded Data Sources**

Following the assessment of accessibility presented in Chapter Four, the routinely recorded data sources included in this study are as follows:

- ***NHS Digital [55]***
  - Recording data for episodes of patient contact with NHS secondary care in England
- ***The Secure Anonymised Information Linkage Databank (SAIL) [100]***
  - Providing access to data for episodes of patient contact with NHS secondary care and in selected cases primary care for patients in Wales
- ***General Practitioners' (GP's), North West England***
  - Recording data for episodes of patient contact with NHS primary care in England
  - Access through *North West eHealth (NWEH)* or General Practice

## **5.6 Ethical and Regulatory Approvals**

This study has been reviewed and approved by the North of Scotland Research Ethics Service (29/01/16, REC reference: 16/NS/0007, Protocol number: UOL001183, IRAS project ID: 189002). A substantial amendment regarding access to primary care data was subsequently approved (19/05/16, REC1, AM01). Regulatory approval has also been provided by the Health Research Authority (05/02/16 REC reference: 16/NS/0007, Protocol number: UOL001183, IRAS project ID: 189002). Research capacity has been confirmed by The Walton Centre for Neurology and Neurosurgery NHS Foundation Trust.

SANAD II has previously been approved by the National Multi-Centre Research Ethics Committee (MREC Reference No: 12/NW/0361) and approved at each research site following independent review by local Research and Development offices.

The University of Liverpool acted as the Sponsor for this study. Delegated responsibilities were assigned to The Walton Centre for Neurology and Neurosurgery NHS Foundation Trust. The University of Liverpool holds Indemnity and insurance cover with Marsh UK LTD, which apply to this study.

This study has been completed during a Clinical Training Fellowship awarded to Dr Graham Powell, funded by the Medical Research Council Hubs for Trials Methodology Research.

## **5.7 Participant Recruitment**

The SANAD II Data Manager identified eligible individuals by review of data recorded for participants enrolled in SANAD II. Individuals' date of birth, date of enrolment and consent documentation were screened and the names and addresses of eligible individuals were retrieved. Four hundred and seventy eligible participants were sent an 'invitation pack' via the postal services. The invitation pack contained a Participant Information Leaflet, Consent Form and pre-paid addressed envelope. The information leaflet detailed the rationale and procedures involved in the retrieval of data from electronic medical records through the included routine sources. Procedures included permission to transfer the identifying variables NHS Number and SANAD II Study Number to NHS Digital, NHS Wales Informatics Service and participants GP's via NWEH, SANAD II Study Number to SAIL and permission to retrieve data from electronic medical records for the time period 2013 – present, matching the duration of SANAD II and transfer to the University of Liverpool. The information leaflet also explained that there was no obligation to take part, participation was entirely voluntary and withdrawal could occur at any time without need to provide explanation. In addition, individuals were assured that their involvement in SANAD II or their routine clinical care would not be affected by participating or not participating in this study and that data would only be retrieved on a single occasion. A proportion of participants in the North West of England were also requested to give permission for the study team and NWEH to approach their GP in order to request access to primary care data. A single 'Reminder: Letter of Invitation' was sent via the postal services accompanied with the Participant Information Leaflet and Consent Form if there had been no response following a period of three weeks. The Participant Information Leaflet and Consent Form are presented in *Appendix C*. Ninety eight individuals (20.9%) consented to participation.

## **5.8 Routinely Recorded Data Source Applications**

### *NHS Digital*

An application for Hospital Episode Statistics (HES) data from the Admitted Patient, Outpatient, Accident and Emergency and Critical Care Datasets was submitted via the newly introduced Data Access Request Service Online Portal (22/04/16). Following review by the Data Access Advisory Group, final approval and issuing of the Data Sharing Agreement was completed (29/06/16) and data for 71 individuals identified by NHS Number was provided (13/07/16). The final cost of the data, including VAT was £10,200.

### *The Secure Anonymised Information Linkage Databank (SAIL)*

An application for data from the Patient Episode Database for Wales, Outpatient, Primary Care and Emergency Department Datasets was submitted (01/02/16). The application was approved following review by the Information Governance Review Panel (26/07/16) and the data for 27 individuals identified by NHS Number was provided following issue of the Data Release Agreement (22/08/16). The final cost of the data, including VAT was £3390.

### *General Practitioners' (GP's), North West England*

An application for primary care data from participants resident in North West England through North West eHealth was planned. However, of 18 patients providing consent, only three were registered to eligible General Practice's enrolled with North West eHealth and with the required Apollo Data Extraction Tool installed. The cost was very high at £16,800 and considered an inefficient use of resources.

A substantial amendment was approved by the North of Scotland Research Ethics Service and Letter of Access provided by the National Institute of Health Research Clinical Research Network to permit the Principal Investigator (Dr Graham Powell) to approach the General Practices of participants' resident in North West England directly. The GP was formally requested to provide permission and access for the Principal Investigator to attend the practice on a single occasion and manually review the participant's electronic medical record. A Letter of Invitation was sent via the postal services accompanied with supporting documentation, including Research Ethics and Health Research Authority Approval Documents, Statement of Activities and Schedule of Events and the Data Extraction Procedure and Form. The data extraction procedure is detailed in *Box 5.1*.

### *Box 5.1: The General Practice Data Extraction Procedure*

- General Practice (GP) confirms agreement to participate in the study
  - o GP returns HRA Statement of Activities and HRA Schedule of Events
- Participant signed Consent Form will be sent to GP via postal service or secure NHS.net email address
- Dr Graham Powell (Principal Investigator) attends GP at a pre-arranged time, presents NHS and University of Liverpool ID badges and NIHR Clinical Research Network Letter of Access
  - o REC / HRA Approval Documents will be available
  - o HRA Statement of Activities / Schedule of Events will be available
- Dr Powell will sign a Confidentiality Agreement / Confidentiality Register, provided by the GP if required
- GP staff member will provide access to a computer terminal and access to the participant's electronic medical record
- Dr Powell will review the participant's electronic medical record
  - o All information in the electronic medical record will be reviewed, including READ codes, free text entries, investigations requested and results and clinic letters including referral letters
- Dr Powell transcribes information that is relevant to the study to the Data Extraction Form, pre-labelled with the participant's Study ID Number. Identifiable information will not be transcribed
- On completion, the GP will have the opportunity to review the Data Extraction Form to confirm their agreement with the information that has been transcribed
- Dr Powell transfers in person the Data Extraction Form to the University of Liverpool. Data is transcribed to the secure study database and the Data Extraction Form is destroyed

The General Practices of 18 participants resident in North West England were approached. Two GP's provided consent and data were extracted for the participants in accordance with the protocol in July 2016. Three GP's refused participation and the remaining 13 GP's provided no response despite repeated attempts at contact.

## **5.9 Data Management**

All personal data in this study has been kept strictly confidential and was handled, stored and destroyed in accordance with the Data Protection Act 1998. The process of identification and invitation of eligible individuals enrolled in SANAD II to participate in this study involved access to the SANAD II database. SANAD II data is managed by the University of Liverpool Clinical Trials Research Centre (CTRC) and existing data security standards were maintained during this study. The SANAD II Data Manager screened date of birth, date of enrolment and consent documentation and the names and addresses of eligible individuals were retrieved. For the 98 participants agreeing to consent to this study, the identifying variables NHS Number and SANAD II Study Number were transferred to NHS Digital and NWIS and SANAD II Study Number to SAIL using a secure electronic transfer system. Subsequently, participants data from electronic medical records accessed through the routine sources was transferred to the University of Liverpool using the NHS Digital Secure File Transfer (SFT) System [187]. 'Study data' consisted of data provided by the routine data sources and data extracted from the SANAD II database. The SANAD II Data Manager received the data from routine sources and linked to data extracted from SANAD II. These datasets were then pseudonymised, identified by a Unique Study Number before being provided to the Principal Investigator (Dr Graham Powell).

Study data were stored using the University of Liverpool Research Data Management Service's *DataStore* [188] and encrypted using industry standard techniques meeting the NHS Information Governance Toolkit standard (8HN20). The data storage location was as follows: *livad.liv.ac.uk\rdm\projectstore\RoutineData*. The Principal Investigator (Dr Graham Powell) acted as *Data Processor*. The University of Liverpool acted as *Data Controller*. The University of Liverpool *Information Security Policy*, informed by the principles set out in ISO 27001 and *Research Data Management Policy* were adhered to throughout the study.

The participant consent forms were the only non-electronic data stored during the study. Consent forms were stored in a locked filing cabinet in the locked SANAD II coordinating centre office (2<sup>nd</sup> Floor, Clinical Sciences Centre, University of Liverpool, L9 7AL).

Study completion is defined by the date of 31/12/18. Fully anonymised study data will be stored in the Research Data Management Service *Archiving Repository* and deleted five years following the date of study completion.

## **5.10 Routinely Recorded Data and Data Coding Systems**

Data routinely recorded in electronic medical records and held in administrative healthcare databases in the United Kingdom are routinely recorded using coding systems. In England, NHS Digital regulates the coding systems and in Wales the NHS Wales Informatics Service. The coding systems between countries are comparable. The NHS data dictionary provides an overview of the coding systems used in the NHS [189].

### ***5.10.1 NHS Digital: Hospital Episode Statistics (HES)***

NHS Digital routinely record data regarding secondary care for patients in England published as Hospital Episode Statistics. In this study, data from four HES datasets were requested:

- *Admitted Patient Care Dataset*
- *Critical Care Dataset*
- *Accident & Emergency Dataset*
- *Outpatient Dataset*

NHS Digital publish data dictionaries for each dataset [190]. In each dictionary, derived from the NHS data dictionary, the individual data fields are presented together with the coded data values for interpretation. For example, from the Admitted Patient Care Dataset:

- *Data field:*     *Admin category at start of episode (ADMINCATST)*
- *Data Values:*    *01 = NHS patient; 02 = Private patient*

The data fields are comparable between datasets, however the available data fields differ, with the Admitted Patient Care Dataset providing the most comprehensive data. The data fields and coded interpretation are presented in detail in the data dictionaries with the exception of clinical data. The Admitted Patient Care and Outpatient Datasets use the International Statistical Classification of Diseases and Related Health Problems (10), discussed in section 5.10.4. However, the Accident & Emergency Dataset records a less detailed diagnostic coding system, presenting diagnosis by broad disease area. For example, '*CNS Disorder*' and '*Central Nervous System Disorder – Epilepsy / Seizure*' were the diagnostic codes most informative to this study. The Critical Care Dataset does not record details regarding diagnosis. Clinical and surgical procedures are recorded using the Office of Population Census and Surveys, Version 4 (OPCS 4) coding system in the Admitted Patient Care and Outpatient Datasets, discussed in section 5.10.4.

### **5.10.2 NHS Wales Informatics Service (NWIS): The Secure Anonymised Information**

#### **Linkage Databank (SAIL):**

Data routinely recorded by NWIS regarding primary and secondary care for patients in Wales were accessed through SAIL. In this study, data from four datasets were requested:

- *Patient Episode Database for Wales*
- *Emergency Department Dataset*
- *Outpatient Dataset*
- *Primary Care Dataset*

NWIS publish data dictionaries for each dataset [191]. In each dictionary, derived from the NHS data dictionary, the individual data fields are presented together with the coded data values for interpretation. Similar to HES datasets, the data fields are comparable between datasets but the availability of data fields differs, with the Patient Episode Database for Wales providing the most comprehensive data. Again similar to HES datasets, diagnostic information is recorded using the International Statistical Classification of Diseases and Related Health Problems (10), with the exception of the Emergency Department Dataset which uses a less detailed diagnostic coding system presenting diagnosis by broad disease area ('*CNS Disorder*', '*Central Nervous System Disorder – Epilepsy / Seizure*'). Finally, clinical and surgical procedures are recorded using the Office of Population Census and Surveys, Version 4 (OPCS 4) coding system in the Patient Episode Database for Wales and Outpatient Datasets.

#### **5.10.3 Primary Care Data**

Primary care data are routinely recorded in electronic medical records by participants General Practitioners (GP). The GP records data both using free text entries and the UK READ coding system, discussed in section 5.10.4 [57]. In this study UK READ codes and free-text entries were reviewed for the two participants in the General Practices where direct access was permitted. All recorded UK READ codes for the relevant time period were also provided for participants registered to General Practices enrolled in the SAIL Primary Care Dataset.

#### **5.10.4 Clinical Coding**

Clinical data including diagnostic data are recorded in electronic medical records using clinical coding systems, differing in the routine datasets accessed during this study between primary and secondary care:

##### **Secondary Care:**

###### *International Statistical Classification of Diseases and Related Health Problems (ICD) 10*

ICD 10 is the 10th revision of the medical classification system by the World Health Organization (WHO). Included are codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. ICD 10 codes are included in both HES and NWIS datasets. ICD 10 codes can be accessed in website form [56]. In addition, NHS Digital provide access to an eBook of ICD 10 codes through their Technology Reference Data Update Distribution (TRUD) service [192].

###### *Office of Population Census and Surveys, Version 4 (OPCS 4)*

OPCS 4 is the coding system used for procedures within NHS hospitals of England, Scotland, Wales and Northern Ireland. OPCS 4 includes codes for surgical operations, procedures and interventions performed in secondary care. NHS Digital provide access to an eBook of OPCS 4 codes through their Technology Reference Data Update Distribution (TRUD) service [192].

##### **Primary Care:**

###### *NHS UK READ Codes Clinical Terms Version 2*

UK READ codes are the standard clinical terminology coding system used in General Practice. The UK READ system includes codes for clinical data (signs, symptoms, observations, investigations, results, diagnoses and diagnostic, therapeutic or surgical procedures) and administrative data (occupation, social circumstances, ethnicity, religion) [57]. The implementation of UK READ codes in General Practice is broader, compared to ICD 10 in secondary care datasets which is generally used only for diagnostic data. UK READ Clinical Terms Version 3 is currently being implemented and may have been used in some of the General Practices providing data for this study. NHS Digital provide access to the UK READ Browser through their Technology Reference Data Update Distribution (TRUD) service [193].

The Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT) is a clinical coding system standardised internationally and including currently active UK READ codes. The UK National Informatics Board has ruled that SNOMED CT will be used as the primary clinical coding system across all healthcare in the UK by 2020 and it is possible that some General Practices included in this study were using this coding system. SNOMED CT is available in browser form [194] or through the NHS Digital Technology Reference Data Update Distribution (TRUD) service [195].

Notably, whether Clinical Terms Version 2, Version 3 or SNOMED CT coding systems were used in the General Practices, the UK READ codes are consistent.

### **5.11 The Routinely Recorded Data**

An inclusive approach to the routinely recorded data variables requested was taken to maximise the data available. This ensured the 'best case' routinely recorded data were available for the analyses.

*Box 5.2* presents a summary of the routinely recorded data requested.

## Box 5.2: Summary of Data Requested from Routine Sources

### Secondary Care Data: HES / SAIL

- **Emergency Care Datasets:**
- *All attendances for study participants during the study period:*
  - Date of attendance
  - Arrival method
  - Clinical Data:
    - 'Reason for attendance' / diagnoses
    - Investigations
    - Treatment
  - Disposal method (discharge, admit)
- **Inpatient Care Datasets:**
- *All attendances for study participants during the study period:*
  - Details of Admission ('Finished Consultant Episodes' and 'Hospital Spells'):
    - Date of admission, date of discharge, method of admission (elective, emergency)
  - Clinical Data:
    - Diagnoses, investigations, treatments, procedures
  - Socioeconomic Data
    - Aggregated data (LSOA), (Welsh) Index of Multiple Deprivation (IMD)
  - Health Economic Data
    - Cost of treatment and procedures (Healthcare Resource Groups (HRG's))
  - Disposal method (discharge, death)
- **Outpatient Care Datasets:**
- *All attendances for study participants during the study period:*
  - Appointment type (urgent / routine new / follow up)
  - Referral Details:
    - Source, urgency, waiting time
  - Clinical Data
    - Diagnoses, investigations, treatment
- **Critical Care Dataset:**
- *All attendances for study participants during the study period:*
  - Admission details, date, duration
  - Clinical details, support provided, Level 2, Level 3 care
  - Disposal Details, discharge destination

### Primary Care Data: General Practitioners' North West England / SAIL

- **Primary Care Dataset:**
- *All attendances for study participants during the study period:*
  - Clinical Data
    - Consultation record, diagnoses, seizures
    - Adverse drug reactions / Adverse events
  - Referrals to neurology and secondary care
  - Prescriptions
  - Diagnostic testing, investigation results (MRI, CT, EEG)
  - Lifestyle information (e.g. smoking and alcohol status, employment status)
  - Socioeconomic Data
    - Non-identifiable, provided at the Lower Layer Super Output Area (LSOA) level

### 5.11.1 NHS Digital: Hospital Episode Statistics (HES)

NHS Digital Hospital Episode Statistics (HES) [55] data were requested and retrieved. The cost was calculated on a not-for-profit, cost recovery basis. The total cost of HES data in this study, excluding VAT was £8,300. *Table 5.1* presents the cost calculation using data provided by NHS Digital.

*Table 5.1: NHS Digital HES Cost Calculation*

Study Element	Cost (£)
Base cost	1000
Release fee	800
Agreement, 3 years	500
Data cost	6000
<b>Total</b>	<b>8300 (+VAT)</b>

The Data Protection Act [50] provides the legal framework for data release. In this study, participant informed written consent provided the valid legal basis for data release. Data minimisation principles were adhered to and therefore only the minimum data required to meet the research objectives were requested. However, as a result of the study objectives, the majority of HES data variables, with the exception of identifiable variables, were requested. A Data Sharing Framework Contract at institutional level and Data Sharing Agreement at study level were required. Data were provided via secure electronic transfer in SQL format. The HES Data Dictionaries [190] provided information to inform the interpretation of included data variables.

*Table 5.2* presents an overview of the HES datasets accessed in this study.

*Table 5.2: An Overview of the HES Datasets*

	APC Dataset	A & E Dataset	OP Dataset	CC Dataset
Number of Participants	43	65	71	1
Number of Attendances	125	230	1193	1
Median (IQR)	2(1-4)	2(1-4)	12(7-20)	N/A
Dates Included	02/04/13 – 21/03/16	02/04/13 – 15/03/16	02/04/13 – 31/03/16	16/12/14

### *Admitted Patient Care Dataset (APC)*

The Admitted Patient Care (APC) Dataset presents data regarding inpatient admissions in England. Admissions are recorded as 'Episodes' of care under a named Consultant. Single or multiple episodes may be included in a single hospital admission, or 'Spell'. A large number of data variables are included; both directly extracted from the SUS dataset and derived from the extracted data. Notable limitations include the absence of data regarding treatments and investigations. Data variables relevant to the objectives of this study are presented in *Table 5.3*.

### *Accident & Emergency Dataset (A&E)*

The Accident & Emergency Dataset (A&E) presents data regarding emergency department attendances in England. A large number of data variables are included; both directly extracted from the SUS dataset and derived from the extracted data. Notable limitations include the non-specific diagnostic coding system and limited detail regarding investigations and treatment. Data variables relevant to the objectives of this study are presented in *Table 5.4*.

### *Outpatient Dataset (OP)*

The Outpatient Dataset (OP) presents data regarding outpatient attendances in England. A large number of data variables are included; both directly extracted from the SUS dataset and derived from the extracted data. Notable limitations include the absence of data regarding treatments and investigations and limited completion of diagnostic information. Data variables relevant to the objectives of this study are presented in *Table 5.5*.

### *Critical Care Dataset (CC)*

The Critical Care Dataset (CC) presents data regarding admission to critical care departments in England. Data variables are directly extracted from the SUS dataset and include information regarding the admission to critical care, treatment administered and discharge details. The limited data variables available are presented in *Table 5.6*.

Table 5.3: An Overview of HES Admitted Patient Care Data Variables

Data Variable	Description
ACTIVAGE	Age
ADMIDATE	Date of admission
ADMIMETH	Admission method
ADMISORC	Admission source
ALCDIAG_4	Alcohol-related diagnosis
ALCFRAC	Alcohol attributable fraction, based on alcohol related diagnosis
CLASSPAT	Type of admission
DIAG_01 - 20	ICD 10 diagnostic code
DISDATE	Date of discharge
DISDEST	Discharge destination
DISMETH	Discharge method
ELECDATE	Date decision to electively admit patient
ELECDUR	Elective admission waiting time
EPIDUR	Episode duration
EPIEND	Date of episode end
EPIORDER	Order of episode
EPISTART	Date of start of episode
ETHNOS	Ethnicity
HRGNHS	Trust generated Healthcare Resource Group code
HRGNHSVN	Healthcare Resource Group Version
IMD04	Index of Multiple Deprivation (IMD) at Super Output Area level
IMD04_DECILE	Interpretation of IMD
IMD04C	IMD crime domain
IMD04ED	IMD education, skills and training domain
IMD04EM	IMD employment domain
IMD04HD	IMD health and disability domain
IMD04HS	IMD barriers to housing and services domain
IMD04I	IMD income domain
IMD04IA	IMD income deprivation affecting older people index
IMD04IC	IMD income deprivation affecting children index
IMD04LE	IMD living environment domain
IMD04RK	IMD overall rank
INTMANIG	Patient intended management
LSOA11	Lower Super Output Area, 2011 Census
MAINSPEF	Main Speciality of Consultant
MSOA11	Middle Super Output Area, 2011 Census
OPDATE_01 - 24	Date of operation
OPERSTAT	Operation status
OPERTN_01 - 24	OPCS4 procedure code
POSOPDUR	Postoperative duration
POSTDIST	Postcode district
PREOPDUR	Preoperative duration
PROCODE	NHS provider code
PROTYPE	Type of provider
PURCODE	Type of commissioner
RURURB_IND	Rural / urban indicator
SEX	Sex
SPELBGIN	Start of hospital spell
SPELDUR	Duration of hospital spell
SPELEND	End of hospital spell
SUSCOREHRG	SUS generated Healthcare Resource Group, spell level
SUSHRG	SUS generated Healthcare Resource Group, episode level
TRETSPEF	Treatment speciality
WAITDAYS	Duration of wait for elective admission

Table 5.4: An Overview of HES Accident & Emergency Data Variables

Data Variable	Description
ACTIVAGE	Age
AEARRIVALMODE	Method of arrival
AEATTENDCAT	Attendance category (planned, unplanned)
AEATTENDDISP	Disposal method
AEDEPTTYPE	Type of emergency care department
AEPATGROUP	Attendance type
ARRIVALDATE	Date of attendance
ARRIVALTIME	Time of attendance
CONCLDUR	Duration of attendance
DIAG_01 - 12	A&E diagnostic code
DIAGA_01 - 12	Anatomical area
DIAGS_01 - 12	Lateralisation
ETHNOS	Ethnicity
HRGNHS	Trust generated Healthcare Resource Group code
HRGNHSVN	Healthcare Resource Group version number
IMD04	Index of Multiple Deprivation (IMD) at Super Output Area level
IMD04_DECILE	Interpretation of IMD
IMD04C	IMD crime domain
IMD04ED	IMD education, skills and training domain
IMD04EM	IMD employment domain
IMD04HD	IMD health and disability domain
IMD04HS	IMD barriers to housing and services domain
IMD04I	IMD income domain
IMD04IA	IMD income deprivation affecting older people index
IMD04IC	IMD income deprivation affecting children index
IMD04LE	IMD living environment domain
IMD04RK	IMD overall rank
INITDUR	Duration from arrival to assessment
INVEST_01 - 12	Investigation code
LSOA11	Lower Super Output Area, 2011 Census
MSOA11	Middle Super Output Area, 2011 Census
POSTDIST	Postcode district
PROCODE	NHS provider code
PROTYPE	Type of provider
PURCODE	Type of commissioner
RURURB_IND	Rural / urban indicator
SEX	Sex
SUSHRG	SUS generated Healthcare Resource Group
SUSHRGVERS	Healthcare Resource Group Version
TREAT_01	Treatment code

Table 5.5: An Overview of HES Outpatient Data Variables

Data Variable	Description
APPTAGE	Age
APPTDATE	Date of attendance
ATENTYPE	Type of attendance
DIAG_01 - 12	ICD 10 diagnostic code
DNADATE	Date of last 'Did Not Attend' or cancellation
ETHNOS	Ethnicity
HRGNHS	Trust generated Healthcare Resource Group code
HRGNHSVN	Healthcare Resource Group version number
IMD04	Index of Multiple Deprivation (IMD) at Super Output Area level
IMD04_DECILE	Interpretation of IMD
IMD04C	IMD crime domain
IMD04ED	IMD education, skills and training domain
IMD04EM	IMD employment domain
IMD04HD	IMD health and disability domain
IMD04HS	IMD barriers to housing and services domain
IMD04I	IMD income domain
IMD04IA	IMD income deprivation affecting older people index
IMD04IC	IMD income deprivation affecting children index
IMD04LE	IMD living environment domain
IMD04RK	IMD overall rank
LSOA11	Lower Super Output Area, 2011 Census
MAINSPEF	Main Speciality of Consultant
MSOA11	Middle Super Output Area, 2011 Census
OPERSTAT	Status of operation, if pending
OPERTN_01 - 24	OPCS4 procedure code
OUTCOME	Outcome of attendance
POSTDIST	Postcode district
PRIORITY	Priority of referral
PROCODE	NHS provider code
PROTYPE	Type of provider
PURCODE	Type of commissioner
REFSOURC	Source of referral
REQDATE	Date referral received
RURURB_IND	Rural / urban indicator
SERVTYPE	Service requested
SEX	Sex
STAFFTYP	Type of healthcare practitioner, grade
SUSHRG	SUS generated Healthcare Resource Group
TRETSPEF	Treatment speciality
WAITDAYS	Time to elective treatment
WAITING	Time from referral to OP review

Table 5.6: An Overview of HES Critical Care Data Variables

<b>Data Variable</b>	<b>Description</b>
ACARDSUPDAYS	Days of advanced cardiovascular support
ARESSUPDAYS	Days of advanced respiratory support
BCARDSUPDAYS	Days of basic cardiovascular support
BRESSUPDAYS	Days of basic respiratory support
CCADMISORC	Source of admission
CCADMITYPE	Type of admission
CCDISDATE	Date of discharge
CCDISDEST	Discharge destination
CCDISLOC	Discharge location
CCDISRDYDATE	Discharge 'ready date'
CCDISSTAT	Discharge status
CCSORCLOC	Source of admission, location
CCSTARTDATE	Date of admission
CCSTARTTIME	Time of admission
CCUNITFUN	Specific critical care unit function
DERMSUPDAYS	Days of dermatological support
GISUPDAYS	Days of gastrointestinal support
LIVERSUPDAYS	Days of liver support
NEUROSUPDAYS	Days of neurological support
ORGSUPMAX	Maximum number of organs supported
RENSUPDAYS	Days of renal support

### 5.11.2 The Secure Anonymised Information Linkage Databank (SAIL)

The NHS Wales Informatics Service (NWIS) datasets were complemented by additional datasets provided by The Secure Anonymised Information Linkage Databank (SAIL) [100]. The cost was calculated on a not-for-profit, cost recovery basis. The total cost of SAIL data in this study, excluding VAT was £2828. Table 5.7 presents the cost calculation using data provided by SAIL.

Table 5.7: SAIL Cost Calculation

Study Element	Timescale	Effort (Days)	Cost (£)
Base cost	n/a	n/a	500
Load data into SAIL	4 weeks	1	291
Create individual level output	2 weeks	5	1455
Data transfer	2 weeks	2	582
<b>Total</b>			<b>2828 (+VAT)</b>

The Data Protection Act [50] provides the legal framework for data release. In this study, participant informed written consent provided the valid legal basis for data release. Data minimisation principles were adhered to and therefore only the minimum data required to meet the research objectives were requested. However, as a result of the study objectives, the majority of data variables, with the exception of identifiable variables, were relevant and were requested. A Data Release Agreement at study level was required. Data were provided via secure electronic transfer in Microsoft Excel format. ‘Metadata’ files were also provided to inform the interpretation of data.

Table 5.8 presents an overview of the SAIL datasets accessed in this study.

Table 5.8: An Overview of the SAIL Datasets

	PEDW Dataset	EDDS Dataset	OP Dataset	GP Dataset
Number of Participants	16	22	27	23
Number of Attendances	34	54	192	5379
Median (IQR)	2(1-3)	1.5(1-3)	7(4-8)	161(137-316)*
Dates Included	11/05/13 – 24/02/16	10/05/13 – 04/12/15	09/05/13 – 24/12/15	30/04/13 – 23/12/15

\*READ Code Entries: may not all represent individual attendances

### *Patient Episode Database for Wales (PEDW)*

The Patient Episode Database for Wales (PEDW) presents data regarding inpatient admissions in Wales and is comparable to the HES APC dataset. Admissions are recorded as 'Episodes' of care under a named Consultant. Single or multiple episodes may be included in a single hospital admission, or 'Spell'. A large number of data variables are included; both directly extracted from the NHS Wales Data Warehouse and derived from the extracted data. Notable limitations include the absence of data regarding treatments and investigations. Data variables relevant to the objectives of this study are presented in *Table 5.9*.

### *Emergency Department Dataset (EDDS)*

The Emergency Department Dataset (EDDS) presents data regarding emergency department attendances in Wales and is comparable to the HES A&E dataset. A large number of data variables are included; both directly extracted from the NHS Wales Data Warehouse and derived from the extracted data. Notable limitations include the non-specific diagnostic coding system and limited detail regarding investigations and treatment. Data variables relevant to the objectives of this study are presented in *Table 5.10*.

### *Outpatient Dataset (OP)*

The Outpatient Dataset (OP) presents data regarding outpatient attendances in Wales and is comparable to the HES OP dataset. A large number of data variables are included; both directly extracted from the NHS Wales Data Warehouse and derived from the extracted data. Notable limitations include the absence of data regarding treatments and investigations and limited completion of diagnostic information. Data variables relevant to the objectives of this study are presented in *Table 5.11*.

### Primary Care (GP) Dataset

The Primary Care (GP) Dataset presents data regarding primary care attendances in Wales for General Practices enrolled within SAIL. Limited data variables are presented including date of event, UK READ Code, UK READ Code description and limited specific values, such as systolic and diastolic blood pressure measurements. Notable limitations include the absence of free text entries and lack of context to the recording of READ Codes. Healthcare events including the receipt of clinical correspondence, investigation results, medication prescriptions and clinical prompts may result in READ code entry, in addition to patient attendance. Data variables are presented in *Table 5.12*.

*Table 5.9: An Overview of SAIL Patient Episode Database for Wales Data Variables*

Data Variable	Description
<b>PEDW Episode</b>	
PROV_UNIT_CD	NHS Provider Code
EPI_NUM	Episode number
EPI_STR_DT	Date of start of episode
EPI_END_DT	Date of end of episode
AGE_EPI_STR_YR	Age at start of episode
CON_SPEC_MAIN_CD	Main speciality of Consultant
CON_SPEC_CD_OF_TREAT	Treatment speciality
EPI_DUR	Duration of episode
DIAG_CD_123	ICD 10 diagnostic code, 3 digits
DIAG_CD_1234	ICD 10 diagnostic code, 4 digits
OPER_CD_123	OPCS4 procedure code, 3 digits
OPER_CD	OPCS4 procedure code
HRG_LOCALPAYMENT_CD	Healthcare Resource Group
HRG_LOCALPAYMENT_DESC	Healthcare Resource Group description
HRG_REFERENCECOST_CD	Healthcare Resource Group reference cost
HRG_REFERENCECOST_DESC	Healthcare Resource Group reference cost description
<b>PEDW Spell</b>	
SPELL_NUM_E	Spell number
GNDR_CD	Gender
ADMIS_DT	Date of admission
ADMIS_MTHD_CD	Admission method
ADMIS_SOURCE_CD	Admission source
INTENDED_MANAGEMENT_CD	Patient intended management
DISCH_DT	Date of discharge
DISCH_MTHD_CD	Discharge method
DISCH_DESTINATION_CD	Discharge destination
DUR_ELECT_WAIT	Elective admission waiting time
PAT_CLASS_CD	Type of admission
SPELL_DUR	Duration of spell
ADMIS_SPEC_CD	Treatment speciality
ADMIS_DEC_DT	Date decision to electively admit patient
LSOA_CD_2001	Lower Super Output Area, 2011 Census

Table 5.10: An Overview of SAIL Emergency Department Data Variables

Data Variable	Description
ADMIN_ARR_DT	Date of attendance
ADMIN_ARR_TM	Time of attendance
AGE	Age
ARRIVAL_MODE	Method of arrival
HEALTH_EVENT_DT	Date of health event
HEALTH_EVENT_TM	Time of health event
ATTEND_GROUP	Attendance type
ATTEND_CATEGORY	Attendance category (planned, unplanned)
DIAG_CD_1 - 6	EDDS diagnostic code
ANAT_AREA_CD_1 - 6	Anatomical area
SIDE_CD_1 - 6	Lateralisation
TREAT_CD_1 - 6	Treatment code
INVEST_CD_1 - 6	Investigation code
ADMIN_END_DT	Date of end of attendance
ADMIN_END_TM	Time of end of attendance
DISCHARGE	Outcome of attendance
LOCATION_TYPE	Location of injury, where relevant
ROAD_USER	Road user status, where relevant
MECH_OF_INJ	Mechanism of injury, where relevant
ACTIVITY	Activity at time of injury, where relevant
SPORT	Sport at time of injury, where relevant
ALCOHOL_IND	Role of alcohol in injury, where relevant
TRIAGE_CAT	Triage category

Table 5.11: An Overview of SAIL Outpatient Data Variables

Data Variable	Description
AGE_AT_APPT	Age
REF_DT	Date of patient referral
CLINICAL_REF_DT	Date of clinical referral
PRIORITY_TYPE_CD	Priority of referral
SOURCE_OF_REF_CD	Source of referral
CON_SPEC_MAIN_CD	Main speciality of Consultant
CON_SPEC_CD_OF_TREAT	Treatment speciality
LOCAL_SPEC_CD	Local sub-speciality
ADMIN_CAT_CD	Patient treatment category
LOC_TYPE_CD	Outpatient clinic location type
MED_STAFF_TYPE_CD	Type of healthcare practitioner, grade
ATTEND_DT	Date of attendance
FIRST_ATTEND_CD	Date of first attendance
ATTEND_CD	Attendance status
OUTCOME_CD	Outcome of attendance
LAST_DNA_CANCEL_DT	Date of last 'Did Not Attend' or cancellation
DIAG_CD_4	ICD 10 diagnostic code
OPER_CD_4	OPCS4 procedure code

*Table 5.12: An Overview of SAIL Primary Care Data Variables*

<b>Data Variable</b>	<b>Description</b>
EVENT_DT	Date of record
EVENT_CD	UK READ Code
TERM_COMBINED	UK READ Code – description
EVENT_VAL	'Number', where relevant, for example blood pressure
SYSTOLIC_VAL	Systolic blood pressure
DIASTOLIC_VAL	Diastolic blood pressure

### 5.11.3 Primary Care Data, North West England

Primary care data were requested directly from the General Practices for the 18 consenting participants' resident in North West England. Two General Practices provided consent and the principal investigator attended the practices and directly extracted relevant data in July 2016 from the EMIS IT system. Three General Practices refused participation and the remaining 13 General Practices provided no response despite repeated attempts at contact.

Participants' complete electronic medical records were reviewed and data extracted. Data reviewed included UK READ Codes, free text entries, investigation results and clinical correspondence. *Table 5.13* presents the data included in participants' primary care electronic medical records.

*Table 5.13: Data Included in Participants Primary Care Electronic Medical Records*

<ul style="list-style-type: none"><li>- Demographic Details:<ul style="list-style-type: none"><li>o Name, Date of Birth, Address, NHS Number</li></ul></li><li>- Active Diagnoses / Problems</li><li>- Significant Past Medical History</li><li>- Family History</li><li>- Medication:<ul style="list-style-type: none"><li>o Acute Prescriptions</li><li>o Repeat Prescriptions</li><li>o Past Prescriptions</li></ul></li><li>- Allergies</li><li>- Health Status:<ul style="list-style-type: none"><li>o Smoking / Alcohol Status</li><li>o Weight, Height</li></ul></li><li>- Planned Events:<ul style="list-style-type: none"><li>o Screening</li><li>o Medication Reviews</li></ul></li><li>- Investigations / Results</li><li>- Medical Record:<ul style="list-style-type: none"><li>o Consultations</li><li>o Clinical Correspondence<ul style="list-style-type: none"><li>▪ Sent</li><li>▪ Received</li></ul></li><li>o UK READ Codes</li></ul></li></ul>
--

#### 5.11.4 Study Participants

A total of 98 individuals consented to participation in this study. Seventy one participants were resident in England and included in the NHS Digital Hospital Episode Statistics Datasets. Twenty seven participants were resident in Wales and included in the Secure Anonymised Information Linkage Databank Datasets. Data recorded during SANAD II were available for all participants. Data were available for all participants from NHS Digital or The Secure Anonymised Information Linkage Databank from January 2013 to the most recently available date, as previously detailed in sections 5.11.1, 5.11.2 and 5.11.3.

##### 5.11.4.1 Gender

Gender was recorded for all participants in SANAD II and there were 55 males and 43 females in the total 98 study participants.

Gender was recorded in the Hospital Episode Statistics (HES) Admitted Patient Care (APC), Accident and Emergency (A&E) and Outpatient (OP) Datasets and data were present for all participants and consistent between datasets. In the Secure Anonymised Information Linkage (SAIL) Databank gender was recorded only in the Patient Episode Database for Wales (PEDW) Dataset. In routine datasets, data regarding gender were available in 87 participants, 11 participants living in Wales did not have an inpatient episode recorded and therefore data regarding gender was missing. In the 87 participants with data available from both sources, gender determined from the routine datasets was compared to the gender recorded in SANAD II, summarised in the cross-tabulation, *Table 5.14*.

*Table 5.14: Gender: Cross-Tabulation*

RCT Data	Routine Data		
	Male	Female	Total
Male	49 (56.3%)	0	49
Female	0	38 (43.7%)	38
Total	49	38	87

#### *5.11.4.2 Age*

Date of birth was recorded for all participants in SANAD II and therefore age in years could be calculated.

Age in years was recorded in the HES APC, A&E and OP Datasets. In the SAIL Databank age in years was recorded in the Emergency Department Dataset (EDDS), PEDW and OP Datasets and was present in routine sources for all participants. Date of birth as an identifiable variable was not required as a result of data minimisation principles and therefore age in years was provided.

Age in years was calculated from the recorded date of birth in SANAD II and adjusted for the year of routine data record, matching the age recorded in routine datasets.

The mean age of the normally distributed sample was 50 (Range: 17-89, Standard Deviation: 20).

## **5.12 Data Processing and Analysis**

### **5.12.1 Overview**

This study assessed the quality and agreement of data routinely recorded in electronic medical records compared with data collected using standard prospective methods in SANAD II. The exploratory nature of the study necessarily resulted in an iterative approach to the analyses. For example, prior to the data being retrieved it was unclear which data variables would be in a comparable format to those recorded on the SANAD II database suitable for a statistical assessment of agreement. *A priori*, the data variables and outcome measures where an assessment of agreement was intended were specified. For example, an assessment of agreement was planned for the variable 'first follow-up seizure' and subsequently outcome measure 'time to first follow-up seizure'. Notably, assessment of the SANAD II outcomes stratified by treatment intervention (antiepileptic drug) under study would be inappropriate as a result of the on-going status of the trial. Therefore, throughout the analyses the included participants in this study were analysed as a complete cohort, without reference to antiepileptic drug or SANAD II study arm.

### **5.12.2 Data Preparation**

The inherent properties of routinely recorded data resulted in a period of data preparation being appropriate. The process included the assessment for duplicate entries and formatting the data for subsequent analysis. All analyses in this study used the prepared datasets. Such 'cleaned' data is the standard for clinical research studies accessing and implementing routinely recorded data. The original dataset and all subsequent formatted datasets were stored, identified by appropriate version numbers. Furthermore, for all analyses data were extracted from all routinely recorded data sources to create a 'best case' dataset. For example, in the assessment of 'total number of follow-up seizures', all datasets for all participants were reviewed to ensure all occurrences of follow-up seizures were identified, with the dataset for each seizure occurrence also recorded. This 'best case' dataset was then used in the assessment of quality and agreement compared with the SANAD II dataset.

### **5.12.3 Assessment of Routinely Recorded Data Quality**

An assessment of 'quality' was appropriate to ensure the data provided in the routinely recorded datasets were comparable to the data collected in SANAD II and that the subsequent assessments of agreement were valid. Quality is a term used throughout this study and the following factors were included in its definition:

- An assessment of the 'comparability' of routinely recorded data variables with data collected using standard prospective methods in SANAD II. This ensured data variables from both datasets were comparable, measuring the same underlying construct and ensured further analyses were valid, including the assessment of agreement.
- An assessment of the 'completeness' of routinely recorded data. This assessed the degree of missing data and resultant systematic bias, compared to data collected using standard prospective methods in SANAD II.

### **5.12.4 Assessment of Agreement**

A statistical assessment of agreement was completed between routinely recorded data and data collected using standard prospective methods in SANAD II for comparable data variables and outcomes measures. Statistical and clinical differences were discussed. Methods for the assessment of agreement for categorical and continuous data are presented. All analyses were performed in *SPSS (Version 22)*.

### *Continuous Data*

To assess agreement between paired (data from different datasets for the same individual) continuous data, Bland Altman methods were employed. Acceptable clinical limits of agreement for each variable or outcome were specified *a priori*. In the first instance, the *Difference* and *Mean* between the datasets were computed. The distribution of data was then assessed using the *Difference* variable through construction of a histogram. For normally distributed data a Paired T Test was performed and for non-normally distributed data a Wilcoxon Signed Rank Test. A P-value <0.05 was taken as indicating a significant difference between the means calculated from the two datasets. Subsequently, Bland Altman Plots were constructed, the *Difference* variable plotted on the Y axis and the *Mean* variable plotted on the X axis. The mean of the *Difference* variable was also plotted. The 95% confidence limits of agreement were calculated by multiplying the standard deviation of the *Difference* variable by 1.96 and subsequently adding or subtracting from the mean of the *Difference* variable. The 95% confidence limits of agreement were then discussed in the context of the specified acceptable clinical limits of agreement [196].

### *Time to Event Data*

Time to event outcomes are included in SANAD II and it was possible to calculate 'time to first seizure' and 'time to 12 month remission' outcomes using routinely recorded datasets. Kaplan Meier curves were created to estimate the survival functions using data from both sources. Subsequently, a Log Rank test was performed with P-value <0.05 indicating a statistically significant difference between the outcomes calculated from the two datasets.

### *Categorical Data*

To assess agreement between paired, nominal categorical datasets, cross tabulations were presented followed by calculation of Cohen's kappa. The interpretation of Cohen's kappa was based on the guidelines from Altman (1999), and adapted from Landis & Koch (1977): 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement. Significance was examined with P<0.05 indicating the level of agreement defined by kappa being significant.

### ***5.12.5 Assessment of Feasibility and Efficiency***

The feasibility and efficiency of accessing and using routinely recorded data was assessed.

Factors important in the assessment of feasibility and efficiency included:

- The resources, including financial and time required during development of the protocol, application and application procedures for routinely recorded data
- The outcome of applications, including final provision of routinely recorded data
- The resources required in data preparation, including data cleaning and formatting
- The attributes of the data provided, including time period coverage, validity and agreement to data collected using standard prospective methods in SANAD II

To inform the assessment of feasibility, a prospective log was recorded throughout the duration of the study. The log included all procedures involved in the application, retrieval and preparation of routinely recorded data.

### **5.13 The Data Variables and Outcome Measures**

Data variables and outcome measures relevant to the following scenarios were assessed:

- *The identification of eligible individuals and recruitment into SANAD II*
- *The follow-up of participants and measurement of SANAD II outcomes*

#### **5.13.1 The SANAD II Data**

##### *5.13.1.1 The Identification of Eligible Individuals and Recruitment into SANAD II*

Individuals eligible for recruitment into SANAD II are identified by the clinician during routine clinical practice. During a clinical review, a diagnosis and classification of epilepsy is made and the decision to commence treatment with antiepileptic drugs discussed and agreed with the patient. Individuals agreeing to participate in SANAD II then complete a SANAD II 'baseline assessment'. During this assessment, the clinician completes a Case Report Form (CRF). Data recorded is summarised in *Box 5.3* and includes the date of first seizure, seizure type, total number of seizures, status and results of investigations such as electroencephalography, medication history, family history and personal medical history including previous neurological insult and occurrence of febrile seizures.

Definite dates of seizures occurring during defined time periods are recorded in SANAD II. In the SANAD II baseline assessment the 'date of first seizure' is recorded together with the total number of seizures prior to the baseline assessment date. However, the dates of all experienced seizures are not recorded. To ensure a comparable analysis, assessments of agreement with routine data have been calculated using only the recorded definite dates of first seizure.

The diagnosis and classification of seizure type and epilepsy syndrome are a mandatory requirement for enrolment in SANAD II. As such the date of 'randomisation', which correlates with the date of baseline assessment, is defined as the date of diagnosis in the SANAD II dataset. Participants are diagnosed based on the accepted definitions of epilepsy, proposed by the International League Against Epilepsy [7, 8]. In accordance with these guidelines, at least two unprovoked seizures occurring greater than 24 hours apart are required for diagnosis and enrolment in SANAD II. The guidelines also state participants may be diagnosed following a single seizure with a probability of subsequent seizures similar to the general recurrence risk following two unprovoked seizures.

However, calculating such probabilities is troublesome in both clinical practice and the research setting and as such two seizures are required for diagnosis and enrolment in SANAD II. During the diagnosis, participants' seizures are classified as 'focal', 'generalised' or 'unclassified'.

The clinical investigations Magnetic Resonance Imaging (MRI) Brain, Computed Tomography (CT) Brain and Electroencephalography (EEG) are recorded in SANAD II at the baseline assessment, although there is no protocol requirement for investigations to be completed. Details recorded in SANAD II for each participant include the investigation, date and result.

#### *5.13.1.2 The Follow-Up of Participants and Measurement of SANAD II Outcomes*

Following recruitment and completion of the SANAD II baseline assessment, 'follow-up assessments' are completed at three, six, 12 months and annually thereafter. During each follow-up assessment the clinician records data onto a CRF. Data recorded is summarised in *Box 5.3* and includes the dates of seizure occurrence and total number of seizures, medication history, adverse events, investigation results, personal medical history and details regarding healthcare resource use.

During the follow-up assessments the 'date of first seizure' and 'date of last seizure' during the defined time period between SANAD II assessments are recorded together with the total number of seizures experienced in this time period. However, the dates of all experienced seizures are not recorded. To ensure a comparable analysis in this study, assessments of agreement with routine data have been calculated using only the recorded definite dates of seizures. This record of seizures permits construction and measurement of a number of the outcomes in SANAD II, such as 'time to 12 month remission from seizures'.

The prescription of Antiepileptic Drugs (AEDs), including initial randomised AED and subsequently prescribed AEDs are recorded during the SANAD II assessments. The named AED, date of prescription and prescribed dosage are recorded together with reason for any dosage alteration. Participants are also requested to complete a number of self-report questionnaires which include details of prescribed AEDs.

Adverse events are recorded during the SANAD II follow-up assessments. Adverse events are defined as any untoward medical occurrence in a subject to whom a medicinal product has been administered, including occurrences which are not necessarily caused by or related to that product. The symptom, severity, date, AED in question, relationship to AED and action taken are recorded during follow-up assessments. Additional details and urgent reporting are required in cases of Serious Adverse Reactions (SAR) or Suspected Unexpected Serious Adverse Reactions (SUSAR). In addition to the assessment of adverse events during follow-up assessments, participants are requested at intervals to complete a number of self-report questionnaires including the *Adverse Events Profile*, providing an additional method for the recording of adverse events.

Episodes of participants' healthcare resource use are recorded in SANAD II and are informative to the health economic analyses, including cost effectiveness analyses. Healthcare resource use may be 'planned' or 'unplanned'. Planned attendances (elective attendances) include routine outpatient follow-up attendances, for example follow-up attendances with the Neurologist. Unplanned attendances (emergency attendances) include emergency, inpatient or primary care attendances, for example, attendances as a result of seizure occurrence or the experience of adverse events.

Planned attendances in SANAD II include the routine SANAD II assessments. Following the SANAD II baseline assessment, the follow-up assessments occur at specified intervals, within the context of routine clinical care. In the majority of participants, it is anticipated such attendances occur in the context of outpatient clinic attendances with the Neurologist. However, follow-up assessments may also occur in dedicated research clinics or opportunistically during unplanned attendances, although such methods would be expected to account for only a minority of SANAD II assessments.

Unplanned attendances in SANAD II may include emergency, inpatient and primary care attendances. Participants are requested to complete self-report questionnaires at specified intervals during follow-up, indicating if they have had any healthcare attendances 'in the last three months'. The healthcare setting and clinical reason for attendance are recorded through the options 'epilepsy-related' or 'non epilepsy-related'.

*Box 5.3 Summary: Data Recorded in SANAD II*

***The Identification of Eligible Individuals and Recruitment into SANAD II***

- Identifying Variables
  - o Name, DOB, NHS Number
- Seizure History
  - o Date of First Seizure, Total Number of Seizures
- Personal History
  - o Neurological Insult
  - o Febrile Seizures
- Family History, Epilepsy
- Investigations
  - o EEG and Imaging (CT or MRI)
- Randomised Antiepileptic Drug

***The Follow-Up of Participants and Measurement of SANAD II Outcomes***

- Review of Medical History
- Further Investigations
  - o EEG and Imaging (CT or MRI)
- Randomised Antiepileptic Drug
  - o Further Antiepileptic Drug Treatment
- Adverse Events
  - o Date, Nature, Severity, Relationship to AED
- Seizure Occurrence
  - o Date of First and Last Seizures Between Follow-Up Assessments
  - o Total Number of Seizures Between Follow-Up Assessments
- Healthcare Resource Use

### **5.13.2 The Routinely Recorded Data**

To permit the assessments of agreement, relevant routinely recorded data were extracted for each assessed data variable and outcome measure. For each variable an algorithmic approach was taken for the definition of relevant clinical events, combining knowledge of the coding systems, clinical behaviours and organisational pathways. An approach utilising the clinical interpretation of routinely recorded data has previously been used in studies assessing seizures [197] and in other disease areas in the UK [198-200].

### **5.13.3 The Identification of Seizure Occurrence in the Routinely Recorded Datasets**

The identification of the occurrence of unprovoked seizures is required for diagnosis of epilepsy and subsequent calculation of the outcomes in SANAD II. In order to permit an assessment of agreement, the occurrence of seizures must be identified and extracted from routinely recorded datasets. In this study an algorithmic approach was developed, informed by the approach taken in a published previous analysis of seizures recorded in NHS Digital Hospital Episode Statistics [197]. In this previous study the occurrence of seizures was identified in the HES Admitted Patient Care (APC) Dataset and subsequent specialist neurology follow-up in the HES Outpatient Dataset was examined. The occurrence of seizures was identified by review of the recorded ICD 10 codes, however, in the Admitted Patient Care Dataset; up to 15 ICD 10 codes are listed for each patient admission. These listed codes include the primary reason for attendance in addition to diagnoses included in an individual's past medical history. This previous study developed an algorithmic approach with the aim of identifying emergency admissions where the occurrence of seizure was the primary reason for attendance, rather than a diagnosis of epilepsy listed as part of the past medical history. 'Epilepsy' and 'seizure' codes were categorised as the primary reason for attendance, meaning the occurrence of seizures, when listed as the first diagnosis ('definite seizure') or second or third diagnosis with a 'probable' supportive code in the first diagnosis position ('probable seizure'). Attendances with epilepsy and seizure codes listed in the second or third diagnostic position with a 'possible' supportive code as the first diagnosis ('possible seizure') or an unrelated code as the first diagnosis ('definitely not seizure') were not categorised as the occurrence of seizure. *Figure 5.1* details the algorithmic approach and *Table 5.15* details the ICD 10 codes included in the definition of 'definite seizure' and 'probable seizure' in this previous study.

Figure 5.1: Algorithm for the Identification of Seizure Occurrence Using the HES Admitted Patient Care Dataset [197]

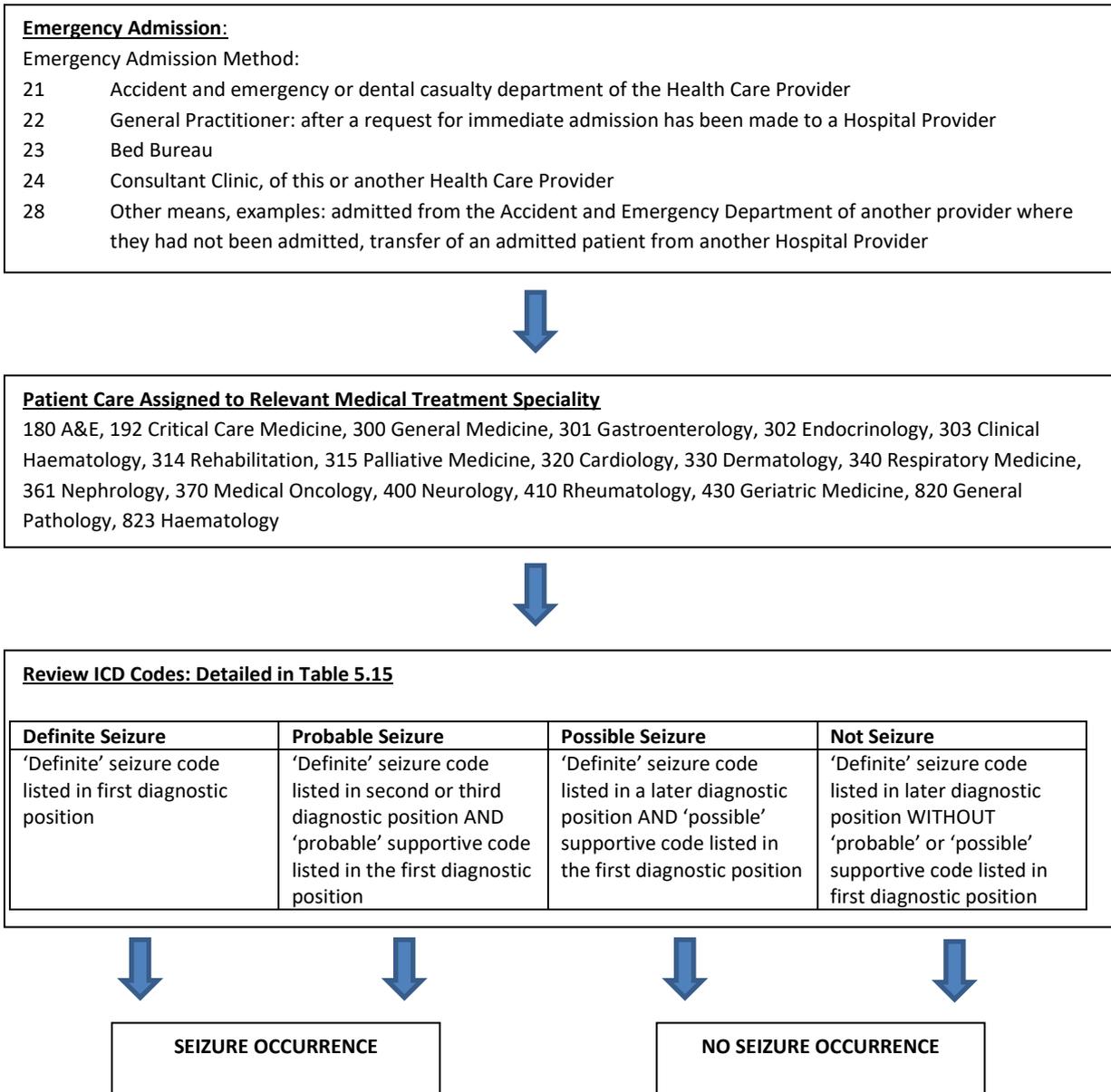


Table 5.15: ICD Codes Included in the Definitions of 'Definite' and 'Probable' Seizure Occurrence [197]

ICD 10 Code	ICD Code Description	Definite	Probable
G400	LOCAL-RELATED (PART) IDIOPATH EPILEP/EPILEP SYND WITH SEIZURE	1	0
G401	LOCAL-RELATED (PART) SYMPTOM EPILEPSY/EPILEPTIC SYND WITH SEIZURE	1	0
G402	LOCAL-RELATED (PART) SYMPTOM EPILEPSY/ EPILEP SYND	1	0
G403	GENERALIZED IDIOPATHIC EPILEPSY AND EPILEPTIC SYNDROMES	1	0
G404	OTHER GENERALIZED EPILEPSY AND EPILEPTIC SYNDROMES	1	0
G405	SPECIAL EPILEPTIC SYNDROMES	1	0
G406	GRAND MAL SEIZURES, UNSPECIFIED (WITH OR WITHOUT PETIT MAL)	1	0
G407	PETIT MAL, UNSPECIFIED, WITHOUT GRAND MAL SEIZURES	1	0
G408	OTHER EPILEPSY	1	0
G409	EPILEPSY, UNSPECIFIED	1	0
G410	GRAND MAL STATUS EPILEPTICUS	1	0
G411	PETIT MAL STATUS EPILEPTICUS	1	0
G412	COMPLEX PARTIAL STATUS EPILEPTICUS	1	0
G418	OTHER STATUS EPILEPTICUS	1	0
G419	STATUS EPILEPTICUS, UNSPECIFIED	1	0
R568	OTHER AND UNSPECIFIED CONVULSIONS	1	0
F019	VASCULAR DEMENTIA, UNSPECIFIED	0	1
F100	MENTAL AND BEHAVIOURAL DISORDERS DUE TO ACUTE INTOXICATION	0	1
F101	MENTAL AND BEHAVIOURAL DISORDERS DUE TO HARMFUL USE OF ALCOHOL	0	1
F102	MENTAL AND BEHAVIOURAL DISORDERS DUE TO ALCOHOL DEPENDENCE	0	1
F103	MENTAL AND BEHAVIOURAL DISORDERS DUE TO WITHDRAWAL OF ALCOHOL	0	1
F104	MENTAL AND BEHAVIOURAL DISORDERS AND DELIRIUM	0	1
F419	ANXIETY DISORDER, UNSPECIFIED	0	1
G439	MIGRAINE, UNSPECIFIED	0	1
H538	OTHER VISUAL DISTURBANCES	0	1
J690	PNEUMONITIS DUE TO FOOD AND VOMIT	0	1
K292	ALCOHOLIC GASTRITIS	0	1
R402	COMA, UNSPECIFIED	0	1
R410	DISORIENTATION, UNSPECIFIED	0	1
R418	OTHER & UNSPEC SYMPTOMS & SIGNS INVOLVING COGNITIVE FUNCTION	0	1
R42X	DIZZINESS AND GIDDINESS	0	1
R451	RESTLESSNESS AND AGITATION	0	1
R51X	HEADACHE	0	1
R55X	SYNCOPE AND COLLAPSE	0	1
R600	LOCALIZED OEDEMA	0	1
R798	OTHER SPECIFIED ABNORMAL FINDINGS OF BLOOD CHEMISTRY	0	1
S000	SUPERFICIAL INJURY OF SCALP	0	1
S001	CONTUSION OF EYELID AND PERIOCCULAR AREA	0	1
S008	SUPERFICIAL INJURY OF OTHER PARTS OF HEAD	0	1
S009	SUPERFICIAL INJURY OF HEAD, PART UNSPECIFIED	0	1
S010	OPEN WOUND OF SCALP	0	1
S018	OPEN WOUND OF OTHER PARTS OF HEAD	0	1
S019	OPEN WOUND OF HEAD, PART UNSPECIFIED	0	1
S099	UNSPECIFIED INJURY OF HEAD	0	1
S308	OTHER SUPERFICIAL INJURIES OF ABDOMEN, LOWER BACK AND PELVIS	0	1
Z038	OBSERVATION FOR OTHER SUSPECTED DISEASES AND CONDITIONS	0	1
Z739	PROBLEM RELATED TO LIFE-MANAGEMENT DIFFICULTY, UNSPECIFIED	0	1

The algorithmic approach developed for this research was informed by the algorithm used in the discussed study yet is more complex, for the following reasons:

**Multiple Diagnostic Coding Systems are Included:**

ICD 10 codes are recorded in the HES Admitted Patient Care Dataset, HES Outpatient Dataset, SAIL Patient Episode Database for Wales and SAIL Outpatient Dataset and are comparable. However, in addition diagnostic codes using a simplified coding system are included in the HES Accident & Emergency Dataset and SAIL Emergency Department Dataset and for a proportion of patients READ codes are recorded in the SAIL Primary Care Dataset. The inclusion of a number of datasets and coding systems necessitated the inclusion of an increased number of codes to define the occurrence of seizures.

**A Single Attendance may be Recorded in More Than One Dataset:**

For example, patients may attend the emergency department and subsequently be admitted, with the single attendance recorded in both datasets. This necessitated a system for stratifying the diagnostic codes. A clinical interpretation was used with codes providing the 'greatest quality', defined as greatest diagnostic detail and specificity, contributing data to the routinely recorded dataset used for the analyses. For example, the ICD code G412 (Complex Partial Status Epilepticus) provides greater diagnostic detail than the HES Accident and Emergency code 241 (CNS Conditions – Epilepsy). In this case the ICD code will be included in the dataset used for the analyses. However, the total number of relevant codes from all datasets for each variable was also presented.

**The Occurrence of Seizures was Identified Before and Following Diagnosis of Epilepsy:**

Routinely recorded data were available for the time periods before and following diagnosis of epilepsy and participant enrolment in SANAD II. The identification of the occurrence of seizures before diagnosis and following diagnosis must be appropriate based on the diagnostic codes available. Diagnostic codes specifying 'epilepsy' or 'seizures' were appropriate for inclusion in the identification of seizures both before and following diagnosis. Diagnostic codes with no specific mention of epilepsy or seizures, specifically the emergency codes 'CNS Disorder' and 'CNS Condition – Unspecified' are not appropriate to identify seizure occurrence prior to diagnosis of epilepsy. However, following clinical discussion and to ensure a sensitive approach, such codes would be sufficiently likely to be recorded and appropriate to identify the occurrence of seizures following diagnosis. The total numbers of seizures identified by such codes for each data variable are presented.

**Multiple Variables Were Assessed, Including the Occurrence of Seizures and Diagnosis and Classification of Epilepsy and Seizures:**

Multiple variables and outcome measures were assessed using the record of 'epilepsy' and 'seizure' codes. In addition to the occurrence of seizures, agreement was assessed for the date of diagnosis of epilepsy and subsequently classification of seizure type (Generalised / Focal / Unclassified). This necessitated development of an additional algorithm using the recorded codes for the identification of 'diagnosis of epilepsy' as a separate entity to the 'occurrence of seizures'.

**A Relevant Code may not Indicate the Occurrence of Seizures in some Datasets:**

This complexity refers to the Outpatient and Primary Care Datasets. An 'epilepsy' or 'seizure' code in the Outpatient Dataset is very unlikely to indicate the occurrence of seizure, although was informative to the diagnosis of epilepsy and classification of seizure type. An 'epilepsy' or 'seizure' code in the Primary Care Dataset can be considered to represent occurrence of seizure, diagnosis of epilepsy and classification of seizure type. The exception is when the date the code is recorded correlates within one month with the date of a relevant attendance in a specialist outpatient clinic. In this case, the General Practitioner is likely to have received clinical correspondence from the specialist clinic and recorded the relevant codes in the patient's electronic medical records. The date recorded may be the date of specialist clinic review or the date the correspondence was received. Such instances were informative to the diagnosis of epilepsy, but were not representative of the occurrence of seizures.

### *The Identification of Seizure Occurrence in the Routinely Recorded Datasets*

In this study, the identification of seizure occurrence in routinely recorded datasets involved the review of specified diagnostic codes, as follows:

- **Inpatient Datasets:**  
**HES Admitted Patient Care Dataset and SAIL Patient Episode Database for Wales**  
Attendances with ICD 10 codes defined as representing 'definite seizure' and 'probable seizure' in the HES Admitted Patient Care and SAIL Patient Episode Database for Wales Datasets were included as representing seizure occurrence
- **Outpatient Datasets:**  
**HES Outpatient and SAIL Outpatient Datasets**  
ICD 10 codes recorded in the HES and SAIL Outpatient Datasets were not included in the assessment of seizure occurrence
- **Emergency Datasets:**  
**HES Accident & Emergency and SAIL Emergency Department Datasets**  
Diagnostic data is recorded in only a minority of attendances and rarely exceeds a single diagnosis in the HES Accident & Emergency and SAIL Emergency Department Datasets Defined codes listed in the first diagnostic position represented seizure occurrence
- **Primary Care Dataset:**  
**SAIL Primary Care Dataset**  
Specified READ codes recorded in the SAIL Primary Care Dataset represented seizure occurrence unless the date correlated within one month with a date of neurological specialist assessment in the Outpatient Datasets

*Figure 5.2* details the algorithmic approach used in this study. *Tables 5.15 and 5.16* detail the diagnostic codes included in the identification of seizure occurrence for the participants in this study.

### *The Assessment of Seizure Occurrence Before and Following Diagnosis of Epilepsy*

The assessment of seizures in participants before diagnosis of epilepsy did not include the emergency codes 'CNS Disorder' and 'CNS Condition – Unspecified'. However, following clinical discussion and to ensure a sensitive approach, such codes would be sufficiently likely to be recorded and appropriate to identify the occurrence of seizures following diagnosis of epilepsy. The total numbers of seizures identified by such codes for each data variable are presented.

### *The Assessment of Relevant Attendances, Not Meeting the Criteria for Seizure Occurrence*

During the assessment of seizure occurrence, the dates of seizures recorded in the SANAD II dataset were used to identify attendances in the routinely recorded datasets within 48 hours of the date of SANAD II recorded seizure. Such attendances may have been inadequately coded or recorded with codes not meeting the criteria for seizure occurrence. The total numbers of such attendances for each data variable are presented.

Figure 5.2: Algorithm for the Identification of Seizure Occurrence

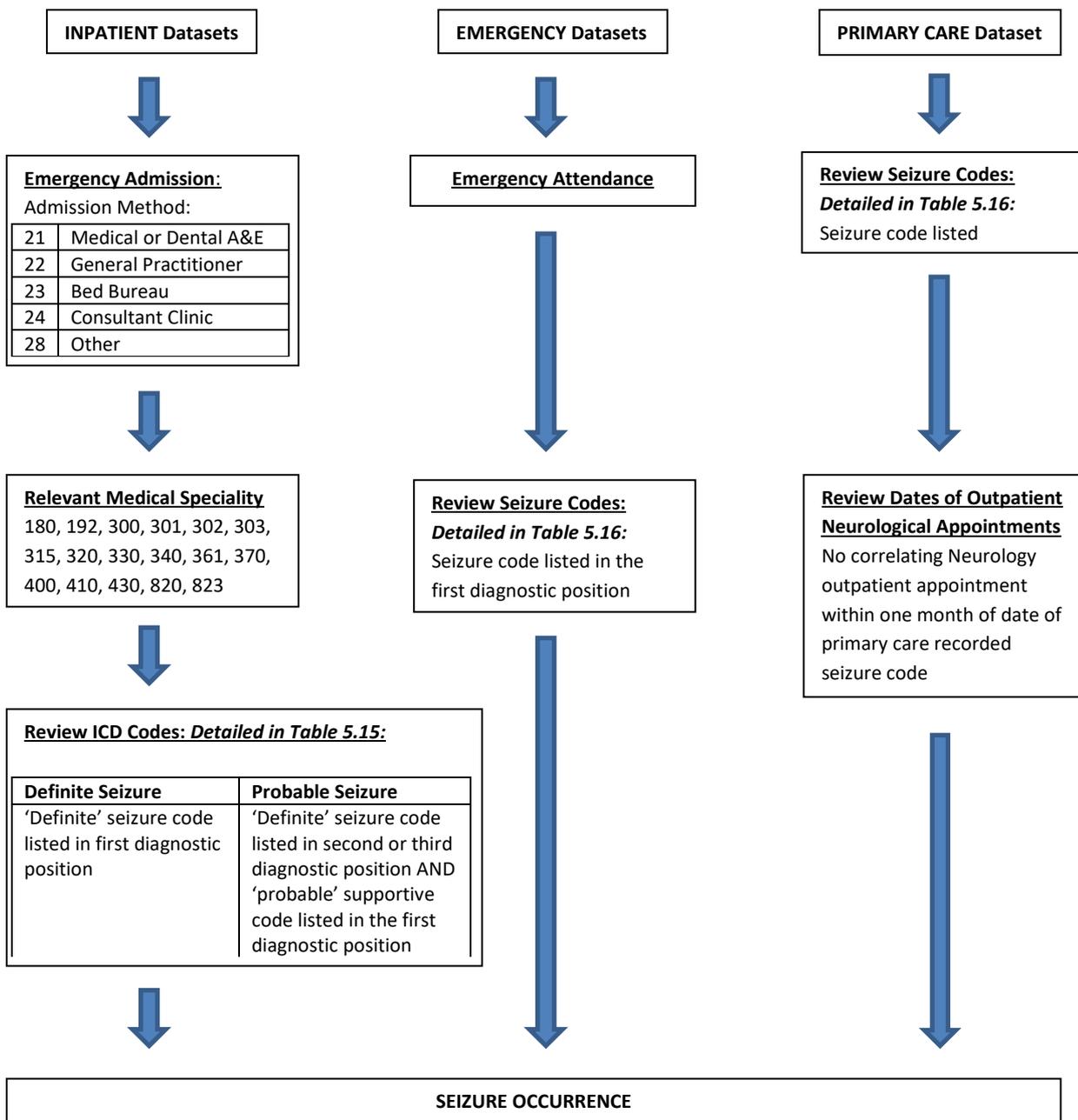


Table 5.16: Emergency and READ Codes Included in the Definition of Seizure Occurrence

Code	Code Description
<b>Emergency Codes</b>	
HES: 241	CNS Conditions - Epilepsy
HES: 24	CNS Disorder*
SAIL: 17A	Seizure / Convulsion
SAIL: 17Z	CNS Condition – Unspecified*
<b>Primary Care READ Codes</b>	
<b>'Seizure' Codes</b>	
1B1W.00	Transient epileptic amnesia
1B27.00	Seizures in response to acute event
1B64.00	Had a convulsion
1B64.11	Convulsion - symptom
282..00	O/E - fit/convulsion
282..11	O/E - a convulsion
282..13	O/E - a seizure
2828	Absence seizure
282Z.00	O/E - fit/convulsion NOS
667D.00	Epilepsy control poor
667T.00	Daily seizures
667V.00	Many seizures a day
667W.00	Emergency epilepsy treatment since last appointment
F132z12	Myoclonic seizure
F250011	Epileptic absences
F250200	Epileptic seizures - atonic
F250300	Epileptic seizures - akinetic
F251200	Epileptic seizures - clonic
F251300	Epileptic seizures - myoclonic
F251400	Epileptic seizures - tonic
F251600	Grand mal seizure
F253.11	Status epilepticus
F254400	Epileptic automatism
F254500	Complex partial epileptic seizure
F255600	Simple partial epileptic seizure
F25H.00	Generalised seizure
F25X.00	Status epilepticus, unspecified
F25y300	Complex partial status epilepticus
F25z.11	Fit (in known epileptic) NOS
Fyu5200	[X]Other status epilepticus
Fyu5900	[X]Status epilepticus, unspecified
R003.00	[D]Convulsions
R003400	[D]Nocturnal seizure
R003y00	[D]Other specified convulsion
R003z00	[D]Convulsion NOS
R003z11	[D]Seizure NOS
Ryu7100	[X]Other and unspecified convulsions
1B63.00	Had a fit
1B63.11	Fit - had one, symptom
282..12	O/E - a fit
2822	O/E - grand mal fit
2823	O/E - petit mal fit
2824	O/E - focal (Jacksonian) fit
2824.11	O/E - Jacksonian fit
2824.12	O/E - focal fit
2825	O/E - psychomotor fit

R003200	[D]Fit
F252.00	Petit mal status
F253.00	Grand mal status
<b>'Epilepsy' Codes</b>	
1030.00	Epilepsy confirmed
667B.00	Nocturnal epilepsy
F035200	Rasmussen syndrome
F132100	Progressive myoclonic epilepsy
F132111	Unverricht - Lundborg disease
F25..00	Epilepsy
F250.00	Generalised non-convulsive epilepsy
F250000	Petit mal (minor) epilepsy
F250100	Pykno-epilepsy
F250400	Juvenile absence epilepsy
F250y00	Other specified generalised non-convulsive epilepsy
F250z00	Generalised non-convulsive epilepsy NOS
F251.00	Generalised convulsive epilepsy
F251000	Grand mal (major) epilepsy
F251011	Tonic-clonic epilepsy
F251500	Tonic-clonic epilepsy
F251y00	Other specified generalised convulsive epilepsy
F251z00	Generalised convulsive epilepsy NOS
F254.00	Partial epilepsy with impairment of consciousness
F254000	Temporal lobe epilepsy
F254100	Psychomotor epilepsy
F254200	Psychosensory epilepsy
F254300	Limbic system epilepsy
F254z00	Partial epilepsy with impairment of consciousness NOS
F255.00	Partial epilepsy without impairment of consciousness
F255000	Jacksonian, focal or motor epilepsy
F255011	Focal epilepsy
F255012	Motor epilepsy
F255100	Sensory induced epilepsy
F255200	Somatosensory epilepsy
F255300	Visceral reflex epilepsy
F255311	Partial epilepsy with autonomic symptoms
F255400	Visual reflex epilepsy
F255500	Unilateral epilepsy
F255y00	Partial epilepsy without impairment of consciousness OS
F255z00	Partial epilepsy without impairment of consciousness NOS
F25A.00	Juvenile myoclonic epilepsy
F25B.00	Alcohol-induced epilepsy
F25C.00	Drug-induced epilepsy
F25D.00	Menstrual epilepsy
F25E.00	Stress-induced epilepsy
F25F.00	Photosensitive epilepsy
F25y.00	Other forms of epilepsy
F25y000	Cursive (running) epilepsy
F25y100	Gelastic epilepsy
F25yz00	Other forms of epilepsy NOS
F25z.00	Epilepsy NOS
Fyu5000	[X]Other generalized epilepsy and epileptic syndromes
Fyu5100	[X]Other epilepsy
SC20000	Traumatic epilepsy

**Legend:** \*: Codes excluded from the assessment of seizures prior to participant diagnosis of epilepsy

**5.13.4 Data Variables Relevant to the Identification of Eligible Individuals and Recruitment into SANAD II**

**5.13.4.1 Seizure Occurrence: Baseline Variables**

Applicable to all variables involving dates of seizure occurrence, a one month (30 days) acceptable clinical limit of agreement was specified *a priori*. This limit was informed by clinical discussion and represents a level inclusive of errors in participant recall together with a clinically acceptable level of disagreement. The baseline variables assessed are detailed in *Table 5.17*.

*Table 5.17: Seizure Occurrence: Baseline Variables*

<b>Variable</b>	<b>Identification in SANAD II</b>	<b>Identification in Routine Data</b>
<p><b>Date of First Seizure</b></p> <p>Defined as the first seizure occurring at any time prior to the baseline SANAD II assessment date</p>	<p>The recorded date of ‘first seizure’</p>	<p>The first identified seizure occurrence prior to the baseline SANAD II assessment date.</p> <p>Thirty six participants (36.7%) had a ‘first seizure’ recorded in the SANAD II dataset occurring prior to 2013. This group of participants were excluded from this assessment due to the lack of availability of routine data coverage for the time period before 2013.</p>
<p><b>Date of First Tonic-Clonic Seizure</b></p> <p>Defined as the first tonic-clonic seizure occurring at any time prior to the baseline SANAD II assessment date</p>	<p>The recorded date of ‘first tonic-clonic seizure’</p>	<p>The first identified tonic-clonic seizure occurrence prior to the baseline SANAD II assessment date.</p> <p>Diagnostic codes were defined <i>a priori</i> to indicate the ‘occurrence of tonic-clonic seizures’. Informed by clinical discussion and to ensure the approach was sensitive, non-specific codes such as ‘Had a Fit’ and ‘Epilepsy Unspecified’ were included to represent tonic-clonic seizures. It was deemed more likely that patients seek medical attention following a tonic-clonic seizure and in such cases, non-specific codes may be recorded.</p> <p>Thirty six participants (36.7%) had a ‘first seizure’ recorded in the SANAD II dataset occurring prior to 2013. This group of participants were excluded from this assessment due to the lack of availability of routine data coverage for the time period before 2013.</p>

#### 5.13.4.2 *The Diagnosis and Classification of Epilepsy and Seizures in the Routinely Recorded Datasets*

The identification of the diagnosis and subsequently classification of epilepsy and seizures in routinely recorded datasets involved the review of specified diagnostic ‘epilepsy’ codes and codes included in the identification of seizure occurrence. The diagnosis of epilepsy in routine datasets is defined as the recording of a single code consistent with a ‘diagnosis of epilepsy’ or the recording of two codes consistent with the ‘occurrence of seizures’. The diagnosis is identified in the routinely recorded datasets as follows:

- **Inpatient Datasets:**

  - HES Admitted Patient Care Dataset and SAIL Patient Episode Database for Wales**

  - Attendance with a single ICD 10 code representing diagnosis of epilepsy or two attendances with codes representing seizure occurrence (‘definite seizure’ or ‘probable seizure’)

- **Outpatient Datasets:**

  - HES Outpatient and SAIL Outpatient Datasets**

  - Review by a neurologist with single ICD 10 code representing diagnosis of epilepsy

- **Emergency Datasets:**

  - HES Accident & Emergency and SAIL Emergency Department Datasets**

  - Two attendances with codes representing seizure occurrence

- **Primary Care Dataset:**

  - SAIL Primary Care Dataset**

  - Single READ code representing diagnosis of epilepsy or two READ codes representing seizure occurrence

For participants meeting the criteria for a diagnosis of epilepsy, the classification of seizure type was determined from a clinical interpretation of the recorded codes. Where ‘focal’ or ‘generalised’ seizures could not be specified, participants were deemed ‘unclassified’.

*Figure 5.3* details the algorithmic approach and *Tables 5.18 and 5.19* detail the codes included in the diagnosis and classification of epilepsy and seizures relevant to the adult participants in this study. The codes included in the definition of seizures have been presented in *Tables 5.15 and 5.16*.

Figure 5.3: Algorithm for the Diagnosis of Epilepsy

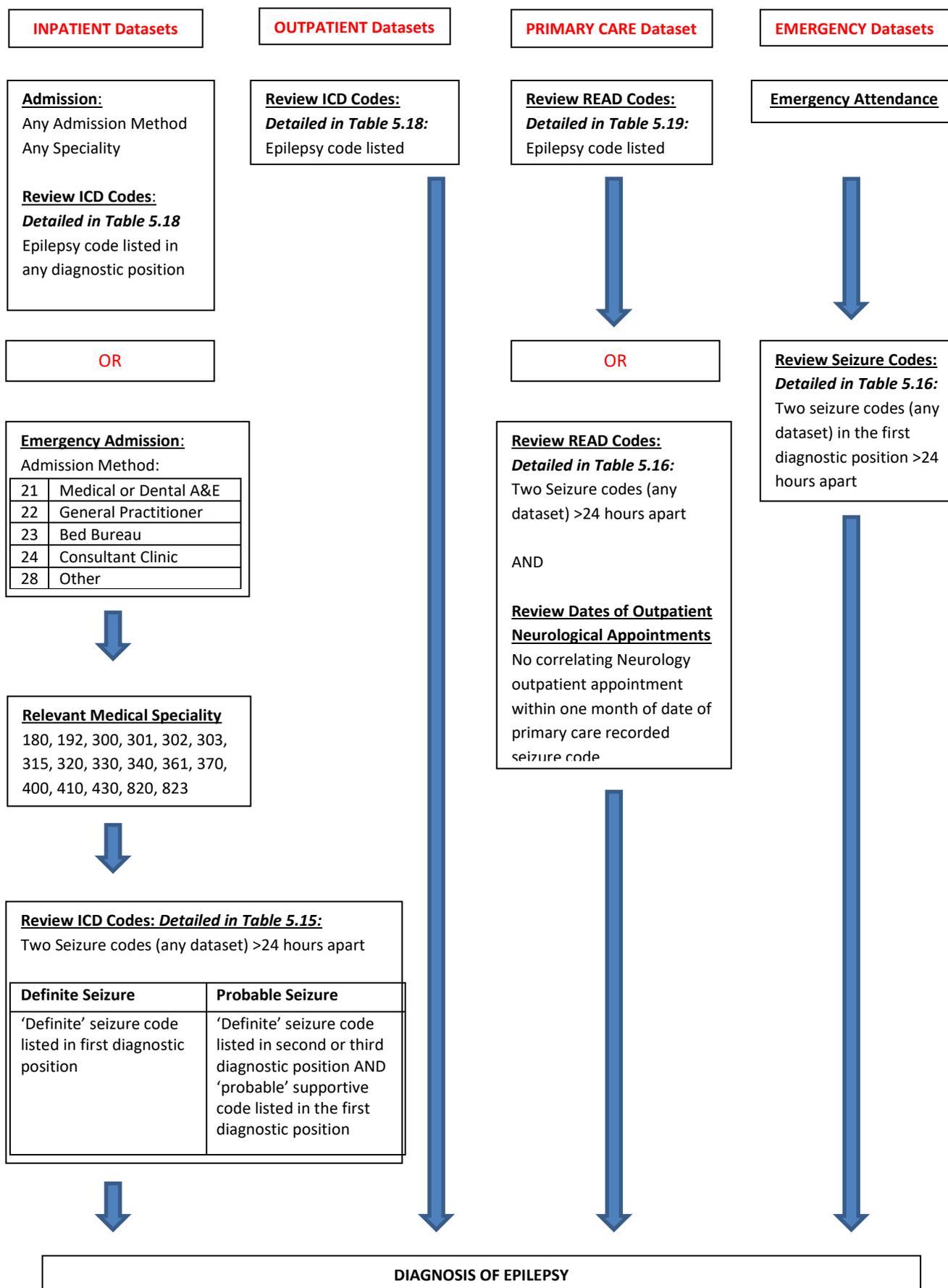


Table 5.18: ICD Codes Included in the Definition of Diagnosis of Epilepsy

ICD 10 Code	ICD Code Description
<b>Focal Epilepsy</b>	
G400	LOCAL-RELATED (PART) IDIOPATH EPILEP/EPILEP SYND WITH SEIZURE
G401	LOCAL-RELATED (PART) SYMPTOM EPILEPSY/EPILEPTIC SYND WITH SEIZURE
G402	LOCAL-RELATED (PART) SYMPTOM EPILEPSY/ EPILEP SYND
<b>Generalised Epilepsy</b>	
G403	GENERALIZED IDIOPATHIC EPILEPSY AND EPILEPTIC SYNDROMES
G404	OTHER GENERALIZED EPILEPSY AND EPILEPTIC SYNDROMES
<b>Unclassified Epilepsy</b>	
G40	EPILEPSY AND RECURRENT SEIZURES
G405	SPECIAL EPILEPTIC SYNDROMES
G408	OTHER EPILEPSY
G409	EPILEPSY, UNSPECIFIED

Table 5.19: READ Codes Included in the Definition of Diagnosis of Epilepsy

Code	Code Description
<b>Focal Epilepsy</b>	
F035200	Rasmussen syndrome
F254.00	Partial epilepsy with impairment of consciousness
F254000	Temporal lobe epilepsy
F254100	Psychomotor epilepsy
F254200	Psychosensory epilepsy
F254300	Limbic system epilepsy
F254z00	Partial epilepsy with impairment of consciousness NOS
F255.00	Partial epilepsy without impairment of consciousness
F255000	Jacksonian, focal or motor epilepsy
F255011	Focal epilepsy
F255200	Somatosensory epilepsy
F255311	Partial epilepsy with autonomic symptoms
F255500	Unilateral epilepsy
F255y00	Partial epilepsy without impairment of consciousness OS
F255z00	Partial epilepsy without impairment of consciousness NOS
F25y100	Gelastic epilepsy
F25y000	Cursive (running) epilepsy
SC20000	Traumatic epilepsy
<b>Generalised Epilepsy</b>	
F250.00	Generalised non-convulsive epilepsy
F250100	Pykno-epilepsy
F250400	Juvenile absence epilepsy
F250y00	Other specified generalised non-convulsive epilepsy
F250z00	Generalised non-convulsive epilepsy NOS
F251.00	Generalised convulsive epilepsy
F251y00	Other specified generalised convulsive epilepsy
F251z00	Generalised convulsive epilepsy NOS
F132100	Progressive myoclonic epilepsy

F132111	Unverricht - Lundborg disease
F25A.00	Juvenile myoclonic epilepsy
Fyu5000	[X]Other generalized epilepsy and epileptic syndromes
<b>Unclassified Epilepsy</b>	
1030.00	Epilepsy confirmed
667B.00	Nocturnal epilepsy
F250000	Petit mal (minor) epilepsy
F251000	Grand mal (major) epilepsy
F25..00	Epilepsy
F251011	Tonic-clonic epilepsy
F251500	Tonic-clonic epilepsy
F255100	Sensory induced epilepsy
F255300	Visceral reflex epilepsy
F255400	Visual reflex epilepsy
F255012	Motor epilepsy
F25B.00	Alcohol-induced epilepsy
F25C.00	Drug-induced epilepsy
F25D.00	Menstrual epilepsy
F25E.00	Stress-induced epilepsy
F25F.00	Photosensitive epilepsy
F25y.00	Other forms of epilepsy
F25yz00	Other forms of epilepsy NOS
F25z.00	Epilepsy NOS
Fyu5100	[X]Other epilepsy

#### 5.13.4.3 Diagnosis and Classification of Epilepsy and Seizures: Variables

An acceptable clinical limit of agreement for the date of diagnosis of epilepsy was specified *a priori* at one month (30 days). This limit was informed by clinical discussion and represents a pragmatic level inclusive of administrative procedures in clinical practice. For example, following a diagnosis during an outpatient assessment, the General Practitioner may not receive correspondence for two weeks at which time a relevant READ code may be recorded. This rationale also explains the one month period permitted following the baseline assessment date, in the 'baseline' diagnosis. The variables assessed are detailed in *Table 5.20*.

Table 5.20: Diagnosis and Classification of Epilepsy and Seizures: Variables

<b>Variable</b>	<b>Identification in SANAD II</b>	<b>Identification in Routine Data</b>
<p><b>Date of Baseline Diagnosis of Epilepsy</b></p> <p>Defined as the first occurrence of a 'diagnosis' at baseline</p>	<p>The recorded date of the SANAD II baseline assessment</p>	<p>Defined as the first occurrence of a 'diagnosis' at any time prior to the baseline SANAD II assessment or within one month subsequently. Thirty six participants (36.7%) had a 'first seizure' recorded in the SANAD II dataset occurring prior to 2013. Twenty of this group (55.5%), not meeting the criteria for a diagnosis of epilepsy were censored from this assessment due to the lack of availability of routine data coverage for the time period before 2013, removing the potential for two seizures to be recorded. This approach ensured a fair comparison and the maximal inclusion of data.</p>
<p><b>Date of All-Time Diagnosis of Epilepsy</b></p> <p>Defined as the first occurrence of a 'diagnosis' at any time</p>	<p>The recorded date of the SANAD II baseline assessment</p>	<p>Defined as the first occurrence of a 'diagnosis' at any time. Thirty six participants (36.7%) had a 'first seizure' recorded in the SANAD II dataset occurring prior to 2013. Seventeen of this group (47.2%), not meeting the criteria for a diagnosis of epilepsy were excluded from this assessment due to the lack of availability of routine data coverage for the time period before 2013, removing the potential for two seizures to be recorded. This approach ensured a fair comparison and the maximal inclusion of data.</p>
<p><b>Baseline Classification of Seizure</b></p> <p>Defined as the 'classification' of seizure in participants meeting the criteria for a baseline diagnosis of epilepsy</p>	<p>The classification (focal, generalised, unclassified) recorded in SANAD II</p>	<p>The classification (focal, generalised, unclassified) derived from the greatest quality recorded seizure and epilepsy diagnostic codes in participants meeting the criteria for a baseline diagnosis of epilepsy. Classification is only assessed in participants meeting the criteria for diagnosis, to ensure a fair comparison to the data recorded in SANAD II.</p>
<p><b>All-Time Classification of Seizure</b></p> <p>Defined as the 'classification' of seizure in participants meeting the criteria for an all-time diagnosis of epilepsy</p>	<p>The classification (focal, generalised, unclassified) recorded in SANAD II</p>	<p>The classification (focal, generalised, unclassified) derived from the greatest quality recorded seizure and epilepsy diagnostic codes in participants meeting the criteria for an all-time diagnosis of epilepsy. Classification is only assessed in participants meeting the criteria for diagnosis, to ensure a fair comparison to the data recorded in SANAD II.</p>

#### *5.13.4.4 The Assessment of Clinical Investigations in the Routinely Recorded Datasets*

Limited data regarding clinical investigations were available in the routinely recorded datasets. The SAIL primary care dataset included all READ codes recorded for each participant for the time period 2013-present. Although MRI, CT or EEG will not be performed in primary care, it was reasonable to expect codes to be entered into participants' electronic medical records on receipt of correspondence from secondary care. READ codes exist for the investigations and results and the primary care dataset had the greatest potential for detailed informative data. Data regarding MRI and CT were available from the HES Accident and Emergency and SAIL Emergency Department Datasets. The record of 'MRI' or 'CT' during an attendance is recorded, but the anatomical site of the imaging is not recorded. Additionally, the investigation results are not recorded and there was no available code for 'EEG'. In both the inpatient and outpatient datasets, despite the availability of ICD codes for the relevant procedures, such codes were not included in the data fields provided in the HES and SAIL datasets.

As a result of the limitations in the routinely recorded data, the 'status' of clinical investigations were compared in this study. Status defines whether an investigation has been 'Performed' or 'Not Performed'. However, narrative analysis of the data regarding the results of the investigations has also been included where this was available.

Identification of the status of the investigations in routinely recorded datasets involved the review of specified investigation codes, defined as follows:

*CT Brain and MRI Brain:*

- **Emergency Datasets:**

- HES Accident & Emergency and SAIL Emergency Department Datasets**

- Record of a relevant code indicating 'MRI' or 'CT' during an attendance with a previously defined episode of seizure occurrence

- Following clinical discussion and as a result of the lack of record of anatomical site, relevant codes recorded during attendances not meeting the criteria for seizure occurrence were not sufficiently likely to represent imaging of the brain

- **Primary Care Dataset:**

- SAIL Primary Care Dataset**

- READ codes representing CT Brain or MRI Brain

*EEG:*

- **Primary Care Dataset:**

- SAIL Primary Care Dataset**

- READ codes representing EEG

An algorithmic approach, detailed in *Figure 5.4* was developed to identify the status of clinical investigations and data were compared to the status recorded in SANAD II. *Table 5.21* details the codes included in the assessment of the status of clinical investigations.

Figure 5.4: Algorithm for the Assessment of Clinical Investigations

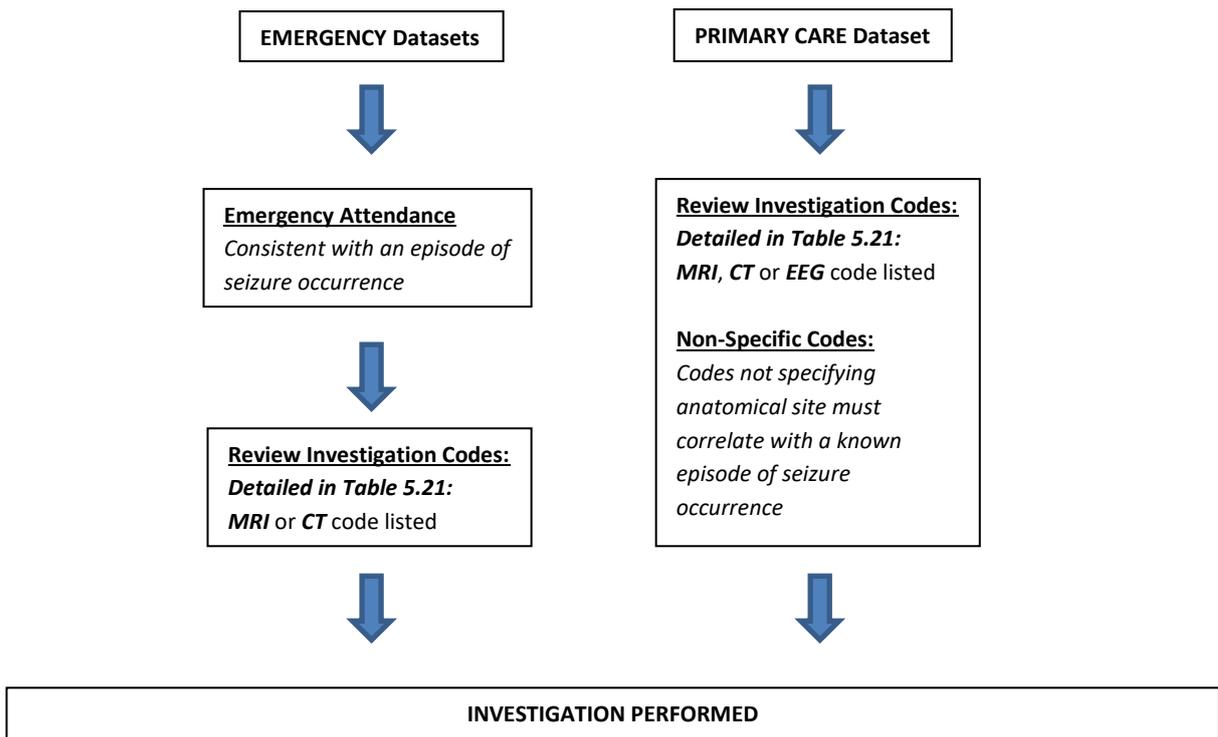


Table 5.21: Investigation Codes Included in the Assessment of Clinical Investigations

Code	Code Description
<b>CT Brain</b>	
HES: 12	Computed Tomography
SAIL: 201	Computed Tomography
READ: 567	Computed Tomography
READ: Y72JC	CT Head
READ: Y72JD	CT Brain
READ: YAYCE	CT of Bone Structures and Cavities of the Head
READ: YAYCB	CT Bone Structures of the Head
READ: YAMGZ	CT Brain Normal
READ: YAQSS	CT Brain Abnormal
READ: 5674	CT Skull
<b>MRI Brain</b>	
HES: 11	Magnetic Resonance Imaging
SAIL: 202	Magnetic Resonance Imaging
READ: 569	Magnetic Resonance: (Imaging) or (Study)
READ: Y7212	MRI of Head
READ: Y7213	MRI of Brain
READ: Y7215	MRI of Brain with Functional Imaging
READ: YB095	MRI Scan Abnormal
READ: YB088	MRI Scan Normal
READ : 5692	Nuclear Magnetic Resonance Scan: Normal
READ: 5693	Nuclear Magnetic Resonance Scan: Abnormal
<b>EEG</b>	
X77i2	Scalp EEG
X77i8	Sleep EEG
X77iL	Video EEG
X77iM	EEG telemetry
X77JD	Asymmetric EEG
X77ii	Ambulatory EEG
X77i8	EEG observations
X77lr	Focal EEG pattern
X77jK	Intraoperative EEG
X77iL	EEG video telemetry
X77ii	AEEG - Ambulatory EEG
X77lr	Localised EEG pattern
X77i9	Generalised EEG pattern
X77ij	Continuous processed EEG
X77i9	Sleep EEG - natural sleep
Xa0ej	Continuous EEG measurements
X77i4	EEG with photic stimulation
70650	EEG - Electroencephalography
X77i6	EEG with drug administration
X77J1	Focal episodic EEG abnormality
X77iz	Focal reduction of EEG activity
X77is	Focal continuous EEG abnormality
X77iA	Sleep EEG - sleep-deprived patient
31130	EEG normal
XM18c	EEG abnormal
X77IA	EEG artefact
X77JH	Generalised EEG frequency asymmetry

X77JG	Generalised EEG amplitude asymmetry
X77Ig	Generalised episodic EEG abnormality
X77II	EEG pattern of uncertain significance
X77Ig	Generalised paroxysmal EEG abnormality
X77IQ	Generalised continuous EEG abnormality
X77i7	EEG during special activation procedure
X77IQ	Generalised non-paroxysmal EEG abnormality
X77i5	EEG with over breathing and photic stimulation
X77IM	Subclinical rhythmical EEG discharges in adults
X77IM	SREDA - Subclinical rhythmical EEG discharges in adults
70650	Electroencephalography
XaPpX	Electroencephalography NEC
70650	EEG - Electroencephalography
XM18c	Electroencephalogram abnormal
X77iM	Electroencephalograph telemetry

### **5.13.5 Data Variables and Outcomes Relevant to the Follow-Up in SANAD II**

#### **5.13.5.1 Seizure Occurrence: Follow-Up Variables**

Applicable to all variables involving dates of seizure occurrence, a one month (30 days) acceptable clinical limit of agreement was specified *a priori*. This limit was informed by clinical discussion and represents a level inclusive of errors in participant recall together with a clinically acceptable level of disagreement. The follow-up variables assessed are detailed in *Table 5.22*.

Table 5.22: Seizure Occurrence: Follow-Up Variables

<b>Variable / Outcome Measure</b>	<b>Identification in SANAD II</b>	<b>Identification in Routine Data</b>
<p><b>Date of First Follow-Up Seizure</b></p> <p>Defined as the first seizure occurrence for the time period from the date of SANAD II randomisation until the final date routine data from all sources was available (31/12/15) or the date of last SANAD II follow-up assessment if the last follow-up assessment occurred prior to 31/12/15.</p> <p>This approach ensured the routine and SANAD II datasets were matched and permitted a fair comparison and the maximal inclusion of available data.</p>	<p>The first seizure recorded following the date of SANAD II randomisation.</p>	<p>The first seizure occurrence identified following the date of SANAD II randomisation.</p>
<p><b>Time to First Follow-Up Seizure</b></p> <p>Defined as the time in days between the date of SANAD II randomisation and the date of first follow-up seizure.</p> <p>Time to first follow-up seizure has been constructed using the 'date of first follow-up seizure' recorded in the SANAD II and identified in the routine datasets, the difference assessed using survival analysis. Participants not experiencing first follow-up seizure were censored using the final date routine data from all sources was available (31/12/15) or the date of last follow-up assessment if the last follow-up assessment occurred prior to 31/12/15. This approach ensured the routine and SANAD II datasets were matched and permitted a fair comparison and the maximal inclusion of available data.</p>		
<p><b>Date of First Follow-Up Tonic-Clonic Seizure</b></p> <p>Defined as the first tonic-clonic seizure occurrence for the time period from the date of SANAD II randomisation until the final date routine data from all sources was available (31/12/15) or the date of last SANAD II follow-up assessment if the last follow-up assessment occurred prior to 31/12/15.</p> <p>This approach ensured the routine and SANAD II datasets were matched and permitted a fair comparison and the maximal inclusion of available data.</p>	<p>The first tonic-clonic seizure recorded following the date of SANAD II randomisation.</p>	<p>The first tonic-clonic seizure occurrence identified following the date of SANAD II randomisation.</p> <p>Diagnostic codes were defined <i>a priori</i> to indicate the 'occurrence of tonic-clonic seizures'. Informed by clinical discussion and to ensure the approach was sensitive, non-specific codes such as 'Had a Fit' and 'Epilepsy Unspecified' were included to represent tonic-clonic seizures.</p>

<p><b>Date 12 Month Remission Achieved</b></p> <p>Defined as the date that remission from seizures has been achieved for a continuous period of 12 months following the date of SANAD II randomisation.</p> <p>For 12 month remission to be achieved there must be at least 12 months of available routine data and a SANAD II follow-up assessment at least 12 months after the date of last seizure. This approach ensured the routine and SANAD II datasets were matched and permitted a fair comparison and the maximal inclusion of available data.</p>	<p>The date that remission from seizures has been achieved for a continuous period of 12 months following the date of SANAD II randomisation.</p> <p>In the SANAD II Dataset, for a date of 12 month remission to be specified, there must be a SANAD II follow-up assessment at least 12 months after the date of last seizure.</p> <p>To ensure a fair comparison, participants in the SANAD II dataset achieving 12 month remission after 31/12/15 will be deemed not to have achieved 12 month remission as a result of the lack of available routine data.</p>	<p>The date that remission from seizures has been achieved for a continuous period of 12 months following the date of SANAD II randomisation.</p> <p>In the routine dataset, for a date of 12 month remission to be specified, there must be at least 12 months of available routine data after the date of last seizure.</p> <p>To ensure a fair comparison, participants in the routine dataset achieving 12 month remission after the date of last SANAD II follow-up assessment, if this last assessment occurred before 31/12/15, will be deemed not to have achieved remission.</p>
<p><b>Time to 12 Month Remission</b></p> <p>Defined as the time in days between the date of SANAD II randomisation and the date 12 month remission is achieved.</p> <p>Time to 12 month remission has been constructed using the 'date 12 month remission is achieved' calculated from the SANAD II and routine datasets, the difference assessed using survival analysis. Participants not achieving remission were censored using the final date routine data from all sources was available (31/12/15) or the date of last follow-up assessment if the last follow-up assessment occurred prior to 31/12/15. This approach ensured the routine and SANAD II datasets were matched and permitted a fair comparison and the maximal inclusion of available data.</p>		
<p><b>Total Number of Follow-Up Seizures</b></p> <p>Defined as the total number of seizures (of any type and tonic-clonic) during the time period following the date of SANAD II randomisation until the final date routine data from all sources was available (31/12/15) or the date of last SANAD II follow-up assessment if the last follow-up assessment occurred prior to 31/12/15.</p> <p>This approach ensured the routine and SANAD II datasets were matched and permitted a fair comparison and the maximal inclusion of available data.</p>	<p>The total number of seizures recorded following the date of SANAD II randomisation.</p> <p>As the dates of 'all' seizures are not recorded in SANAD II, for this assessment only the 'definite' dates of seizures recorded in the SANAD II dataset have been included.</p>	<p>The total number of seizure occurrences identified following the date of SANAD II randomisation.</p>

### *5.13.5.2 The Assessment of Antiepileptic Drugs in the Routinely Recorded Datasets*

Limited data regarding AEDs were available in the routinely recorded datasets. There were no available codes for the prescription, dose or indication in the ICD 10 or Emergency Dataset coding systems. Therefore, data retrieved from the Inpatient, Emergency, Outpatient or Critical Care datasets were not informative to the assessment of this variable. The SAIL Primary Care Dataset included all READ codes recorded for each participant for the time period 2013-present. READ codes were available for medications including AEDs identified by both generic and trade name including dosage of the tablet and are recorded in primary care electronic medical records with each prescription. However, the prescribed dose is not adequately recorded. For example, a participant may have a READ code 'Lamotrigine 25mg' but the prescribed dosage is not recorded within the READ code system and could for example include 25mg once a day, 25mg twice a day, 50mg in the morning and 25mg at night.

#### *The Identification of the Date of AED First Prescription in the Routinely Recorded Datasets*

As a result of the limitations in the routinely recorded data, the 'date of AED first prescription' was assessed, including the initial prescription of randomised AED and subsequent add-on or alternative monotherapy AED prescriptions. The date of AED first prescription is defined as the earliest date evidence of AED prescription is identified. The assessment has been limited to the 23 participants where data were available in the SAIL Primary Care Dataset. The dates of AED first prescription identified from routinely recorded datasets were compared to the data regarding AEDs in SANAD II.

Identification of the date of AED first prescription in routinely recorded datasets involved the review of specified READ codes:

- **Primary Care Dataset:**

- SAIL Primary Care Dataset**

- Date of first READ code representing prescription of an AED

Recorded READ codes were reviewed and both generic and trade AED names, detailed in *Table 5.23* were screened.

An acceptable clinical limit of agreement has been specified at three months (90 days). This time period allows for the variability in prescribing practice. For example, a possible scenario may include a patient receiving a prescription during the SANAD II baseline assessment and a two month medication supply from the hospital pharmacy. A prescription may not be evident until the third month in the Primary Care Dataset when the patient requires a re-supply of the AED and a repeat prescription is issued.

### *Compliance*

Additionally, data regarding compliance can be inferred from the Primary Care Dataset based on regularity of repeat prescription (1, 2 or 3 monthly). Such data is not recorded in the SANAD II dataset and an interpretation of the compliance derived from the Primary Care Dataset is presented.

*Table 5.23: Antiepileptic Drugs, Generic and Trade Names*

<b>Antiepileptic Drugs: Trade Name</b>	<b>Antiepileptic Drugs: Generic Name</b>
Epilim, Epilim Chrono	Sodium Valproate
Emeside/Zarontin	Ethosuximide
Keppra	Levetiracetam
Lamictal	Lamotrigine
Tegretol, Tegretol Retard	Carbamazepine
Epanutin	Phenytoin
Frisium	Clobazam
Fycompa	Perampanel
Luminal	Phenobarbital
Lyrica	Pregabalin
Neurontin	Gabapentin
Rivotril	Clonazepam
Topamax	Topiramate
Trileptal	Oxcarbazepine
Vimpat	Lacosamide
Zebinix	Eslicarbazepine
Zonegran	Zonisamide
Diacomit	Stiripentol
Diamox	Acetazolamide
Gabitril	Tiagabine
Inovelon	Rufinamide
Mysoline	Primidone
Nootropil	Piracetam
Sabril	Vigabatrin
Trobalt	Retigabine
Briviact	Brivaracetam

### 5.13.5.3 *The Assessment of Adverse Events in the Routinely Recorded Datasets*

Potentially 'any' clinical symptom or diagnosis may represent an adverse event. As such the ICD 10, emergency and READ coding systems may all contain relevant diagnostic information and data from the Inpatient, Outpatient, Emergency and Primary Care Datasets may include data informative to this assessment.

#### *The Identification of Adverse Events in the Routinely Recorded Datasets*

The occurrence of adverse events, date and clinical symptoms or diagnoses, were assessed. In the routinely recorded datasets, adverse events were identified with the occurrence of a diagnostic code consistent with 'adverse event' or the occurrence of a diagnostic code clinically consistent with an adverse event recorded in SANAD II and occurring within 90 days of the SANAD II recorded adverse event. Diagnostic codes indicating specifically 'adverse events' are detailed in *Table 5.24*. Diagnostic codes clinically consistent with adverse events recorded in SANAD II were dependent on the specific adverse event. A clinical interpretation of the diagnostic codes and symptoms recorded as adverse events in SANAD II was taken to identify consistent adverse event occurrences in the routinely recorded datasets. Adverse events frequently reported or serious in severity for the AEDs included in SANAD II are summarised in *Table 5.25* and the record of such events were sought in the routinely recorded datasets, in addition. Furthermore, healthcare attendances with inadequate diagnostic information, but within 90 days of the date of adverse event recorded in SANAD II were noted. Finally, diagnostic data occurring at any time that may provide an alternative explanation for the recorded adverse event were also extracted.

The acceptable clinical limit of agreement has been specified at 90 days, informed by clinical discussion to account for variability in healthcare attendance, recall of onset dates and duration of adverse events. For example, a possible scenario may include a patient experiencing a chronic mild symptom and reporting it to their General Practitioner who advises symptomatic relief, continuation of treatment and discussion with the Neurologist during the planned SANAD II follow-up assessment two months later where the symptom may be recorded as an adverse event in the SANAD II dataset. Adverse events were identified in routinely recorded datasets using an algorithmic approach detailed in *Figure 5.5* and compared to data collected using standard methods in SANAD II.

The identification of adverse events in routinely recorded datasets involved the review of diagnostic codes:

- **Inpatient Datasets:**  
**HES Admitted Patient Care Dataset and SAIL Patient Episode Database for Wales**  
Attendance with an ICD 10 code defining 'adverse event' or attendance with an ICD 10 code clinically consistent with an adverse event recorded in SANAD II and occurring within 90 days of the SANAD II recorded adverse event
- **Outpatient Datasets:**  
**HES Outpatient and SAIL Outpatient Datasets**  
Attendance with an ICD 10 code defining 'adverse event' or attendance with an ICD 10 code clinically consistent with an adverse event recorded in SANAD II and occurring within 90 days of the SANAD II recorded adverse event
- **Emergency Datasets:**  
**HES Accident & Emergency and SAIL Emergency Department Datasets**  
Attendance with an emergency code defining 'adverse event' or attendance with an emergency code clinically consistent with an adverse event recorded in SANAD II and occurring within 90 days of the SANAD II recorded adverse event
- **Primary Care Dataset:**  
**SAIL Primary Care Dataset**  
Attendance with a READ code defining 'adverse event' or attendance with a READ code clinically consistent with an adverse event recorded in SANAD II and occurring within 90 days of the SANAD II recorded adverse event

Figure 5.5: Algorithm for the Identification of Adverse Events

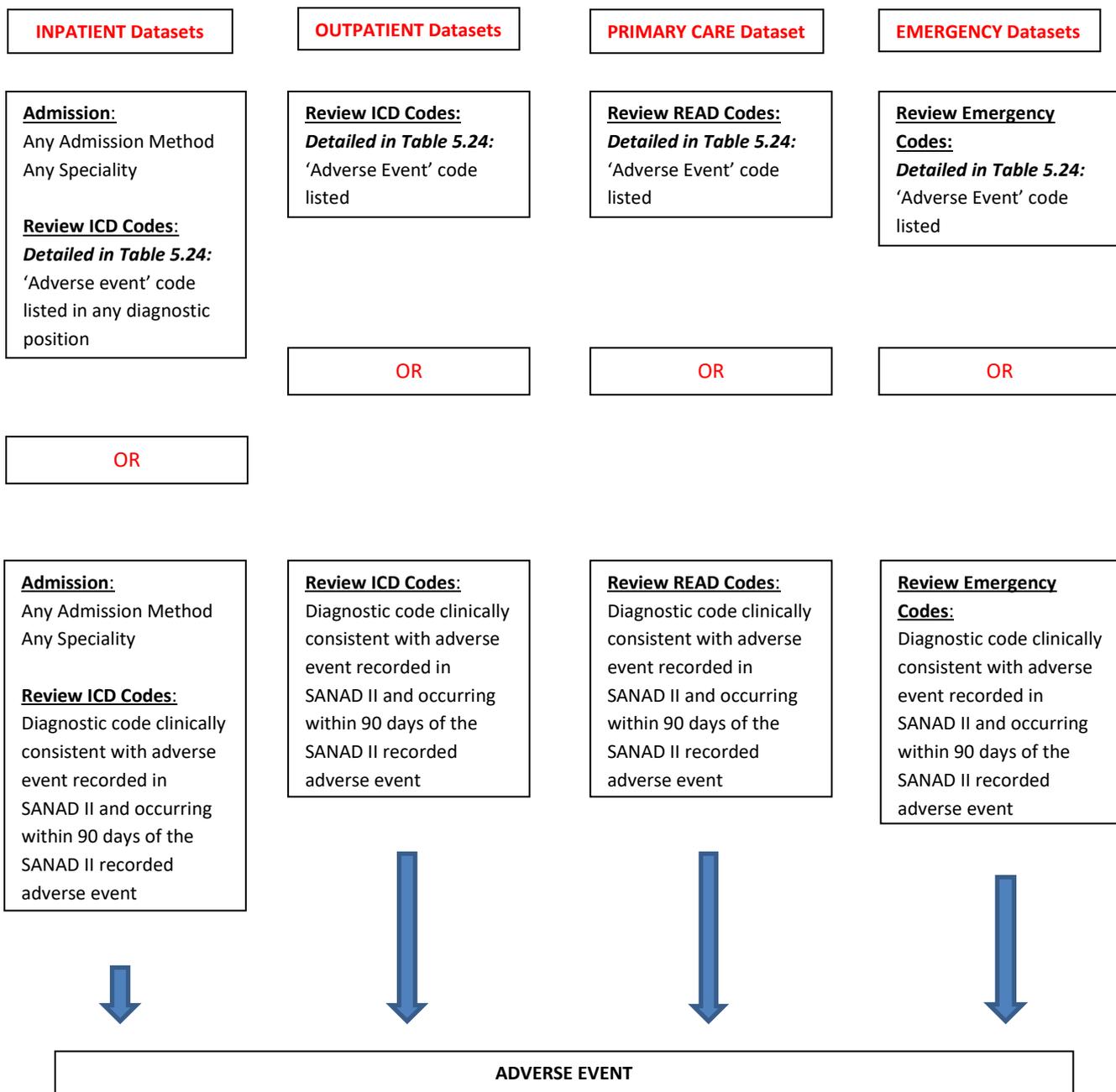


Table 5.24: Diagnostic Codes Indicating 'Adverse Events'

Code	Code Description
<b>ICD 10 Codes</b>	
T88	Other complications of surgical and medical care, not elsewhere classified
T88.6	Anaphylactic reaction due to adverse effect of correct drug or medicament properly administered
T88.7	Unspecified adverse effect of drug or medicament
T88.8	Other specified complications of surgical and medical care, not elsewhere classified
T88.9	Complication of surgical and medical care, unspecified
T42.0	Poisoning by, adverse effect of and underdosing of hydantoin derivatives
T42.1	Poisoning by, adverse effect of and underdosing of iminostilbenes
T42.2	Poisoning by, adverse effect of and underdosing of succinimides and oxazolidinediones
T42.3	Poisoning by, adverse effect of and underdosing of barbiturates
T42.4	Poisoning by, adverse effect of and underdosing of benzodiazepines
T42.5	Poisoning by, adverse effect of and underdosing of mixed antiepileptics
T42.6	Poisoning by, adverse effect of and underdosing of other antiepileptic and sedative-hypnotic drugs
T42.7	Poisoning by, adverse effect of and underdosing of unspecified antiepileptic and sedative-hypnotic drugs
<b>Emergency Code</b>	
32	Allergy (anaphylaxis)
<b>READ Codes</b>	
SN5	Adverse effects NEC
SN5z	Adverse effects NOS
TJ632	Adverse reaction to carbamazepine
TJ63	Adverse reaction to other anticonvulsant
TJ61	Adverse reaction to hydantoin derivative
TJHyz	Adverse reaction to other drug or medicine NOS
TJHz.	Adverse reaction to drug or medicinal substance NOS
TJ6	Adverse reaction to anticonvulsants and anti-parkinsonism drugs
TJ6z	Adverse reaction to anticonvulsant and antiparkinsonism drugs NOS
Xa5Jh	Lamotrigine adverse reaction
TJ632	Adverse reaction to carbamazepine
TJ633	Adverse reaction to sodium valproate
TJ610	Adverse reaction to phenytoin
TJ70	Adverse reaction to barbiturate
TJ94	Adverse reaction to benzodiazepine-based tranquilliser

Table 5.25: Frequent or Serious Adverse Events for Antiepileptic Drugs Included in SANAD II

<p><b>Lamotrigine:</b></p> <ul style="list-style-type: none"> <li>- Rash</li> <li>- Stevens Johnson Syndrome / Toxic Epidermal Necrolysis</li> <li>- Dizziness / Vertigo</li> <li>- Fatigue / Drowsiness</li> <li>- Nausea / GI Disturbance</li> <li>- Headache</li> <li>- Tremor</li> <li>- Incoordination / Ataxia</li> <li>- Deranged Liver Function Tests / Enzymes</li> <li>- Behavioural Disturbance</li> <li>- Confusion / Poor Memory</li> <li>- Anaphylaxis</li> </ul>
<p><b>Levetiracetam:</b></p> <ul style="list-style-type: none"> <li>- Fatigue / Drowsiness</li> <li>- Headache</li> <li>- Behavioural Disturbance</li> <li>- Suicidal Thoughts</li> <li>- Tremor</li> <li>- Incoordination / Ataxia</li> <li>- Confusion / Poor Memory</li> <li>- Nausea / GI Disturbance</li> <li>- Rash</li> <li>- Weight Gain</li> <li>- Dizziness / Vertigo</li> <li>- Anaphylaxis</li> </ul>
<p><b>Zonisamide:</b></p> <ul style="list-style-type: none"> <li>- Dizziness / Vertigo</li> <li>- Fatigue / Drowsiness</li> <li>- Headache</li> <li>- Rash</li> <li>- Insomnia</li> <li>- Nausea / GI Disturbance</li> <li>- Anaphylaxis</li> </ul>
<p><b>Sodium Valproate:</b></p> <ul style="list-style-type: none"> <li>- Nausea / GI Disturbance</li> <li>- Fatigue / Drowsiness</li> <li>- Weight Gain</li> <li>- Behavioural Disturbance / Aggression</li> <li>- Confusion / Poor Memory</li> <li>- Incoordination / Ataxia</li> <li>- Tremor</li> <li>- Hyperammonaemia</li> <li>- Thrombocytopenia</li> <li>- Anaphylaxis</li> </ul>

#### 5.13.5.4 *The Assessment of Healthcare Resource Use in the Routinely Recorded Datasets*

Data regarding healthcare resource use for both 'planned' and 'unplanned' healthcare attendances were compared between the SANAD II and routinely recorded datasets.

##### *The Identification of Planned Healthcare Attendances in the Routinely Recorded Datasets*

The SANAD II baseline and follow-up assessments were representative of planned healthcare attendances. In the routinely recorded datasets, SANAD II assessments were identified as a date of relevant healthcare attendance within one month of the SANAD II assessment date. Where multiple relevant attendances occur within one month, the date in closest proximity to the SANAD II assessment date was recorded. In the first instance the outpatient datasets were reviewed. Subsequently, for SANAD II assessments not identified in the outpatient datasets, the emergency and inpatient datasets were reviewed. As the SANAD II assessment would occur opportunistically and not depend on clinical reason for attendance or admission speciality, such attendances were not restricted by diagnostic code or admission speciality.

*A Priori*, an acceptable clinical limit of agreement has been specified at one month (30 days). This permits time for the completion of the SANAD II documentation that may not always be completed at the time of the trial assessment for logistical reasons, for example participant assessment during a busy outpatient clinic.

Planned healthcare attendances were identified in the routinely recorded datasets as follows:

- **Outpatient Datasets:**

- HES Outpatient and SAIL Outpatient Datasets**

- Attendance with listed medical speciality within 30 days of SANAD II assessment date:

- 301, General Medicine, 400, Neurology, 401, Clinical Neuro-Physiology, 420, Paediatrics, 421, Paediatric Neurology*

**For SANAD II Attendances Not Identified:**

- **Inpatient Datasets:**

**HES Admitted Patient Care Dataset and SAIL Patient Episode Database for Wales**

- Attendance within 30 days of SANAD II assessment date

- **Emergency Datasets:**

**HES Accident & Emergency and SAIL Emergency Department Datasets**

- Attendance within 30 days of SANAD II assessment date

The availability of data in the outpatient datasets for some participants extended into 2016 and dates of outpatient attendances identified in 2016 were included. However, where dates of SANAD II assessment in 2016 are identified in the SANAD II dataset, but not the outpatient dataset and there are no other attendances in 2016 in the outpatient dataset, such dates were excluded from the analysis as it was likely that relevant routinely recorded data were not available for the time period.

*The Identification of Unplanned Healthcare Attendances in the Routinely Recorded Datasets*

During SANAD II participants are requested to complete self-report questionnaires at specified intervals during follow-up, indicating if they have had any healthcare attendances 'in the last three months'. Episodes of healthcare attendances reported in SANAD II for the specified three month time periods were sought in the routinely recorded datasets. The date and clinical reason for attendance were extracted.

The exact dates of healthcare attendances are not recorded in the SANAD II dataset and therefore a comparison of the dates of attendance would not be valid. Therefore, the total numbers of healthcare attendances reported in SANAD II and extracted from routine datasets for the equivalent time periods were compared.

Unplanned healthcare attendances were identified in the routinely recorded datasets as follows:

**Where Emergency Department Attendances were Reported in SANAD II:**

- **Emergency Datasets:**

**HES Accident & Emergency and SAIL Emergency Department Datasets**

- Total attendances and diagnoses during the 90 days prior to the date of completion of the SANAD II self-report questionnaire

**Where Inpatient Admissions were Reported in SANAD II:**

- **Inpatient Datasets:**

**HES Admitted Patient Care Dataset and SAIL Patient Episode Database for Wales**

- Total attendances and diagnoses during the 90 days prior to the date of completion of the SANAD II self-report questionnaire

The availability of data in selected routine datasets for some participants extended into 2016 and dates of healthcare attendances identified in 2016 were included. However, where dates of SANAD II reported attendances in 2016 are not identified in the routine datasets and there were no other attendances in 2016, such attendances were excluded as it was likely that relevant routinely recorded data was not available for the equivalent time period.

In addition to emergency department attendances and inpatient admissions, participants in SANAD II are requested to provide details regarding primary care attendances. However, comparable data regarding primary care attendances could not be accurately or reliably identified from the Primary Care Dataset. The Primary Care Dataset includes READ codes and date of entry, but without context. Healthcare events including the receipt of clinical correspondence, investigation results, medication prescriptions and clinical prompts may result in READ code entry, in addition to patient attendance. It is therefore not possible to accurately and reliably identify episodes of patient attendance.

## **5.14 Conclusions**

In this chapter the methods for the assessment of quality and agreement between routinely recorded data and data collected using standard prospective methods have been presented. In the following Chapter Six, the assessment of seizure occurrence, diagnosis and classification of epilepsy and seizures in routinely recorded datasets will be examined, relevant to the identification and recruitment of individuals eligible for inclusion in SANAD II. In Chapter Seven, variables and outcome measures relevant to the follow- up of participants in SANAD II will be examined. Finally, in Chapter Eight, the feasibility and efficiency of accessing and using routinely recorded data for participants in SANAD II is discussed and recommendations for improvement proposed.

## **Chapter Six**

### **Results: The Assessment of Seizures and Data Variables Relevant to the Identification of Eligible Individuals and Recruitment into SANAD II**

#### **6.1 Introduction**

In this chapter, the assessment of seizure occurrence, diagnosis and classification of epilepsy and seizures in routinely recorded datasets is examined, relevant to the identification and recruitment of individuals eligible for SANAD II. Using the algorithmic approaches developed and presented in Chapter Five, the codes recorded for participants in the datasets together with an assessment of the quality of the available data is presented. Subsequently, an assessment of the agreement between routinely recorded data and data collected using standard prospective methods for variables relevant to the identification and recruitment of individuals was completed. Relevant variables included date of first seizure, date of diagnosis and classification of epilepsy and seizures. Finally, the record of clinical investigations was examined, including an assessment of the quality of the available data and agreement between datasets for the status (performed / not performed) of MRI Brain, CT Brain and EEG.

## **6.2 The Identification of Seizure Occurrence in the Routinely Recorded Datasets**

The identification of seizure occurrence is essential for the diagnosis of epilepsy and enrolment into SANAD II and subsequently calculation of the trial outcomes.

Records of seizure occurrence were identified in the HES APC and A&E and the SAIL PEDW, EDDS and GP Datasets. Both 'seizure' codes and 'epilepsy' codes were recorded in the routine datasets and included as representing seizure occurrence in accordance with the developed algorithm. The codes present in the routine datasets for the participants in this study indicating seizure occurrence are presented in *Table 6.1*.

There were a total 116 healthcare attendances using 137 diagnostic codes meeting the criteria for seizure occurrence identified in the routinely recorded datasets for the 98 participants included in the study. It was possible in only a minority of cases to define seizure type, using the code recorded. In all coding systems and routinely recorded datasets, the most commonly recorded codes were non-specific 'seizure' and 'epilepsy' codes. The most common ICD 10 code was 'Unspecified Convulsions (R568)', emergency code 'CNS Conditions – Epilepsy (HES 41, SAIL 17A)' and READ code 'Convulsion NOS (R003z)' and 'Had a Fit (IB63)'. The frequencies with which each code was recorded are presented in *Table 6.1*.

Table 6.1: The Occurrence of Seizures in Routinely Recorded Datasets

Code	Description	Total Records
<b>ICD Codes</b>		
G40	EPILEPSY AND RECURRENT SEIZURES*	1
G401	LOCAL-RELATED (PART) SYMPTOM EPILEPSY/EPILEPTIC SYND WITH SEIZURE	1
G402	LOCAL-RELATED (PART) SYMPTOM EPILEPSY/ EPILEP SYND	1
G403	GENERALIZED IDIOPATHIC EPILEPSY AND EPILEPTIC SYNDROMES*	5
G409	EPILEPSY, UNSPECIFIED*	9
G412	COMPLEX PARTIAL STATUS EPILEPTICUS	1
R568	OTHER AND UNSPECIFIED CONVULSIONS*	27
<b>Emergency Codes</b>		
HES: 241	CNS Conditions – Epilepsy*	25
HES: 24	CNS Disorder*	25
SAIL: 17A	Seizure / Convulsion*	8
SAIL: 17Z	CNS Condition – Unspecified*	5
<b>Primary Care READ Codes</b>		
2828	Absence Seizure	2
F251600	Grand mal seizure*	2
1B63.00	Had a fit*	7
R003400	[D]Nocturnal seizure*	1
R003z00	[D]Convulsion NOS*	7
282..00	O/E - fit/convulsion*	1
F25z.11	Fit (in known epileptic) NOS*	1
F254500	Complex partial epileptic seizure	1
R003200	[D]Fit*	1
F251200	Epileptic seizures – clonic*	1
F25..00	Epilepsy*	1
F254000	Temporal lobe epilepsy	2
F255000	Jacksonian, focal or motor epilepsy	1
Fyu5100	[X]Other epilepsy*	1

**Legend:** \*: Codes indicating tonic-clonic seizures

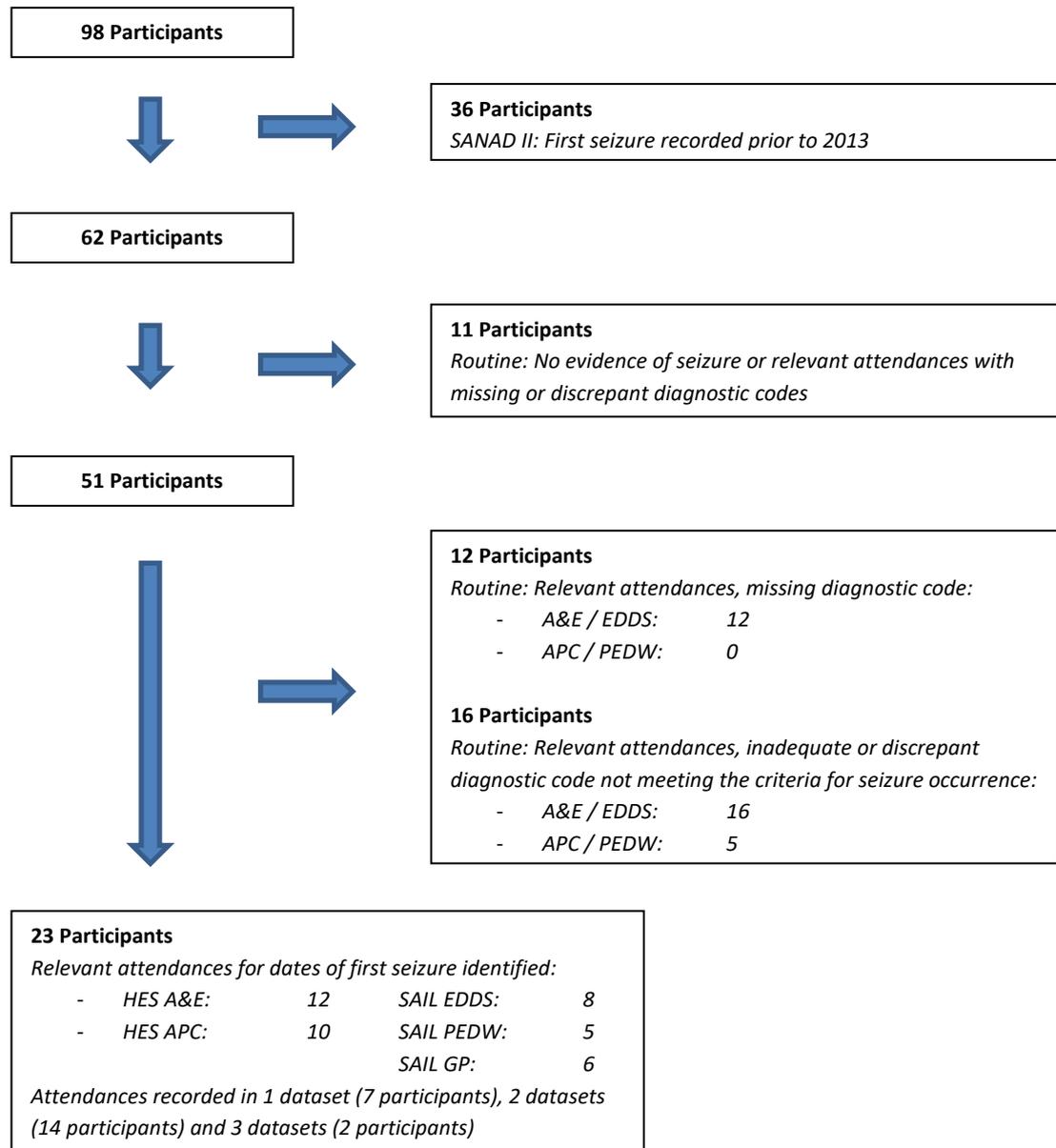
### **6.3 Date of First Seizure**

Thirty six participants had a 'first seizure' recorded in the SANAD II dataset occurring prior to 2013. This group of participants were excluded from this assessment due to the lack of availability of routine data coverage for the time period before 2013. The remaining sixty two participants had a first seizure occurrence recorded in SANAD II.

In the routine datasets a first seizure occurrence was identified in 23 of 62 participants. The identification of relevant participants is presented in *Figure 6.1*. The most common codes were the ICD code 'Unspecified Convulsions (R568)' occurring in 13 participants, and the emergency code 'CNS Conditions, Epilepsy' occurring in 10 participants. All seizure occurrences identified in datasets using the ICD 10 coding system were classified as 'definite' using the developed algorithm.

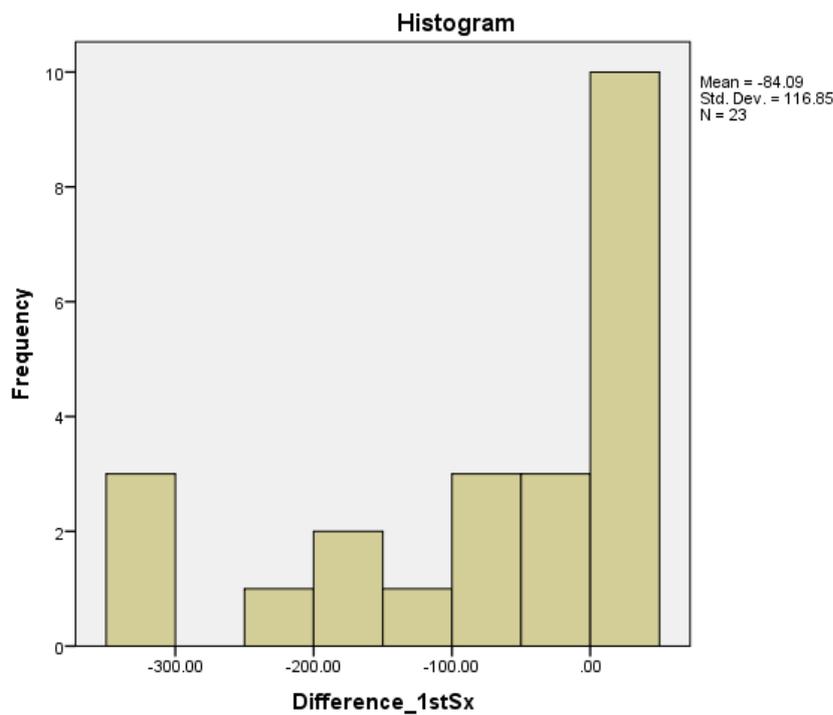
Sixteen participants had a relevant attendance within 48 hours of a definite seizure recorded in SANAD II but with inadequate or discrepant diagnostic codes not meeting the criteria for seizure occurrence. Codes included 'CNS, Non-Epilepsy' in the emergency datasets and 'Disorientation' and 'Confusion' in the inpatient datasets. Of the 16 participants, six had relevant attendances inadequately coded 'CNS Disorder' and 'CNS Condition – Unspecified' in the emergency datasets. Following clinical discussion, it was deemed not appropriate to include such non-specific codes to indicate seizure occurrence prior to diagnosis of epilepsy.

Figure 6.1: The Identification of the Date of First Seizure in Routine Datasets



The difference in days between the dates of first seizure from SANAD II subtracted from the dates from routine datasets was calculated and the assumption of normal distribution around the mean assessed. The distribution displays a negative skew on inspection, detailed in *Figure 6.2*.

*Figure 6.2: The Difference in Days Between the Date of First Seizure*

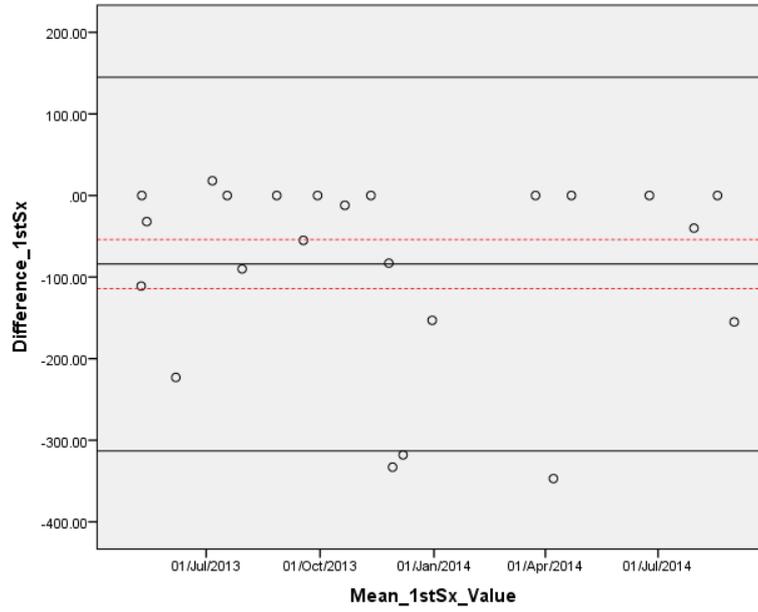


A Wilcoxon Signed Rank Test has been performed in SPSS. The significant result ( $P=0.002$ ) indicates that the mean dates of first seizure calculated from SANAD II and routine datasets are significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 6.2* and *Figure 6.3*.

The Bland Altman Plot demonstrates that agreement is poor. The 95% confidence limits of agreement between the dates of first seizure are 145 and -313 days, clearly in excess of the specified 30 day clinically acceptable limit indicated by the red dashed lines. The mean of the difference between the dates is -84, indicating that the date of first seizure is identified in the SANAD II dataset a mean of 84 days earlier.

Figure 6.3: Date of First Seizure: Bland Altman Plot

Mean	-84.09
Upper 95% Confidence Limit of Agreement	144.94
Lower 95% Confidence Limit of Agreement	-313.12



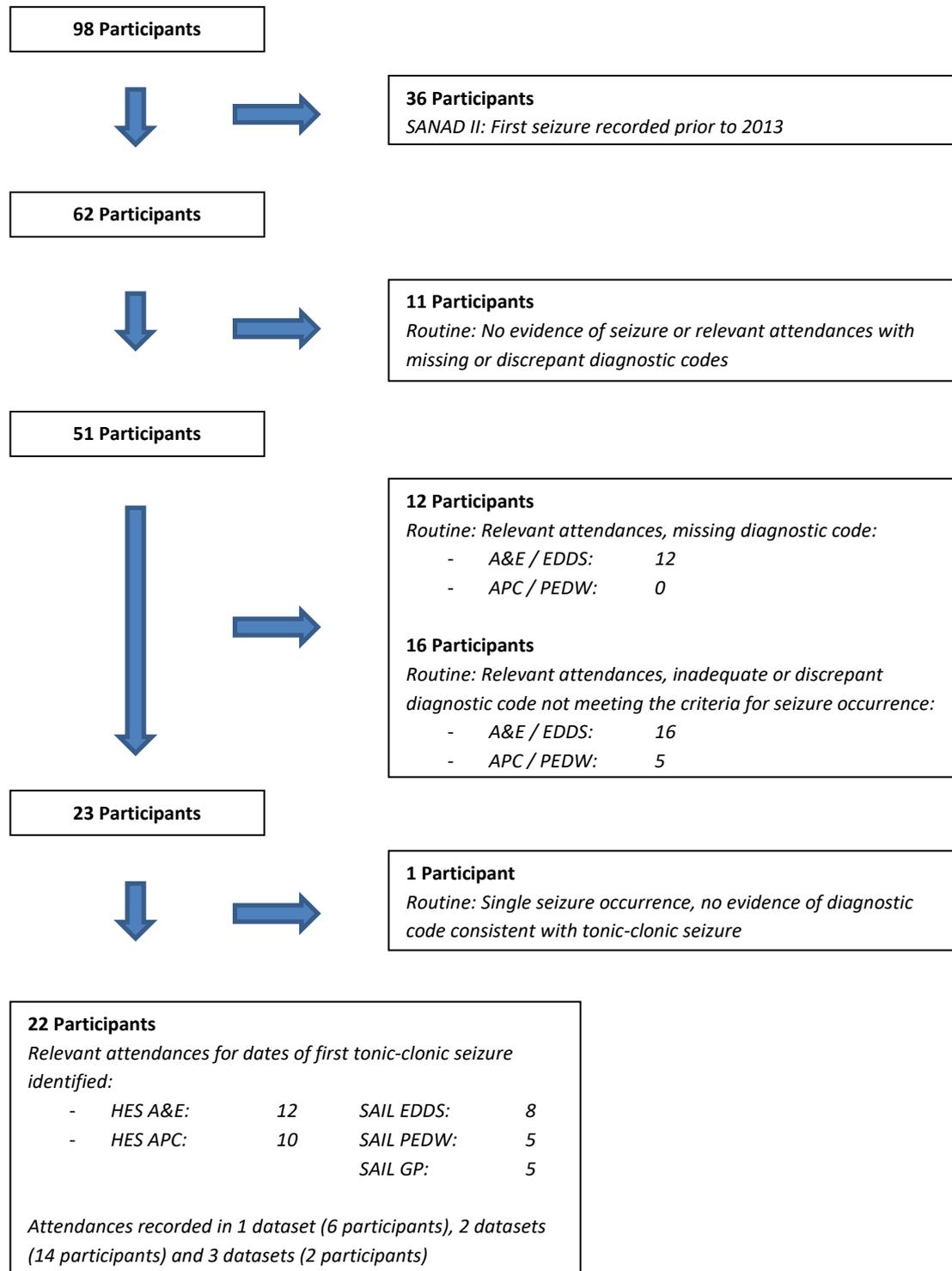
#### **6.4 Date of First Tonic-Clonic Seizure**

Thirty six participants had a 'first seizure' recorded in the SANAD II dataset occurring prior to 2013 and were excluded. Of the remaining 62 participants, 43 had a first tonic-clonic seizure occurrence recorded in SANAD II.

In the routine datasets a first tonic-clonic seizure occurrence was identified in 22 participants. The identification of relevant participants is presented in *Figure 6.4*. The most common codes were the ICD code 'Unspecified Convulsions (R568)' occurring in 13 participants, and the emergency code 'CNS Conditions, Epilepsy' occurring in 10 participants. In eight participants without a first tonic-clonic seizure in the SANAD II dataset, a first tonic-clonic seizure occurrence was identified in routine datasets.

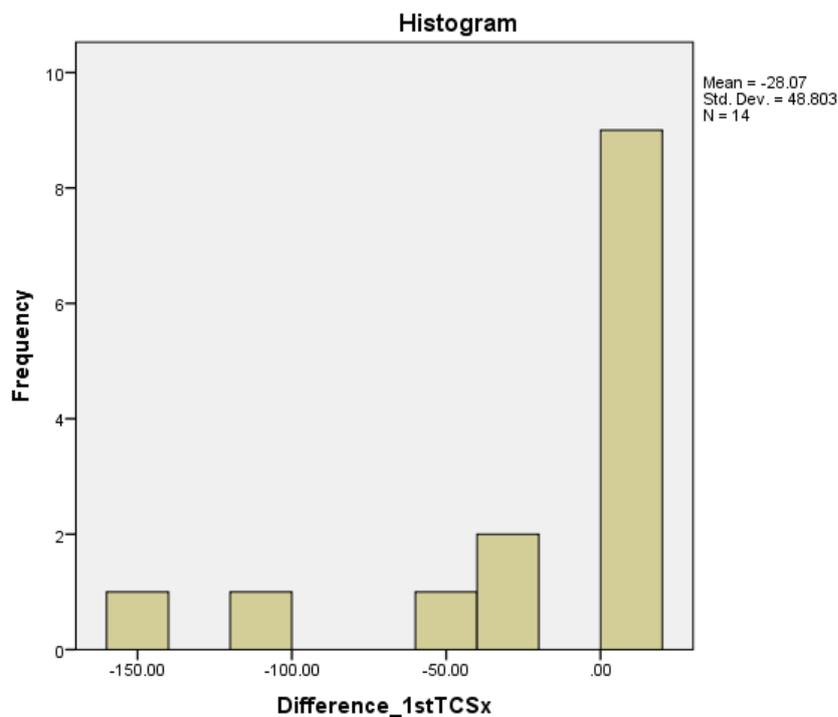
Sixteen participants had a relevant attendance within 48 hours of a definite seizure recorded in SANAD II but with inadequate or discrepant diagnostic codes not meeting the criteria for seizure occurrence. Codes included 'CNS, Non-Epilepsy' in the emergency datasets and 'Disorientation' and 'Confusion' in the inpatient datasets. Of the 16 participants, six had relevant attendances inadequately coded 'CNS Disorder' and 'CNS Condition – Unspecified' in the emergency datasets. Again, it was deemed not appropriate to include such non-specific codes to indicate seizure occurrence prior to the diagnosis of epilepsy.

Figure 6.4: The Identification of the Date of First Tonic-Clonic Seizure in Routine Datasets



The difference in days between the dates of first tonic-clonic seizure from SANAD II subtracted from the dates from routine datasets was calculated and the assumption of normal distribution around the mean assessed. The distribution displays a negative skew on inspection, detailed in *Figure 6.5*.

*Figure 6.5: The Difference in Days Between the Date of First Tonic-Clonic Seizure*

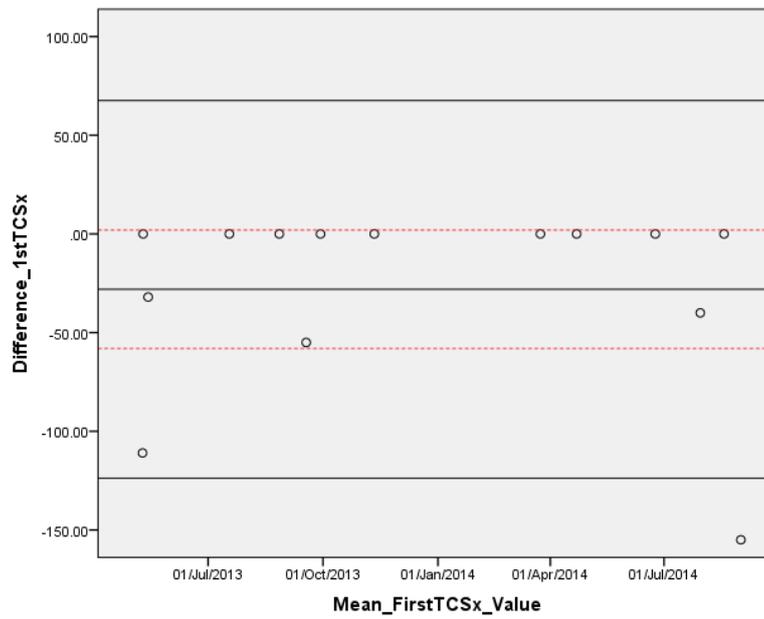


A Wilcoxon Signed Rank Test has been performed in SPSS. The significant result ( $P=0.043$ ) indicates that the mean dates of first seizure calculated from SANAD II and routine datasets are significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 6.2* and *Figure 6.5*.

The Bland Altman Plot demonstrates that agreement is poor. The 95% confidence limits of agreement between the dates of first seizure are 68 and -124 days, clearly in excess of the specified 30 day clinically acceptable limit indicated by the red dashed lines although improved compared to the agreement for first seizures of any type. The mean of the difference between the dates is -28, indicating that the date of first seizure is identified in the SANAD II dataset a mean of 28 days earlier.

Figure 6.6: Date of First Tonic-Clonic Seizure: Bland Altman Plot

Mean	-28.07
Upper 95% Confidence Limit of Agreement	67.58
Lower 95% Confidence Limit of Agreement	-123.72



## **6.5 Conclusions: Date of First Seizure and First Tonic-Clonic Seizure**

The quality and agreement for variables and outcome measures involving record of the occurrence of first seizures was poor. Dates of first tonic-clonic seizure were also poorly recorded, although data are more complete in the routinely recorded datasets compared to the analysis of first seizure of any type. Attendances as a result of seizure occurrence are either missing or do not include diagnostic codes consistent with seizure.

For the limited number of participants where first seizures were identified, the agreement for the date of occurrence compared to the date collected using standard prospective methods in SANAD II was poor, with first seizure occurrences identified earlier in the SANAD II dataset. The delay in identification of first seizure occurrence has implications for the utility of routinely recorded data. In epilepsy research, routinely recorded data may be limited for the identification of eligible individuals for recruitment into prospective trials. In SANAD II for example, individuals must not have commenced treatment with antiepileptic drugs; this is less likely if there is a delay in identification. In clinical practice, the missing routinely recorded data is perhaps of greater importance, with impacts on the incidence and prevalence rates if data are used for disease monitoring purposes.

Explanations for these findings may include inaccurate recording of diagnostic codes in routinely recorded datasets or inaccurate initial clinical diagnosis of seizure. Furthermore, participants may not have sought medical attention and therefore would not have attendances recorded in routine datasets, although following a first occurrence of seizure and particularly tonic-clonic seizure, it would be unlikely for an individual not to seek medical attention.

Table 6.2: The First Seizure and First Tonic-Clonic Seizure: Descriptive Statistics and Agreement

		<b>First Seizure (All Types)</b>	<b>First Tonic-Clonic Seizure</b>
<b>RCT Data Patients</b>	Patients with Seizures	62 (100%)	43 (69.4%)
	Total Eligible Patients	62	62
<b>Routine Data Patients</b>	Patients with Seizures	23 (37.1%)	22 (35.5%)
	Total Eligible Patients	62	62
Total Paired Patients		23 (37.1%)	14 (22.6%)
<b>Routine Datasets</b>	Total Seizures in Dataset (Total Seizures in 'Greatest Detail': Mutually Exclusive)	HES: Admitted Patient Care	10(10)
		HES: Accident and Emergency	12(2)
		HES: Outpatient	0(0)
		HES: Adult Critical Care	0(0)
		SAIL: Patient Episode Database for Wales	5(5)
		SAIL: Emergency	8(3)
		SAIL: Outpatient	0(0)
		SAIL: Primary Care	6(3)
<b>Assessment of Agreement</b>	<b>RCT Data</b>	Mean	15/10/13
		Range	15/02/13 – 18/08/14
	<b>Routine Data</b>	Mean	07/01/14
		Range	10/05/13 – 17/11/14
	Test for Significance		Wilcoxon Signed Rank
	Significance		P=0.002
		Wilcoxon Signed Rank	
		P=0.043	

## **6.6 The Identification of Diagnosis and Classification of Epilepsy and Seizures in the Routinely Recorded Datasets**

A diagnosis of epilepsy is a mandatory requirement for enrolment in SANAD II and participants are diagnosed based on the accepted definitions of epilepsy, proposed by the International League Against Epilepsy [7, 8].

In the routinely recorded datasets, a diagnosis of epilepsy was identified using both algorithmic methods; the occurrence of two episodes of seizure greater than 24 hours apart or the record of an 'epilepsy' code consistent with a diagnosis of epilepsy. Episodes of seizure occurrence were identified in the HES APC and A&E and the SAIL PEDW, EDDS and GP Datasets and have previously been presented in *Table 6.1*. For participants meeting the criteria for a diagnosis of epilepsy in the routinely recorded datasets, the classification of seizure type was subsequently derived through interpretation of the codes of greatest detail. For participants meeting the criteria for a diagnosis through record of an 'epilepsy' code, *Table 6.3* presents the codes recorded for the participants in this study.

Forty seven of the 98 participants included in this study met the criteria for diagnosis of epilepsy in the routinely recorded datasets using all available data. Thirteen participants within this group met the criteria for diagnosis through the record of two or more episodes of seizure occurrence. It was possible in only a minority of cases to adequately define classification of seizure type, dependant on the diagnostic code recorded. In all coding systems and datasets, the most commonly recorded codes indicating seizure occurrence were non-specific codes with limited diagnostic and clinical value. The most common code indicating the diagnosis of epilepsy was the ICD code 'Focal Epilepsy, Simple Partial (G401)' identified in the HES OP dataset in 13 participants followed by the READ codes 'Epilepsy (F25)' and 'Epilepsy; Jacksonian, Focal, Motor (F2550)', occurring in the SAIL GP dataset in eight and four participants respectively. In the routinely recorded datasets, the codes indicating diagnosis and classification of epilepsy and seizures and the frequency each are recorded are presented in *Table 6.3*.

*Table 6.3: The Diagnosis and Classification of Epilepsy and Seizures in Routinely Recorded Datasets*

<b>Code</b>	<b>Description</b>	<b>Total Records</b>
<b>Focal Epilepsy</b>		
G401	LOCAL-RELATED (PART) SYMPTOM EPILEPSY/EPILEPTIC SYND WITH SEIZURE	13
G402	LOCAL-RELATED (PART) SYMPTOM EPILEPSY/ EPILEP SYND	1
F254000	Temporal lobe epilepsy	2
F255000	Jacksonian, focal or motor epilepsy	4
<b>Generalised Epilepsy</b>		
G403	GENERALIZED IDIOPATHIC EPILEPSY AND EPILEPTIC SYNDROMES	3
<b>Unclassified Epilepsy</b>		
G40	EPILEPSY AND RECURRENT SEIZURES	3
G409	EPILEPSY, UNSPECIFIED	2
F25..00	Epilepsy	8

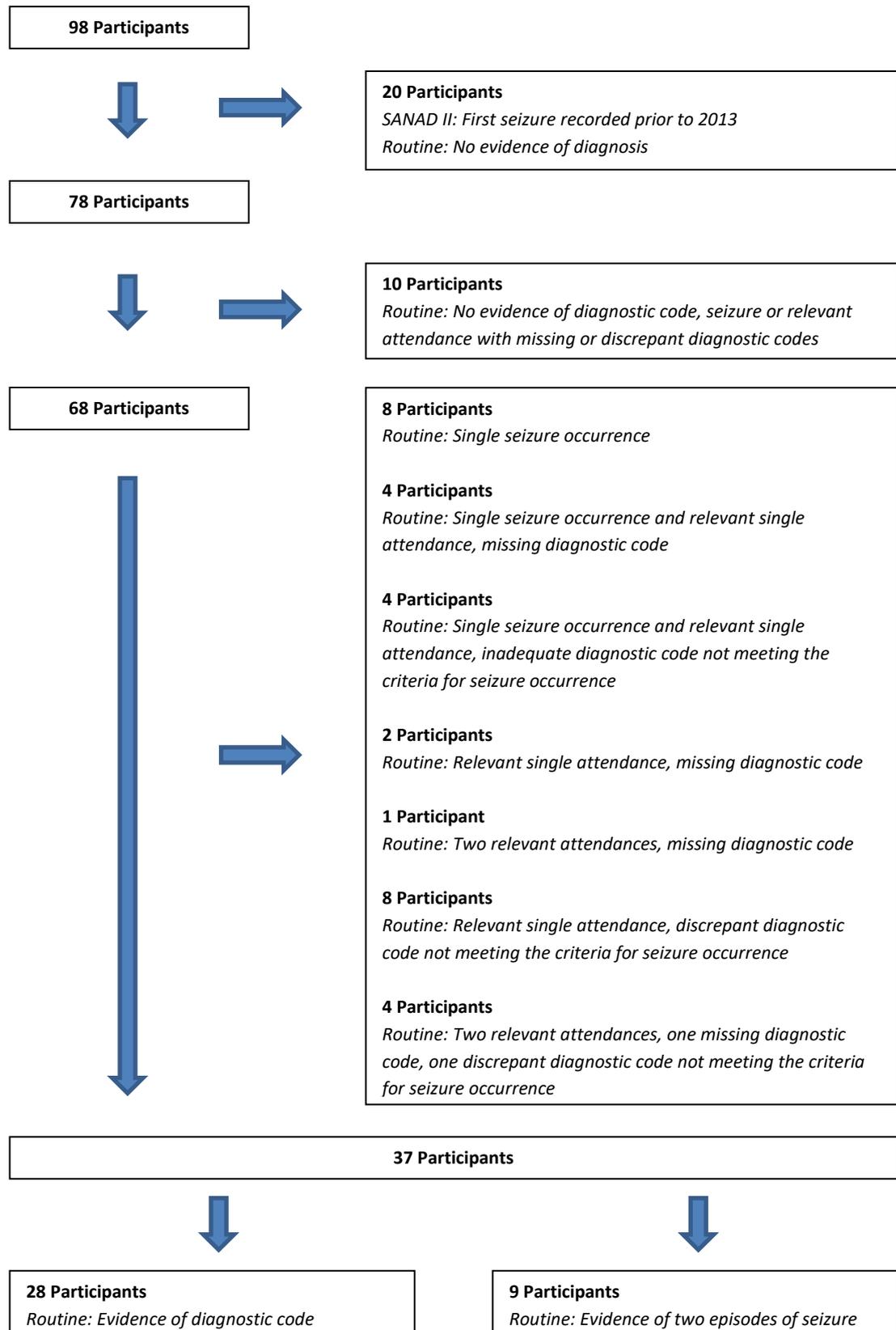
## **6.7 Diagnosis of Epilepsy (Baseline)**

Thirty six participants had a 'first seizure' recorded in the SANAD II dataset occurring prior to 2013. Within this group, in the routine datasets 20 participants did not meet the criteria for a baseline diagnosis of epilepsy and were excluded from further analysis due to the lack of availability of routine data coverage for the time period before 2013, removing the potential for two seizures to be recorded. Seventy eight participants were included in the assessment and had a diagnosis of epilepsy in the SANAD II dataset.

In the routine datasets a baseline diagnosis of epilepsy was identified in 41 participants. The identification of the date of baseline diagnosis is presented in *Figure 6.7*. Nine participants met the criteria for a baseline diagnosis through the record of two episodes of seizure occurrence. Twenty eight participants met the criteria for a baseline diagnosis through the record of a code consistent with a diagnosis of epilepsy. A diagnostic code was recorded in the primary care dataset in 13 participants, outpatient datasets in 10 participants, inpatient datasets in four participants and emergency care dataset in one participant. Notably, primary care data were available for 28 participants, in 13 of which a diagnosis was available. Outpatient data were available for all 98 participants yet a diagnostic code was recorded in only 10, all recruited from a single centre.

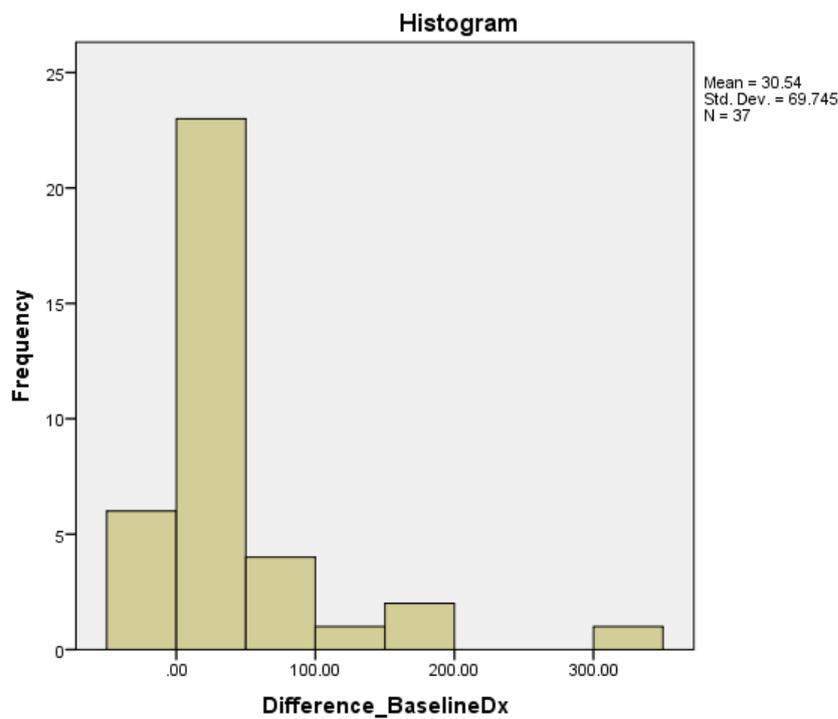
As detailed in *Figure 6.7*, 11 participants had a relevant healthcare attendance within 48 hours of a definite seizure recorded in SANAD II but with missing diagnostic information. All such attendances were recorded in the emergency datasets. Sixteen participants had a relevant attendance but with inadequate diagnostic information, not meeting the criteria for seizure occurrence. Such attendances were most commonly recorded in the emergency datasets and were coded 'CNS Disorder' and 'CNS Condition – Unspecified'. Following clinical discussion, it was deemed not appropriate to include such non-specific codes to represent seizure occurrence prior to the diagnosis of epilepsy.

Figure 6.7: The Identification of the Date of Baseline Diagnosis in Routine Datasets



The difference in days between the dates of diagnosis from SANAD II subtracted from the dates of diagnosis from routine datasets was calculated and the assumption of normal distribution around the mean assessed. The distribution displays a positive skew on inspection, detailed in *Figure 6.8*.

*Figure 6.8: The Difference in Days Between the Dates of Baseline Diagnosis of Epilepsy*



A Wilcoxon Signed Rank Test has been performed in SPSS. The significant result ( $P=0.004$ ) indicates that the mean dates of diagnosis calculated from SANAD II and routine datasets are significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 6.4* and *Figure 6.9*.

The Bland Altman Plot demonstrates that agreement is poor. The 95% confidence limits of agreement between the dates of baseline diagnosis are 167 and -106 days, clearly in excess of the specified 30 day clinically acceptable limit indicated by the red dashed lines. The mean of the difference between the dates is 31, indicating that the date of baseline diagnosis is identified in the routine datasets a mean of 31 days earlier.



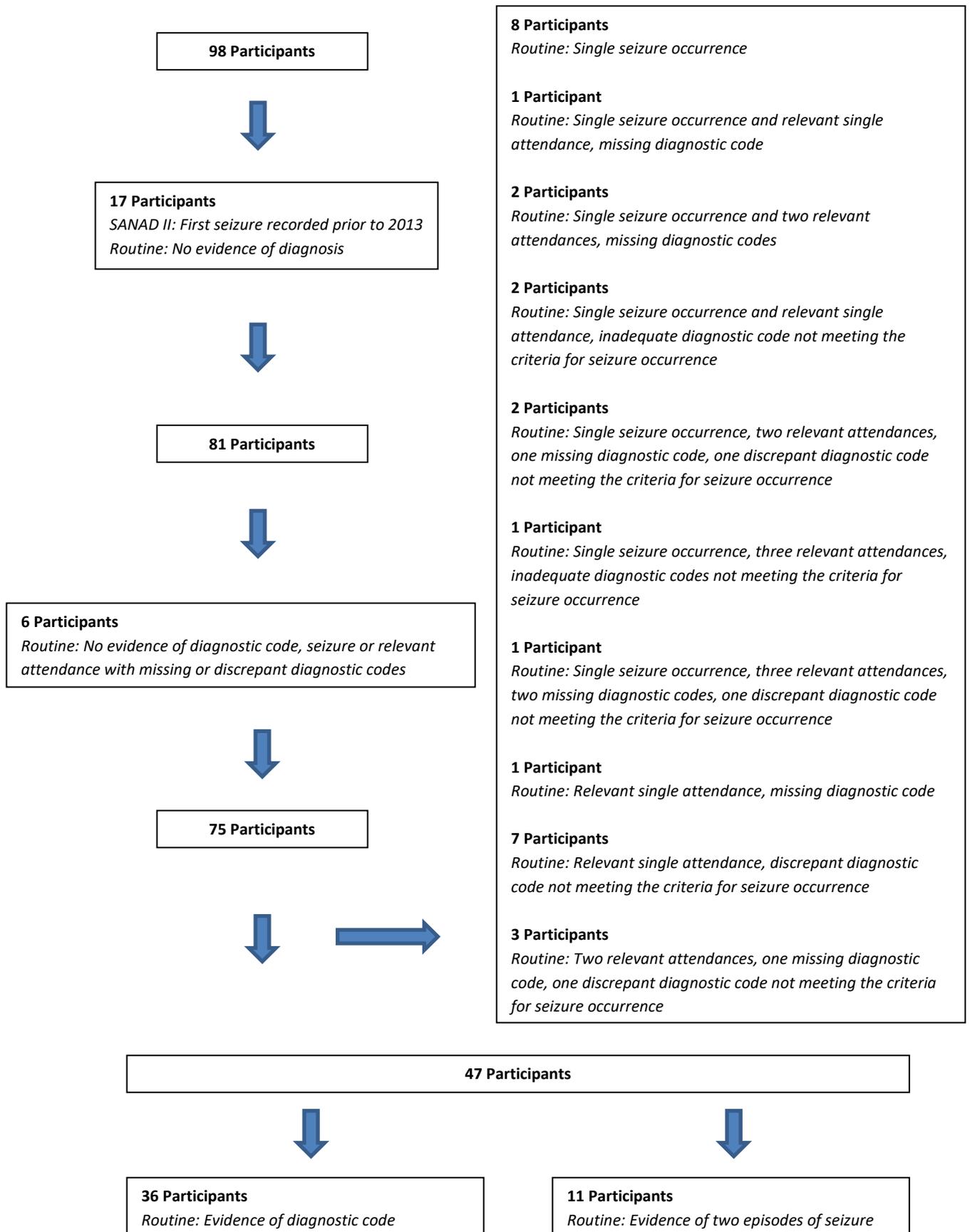
## **6.8 Diagnosis of Epilepsy (All-Time)**

Thirty six participants had a 'first seizure' recorded in the SANAD II dataset occurring prior to 2013. Within this group, in the routine datasets 17 participants did not qualify for an all-time diagnosis of epilepsy and were excluded from further analysis due to the lack of availability of routine data coverage for the time period before 2013, removing the potential for two seizures to be recorded. Eighty one participants were included in the assessment and had a diagnosis of epilepsy in the SANAD II dataset.

In the routine datasets an all-time diagnosis of epilepsy was identified in 47 participants. The identification of the date of all-time diagnosis is presented in *Figure 6.10*. Eleven participants met the criteria for a diagnosis through the record of two episodes of seizure occurrence. Thirty six participants met the criteria for a diagnosis through the record of a code consistent with a diagnosis of epilepsy. A diagnostic code was recorded in the primary care dataset in 14 participants, outpatient datasets in 15 participants, inpatient datasets in six participants and emergency care dataset in one participant. Notably, primary care data was available for 28 participants, in 14 of which a diagnosis was available. Outpatient data was available for all 98 participants yet a diagnostic code was recorded in only 15.

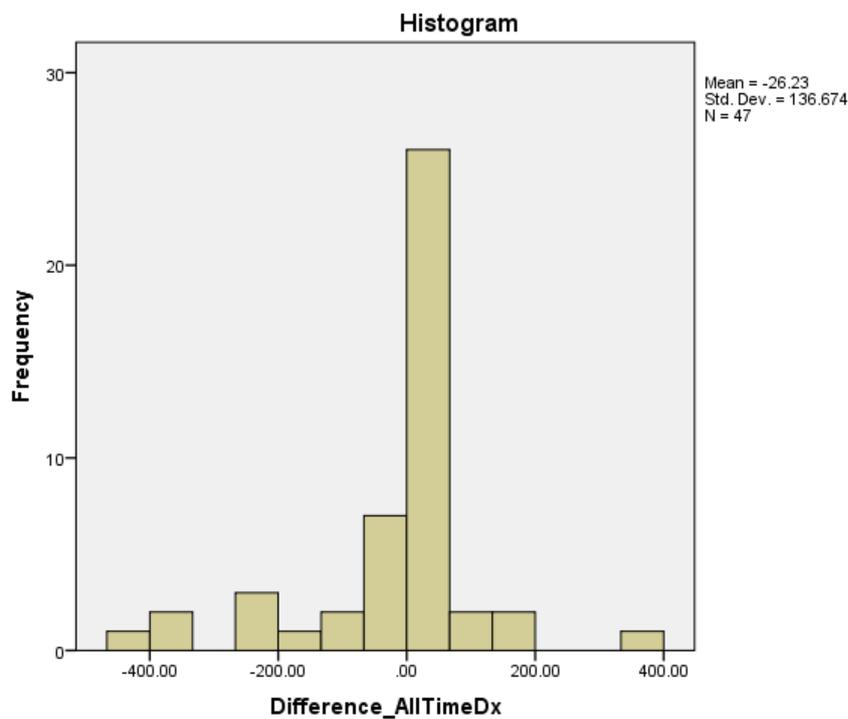
As detailed in *Figure 6.10*, 10 participants have a relevant healthcare attendance within 48 hours of a definite seizure recorded in SANAD II but with missing diagnostic information. All such attendances were recorded in the emergency datasets. Sixteen participants have a relevant attendance but with inadequate diagnostic information, not meeting the criteria for seizure occurrence. Such attendances were most commonly recorded in the emergency datasets and were coded 'CNS Disorder' and 'CNS Condition – Unspecified'. These results are very similar to those found for diagnosis of epilepsy at baseline and similarly, it was deemed not appropriate to include non-specific codes to represent seizure occurrence prior to the diagnosis of epilepsy.

Figure 6.10: The Identification of the Date of All-Time Diagnosis in Routine Datasets



The difference in days between the dates of diagnosis from SANAD II subtracted from the dates of diagnosis from routine datasets was calculated and the assumption of normal distribution around the mean assessed. The distribution is normal to inspection, detailed in *Figure 6.11*.

*Figure 6.11: The Difference in Days Between the Dates of All-Time Diagnosis of Epilepsy*



A Paired T-Test has been performed in SPSS. The non-significant result ( $P=0.195$ ) indicates that the mean dates of diagnosis calculated from SANAD II and routine datasets are not significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 6.4* and *Figure 6.12*.

Despite the lack of significant difference between the means of the all-time diagnosis calculated using the SANAD II and routine datasets, the Bland Altman Plot demonstrates that agreement is poor. The 95% confidence limits of agreement between the dates of all-time diagnosis are 242 and -294 days, clearly in excess of the specified 30 day clinically acceptable limit indicated by the red dashed lines. The mean of the difference between the dates is -26, indicating that the date of all-time diagnosis is identified in the SANAD II dataset a mean of 26 days earlier.

Figure 6.12: All-Time Diagnosis of Epilepsy: Bland Altman Plot

Mean	-26.23
Upper 95% Confidence Limit of Agreement	241.64
Lower 95% Confidence Limit of Agreement	-294.10

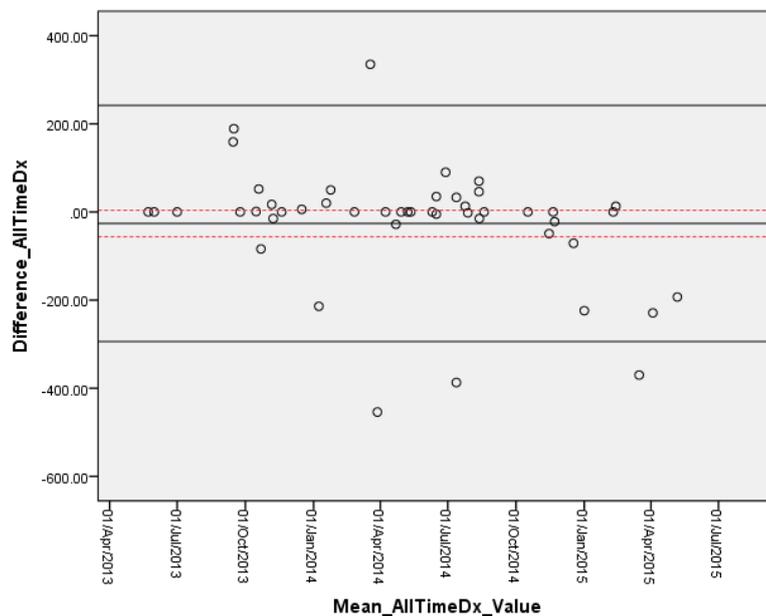


Table 6.4: The Diagnosis of Epilepsy: Descriptive Statistics and Agreement

		Baseline Diagnosis	All-Time Diagnosis
RCT Data Patients	Total Diagnoses	78 (100%)	81 (100%)
	Total Eligible Patients	78	81
Routine Data Patients	Total Diagnoses	37 (47.4%)	47 (58.0%)
	Total Eligible Patients	78	81
Total Paired Diagnoses		37 (47.4%)	47 (58.0%)
Assessment of Agreement	RCT Data	Mean	23/04/14
		Range	23/05/13 – 19/02/15
	Routine Data	Mean	24/03/14
		Range	23/05/13 – 09/02/15
		Test for Significance	Wilcoxon Signed Rank
		Significance	P=0.004
		Paired T Test	
		P=0.195	

### **6.9 Classification of Seizures (Baseline)**

Thirty seven participants met the criteria for a baseline diagnosis of epilepsy in routine datasets. Of those, in the SANAD II dataset 33 participants were classified as focal, three participants as generalised and one participant unclassified. In the routine dataset, 18 participants were unclassified, 17 participants classified as focal and two were classified as generalised. In the 17 participants classified as focal seizure in the routine dataset, all were classified as focal in the SANAD II dataset. One of the two participants classified as generalised in the routine dataset was classified as focal in the SANAD II dataset. The results have been summarised in *Table 6.5*.

In the routine dataset, eight participants of the nine meeting the criteria for a baseline diagnosis of epilepsy through the record of two episodes of seizure occurrence had classifications in disagreement to the classification recorded in the SANAD II dataset. All such participants were deemed unclassified in the routine datasets as a result of limited diagnostic data. Ten of the 28 participants meeting the criteria for a baseline diagnosis of epilepsy through the record of a diagnostic code consistent with a diagnosis of epilepsy had classifications in disagreement to the classification recorded in the SANAD II dataset. Of these 10 participants, six diagnostic codes were recorded in the primary care dataset, one in the emergency care dataset and three in the inpatient datasets. Notably, in the 10 participants with diagnostic data recorded in the outpatient dataset, all classifications were in agreement. As previously discussed, all such participants were recruited in a single centre.

Agreement has been assessed using cross tabulation and Cohens Kappa. The value of Kappa is 0.151 (P=0.018) indicating 'none to slight' agreement between baseline classification of seizures determined from SANAD II and routine datasets. Results are presented in *Table 6.7*.

Table 6.5: Baseline Classification: Cross-Tabulation

RCT Data		Routine Data			
		Focal	Generalised	Unclassified	Total
Focal		17 (46.0%)	1 (2.7%)	15 (40.6%)	33 (89.2%)
Generalised		0	1 (2.7%)	2 (5.4%)	3 (8.1%)
Unclassified		0	0	1 (2.7%)	1 (2.7%)
Total		17 (46.0%)	2 (5.4%)	18 (48.6%)	37

### **6.10 Classification of Seizures (All-Time)**

Forty seven participants met the criteria for an all-time diagnosis of epilepsy in routine datasets. Of those, in the SANAD II dataset 43 participants were classified as focal, three participants as generalised and one participant unclassified. In the routine datasets, 22 participants were unclassified, 22 participants classified as focal and three were classified as generalised. In the 22 participants classified as focal seizure in the routine dataset, all were classified as focal in the SANAD II dataset. Two of the three participants classified as generalised in the routine dataset were classified as focal in the SANAD II dataset, with both diagnostic codes recorded in the inpatient datasets. The results have been summarised in *Table 6.6*.

In the routine dataset, nine participants of the 11 meeting the criteria for an all-time diagnosis of epilepsy through the record of two episodes of seizure occurrence had classifications in disagreement to the classification recorded in the SANAD II dataset. All such participants were deemed unclassified in the routine datasets as a result of limited diagnostic data. Thirteen of the 36 participants meeting the criteria for an all-time diagnosis of epilepsy through the record of a diagnostic code consistent with a diagnosis of epilepsy had classifications in disagreement to the classification recorded in the SANAD II dataset. Of these 13 participants, eight diagnostic codes were recorded in the primary care dataset, one in the emergency care dataset and four in the inpatient datasets. Notably, in the 15 participants with diagnostic data recorded in the outpatient dataset, 13 classifications were in agreement. The two participants not in agreement both had a record of the ICD code 'Epilepsy and Recurrent Seizures (G40)' and were therefore deemed unclassified.

Agreement has been assessed using cross tabulation and Cohens Kappa. The value of Kappa is 0.123 (P=0.019) indicating 'none to slight' agreement between baseline classification of seizures determined from SANAD II and routine datasets. Results are presented in *Table 6.7*.

Table 6.6: All-Time Classification: Cross-Tabulation

RCT Data		Routine Data			
		Focal	Generalised	Unclassified	Total
	Focal	22 (46.8%)	2 (4.3%)	19 (40.4%)	43 (91.5%)
	Generalised	0	1 (2.1%)	2 (4.3%)	3 (6.4%)
	Unclassified	0	0	1 (2.1%)	1 (2.1%)
	Total	22 (46.8%)	3 (6.4%)	22 (46.8%)	47

Table 6.7: The Classification of Seizures: Agreement

		Baseline Classification	All-Time Classification
Total Paired Classifications		37	47
Assessment of Agreement	Cohens Kappa	0.151	0.123
	Significance	P=0.018	P=0.019
	Interpretation	'None to Slight'	'None to Slight'

### **6.11 Conclusions: Diagnosis and Classification of Epilepsy and Seizures**

These results indicate that the quality of routinely recorded data and agreement to data collected using standard prospective methods are poor for the variables date of diagnosis of epilepsy and classification of seizures. Diagnostic codes consistent with a 'diagnosis of epilepsy' are present in less than half of the participants. Furthermore, the poor record of seizure occurrence results in even fewer participants having evidence of two episodes of seizure occurrence required for diagnosis. Where diagnosis was possible, agreement for the date of diagnosis was poor with participants diagnosed at an earlier interval in the routine datasets at baseline, likely explained by the coding of seizure occurrence with a diagnostic yet non-specific 'epilepsy' code, either as a result of the lack of more specific codes, for example in the Emergency Datasets, or a result of inaccurate recording of codes in routinely recorded datasets or inaccurate initial clinical diagnosis by non-specialist physicians. In those participants meeting the criteria for diagnosis, there was poor agreement for the classification of seizures. This is explained by the majority of participants being deemed 'unclassified' as a result of inadequate clinical detail in the routine data. This has implications for the utility of routinely recorded data to identify individuals diagnosed with epilepsy to assist with trial recruitment in epilepsy research or identify the incidence and prevalence rates if data are used for disease monitoring purposes.

The Outpatient Dataset had record of diagnostic 'epilepsy' codes but only in a small minority of participants. In one recruitment centre, all participants had a diagnostic code recorded in the Outpatient Dataset following Neurological review. This is a result of a local electronic proforma completed by the clinician following the clinic attendance. In the absence of the ability to incorporate antiepileptic drug prescribing into the algorithm to identify diagnosis of epilepsy, Neurological outpatient review followed by record of a diagnostic 'epilepsy' code selected by the clinician perhaps represents the most accurate method to identify the diagnosis of epilepsy in the included routine datasets. Furthermore, for the small number of participants with diagnosis identified in the Outpatient Dataset in this study, the diagnostic codes recorded permitted more detailed classification of seizures.

## **6.12 The Identification of Clinical Investigations in the Routinely Recorded Datasets**

The clinical investigations Magnetic Resonance Imaging (MRI) Brain, Computed Tomography (CT) Brain and Electroencephalography (EEG) are recorded at recruitment into SANAD II.

Clinical investigations were identified in the HES A&E and the SAIL EDDS and GP datasets. There were 27 records of CT, nine records of MRI and eight records of EEG. The codes present in the routine datasets for the participants in this study indicating MRI Brain, CT Brain and EEG are presented in *Table 6.8*.

*Table 6.8: The Identification of Clinical Investigations in Routinely Recorded Datasets*

<b>Code</b>	<b>Description</b>	<b>Total Records</b>
<b><i>CT Brain</i></b>		
HES: 12	Computed Tomography	17
SAIL: 201	Computed Tomography	4
READ: 567	Computed Tomography	2
READ: Y72JD	CT Brain	1
READ: YAMGZ	CT Brain Normal	2
READ: 5674	CT Skull	1
<b><i>MRI Brain</i></b>		
READ: 569	Magnetic Resonance: (Imaging) or (Study)	3
READ: Y7213	MRI of Brain	2
READ : 5692	Nuclear Magnetic Resonance Scan: Normal	3
READ: 5693	Nuclear Magnetic Resonance Scan: Abnormal	1
<b><i>EEG</i></b>		
X77iL	Video EEG	1
31130	EEG normal	6
31C	EEG Observations	1

### **6.13 Magnetic Resonance Imaging (MRI)**

In the SANAD II Dataset, 72 participants underwent MRI and data were missing for seven participants, excluded from this analysis. In the HES and SAIL Emergency Datasets no records of MRI were identified. In the SAIL Primary Care Dataset nine records of MRI were identified in seven participants. Results were available for two participants and were consistent with the results recorded in the SANAD II dataset. In two of the seven participants an MRI was recorded in the Primary Care Dataset and not the SANAD II Dataset. In seven of the nine records, anatomical area was not specified and the codes included 'Magnetic Resonance Imaging or Study' and 'Nuclear Magnetic Resonance Scan: Normal'.

Agreement has been assessed using cross tabulation and calculation of Cohens Kappa. The value of Kappa is -0.016 (P=0.602) indicating no agreement between the status of MRI determined from SANAD II and routine datasets. Results are presented in *Table 6.9 and 6.12*.

*Table 6.9: MRI: Cross-Tabulation*

RCT Data	Routine Data		
		MRI Performed	MRI Not Performed
MRI Performed	5 (5.5%)	67 (73.6%)	72 (79.1%)
MRI Not Performed	2 (2.2%)	17 (18.7%)	19 (20.9%)
Total	7 (7.7%)	84 (92.3%)	91

### **6.14 Computed Tomography (CT)**

In the SANAD II Dataset, 33 participants underwent CT and data were missing for seven participants, excluded from this analysis. In the HES Emergency Dataset 17 records of CT were identified in 17 participants, although results were not available. In seven of the 17 participants a CT was recorded in the HES Emergency Dataset but not the SANAD II Dataset. In the SAIL Datasets, 10 records of CT were identified. Four records of CT were recorded in four participants in the SAIL Emergency Department Dataset, of these two participants did not have CT recorded in the SANAD II Dataset. Six records of CT in six participants were recorded in the SAIL Primary Care Dataset, of these two had results available that were in agreement with results recorded in the SANAD II Dataset. In total, 23 of the 27 records anatomical area was not specified and the code 'Computed Tomography' was recorded.

Agreement has been assessed using cross tabulation and calculation of Cohens Kappa. The value of Kappa is 0.406 ( $P < 0.001$ ) indicating 'fair' agreement between the status of CT determined from SANAD II and routine datasets. Results are presented in *Table 6.10 and 6.12*.

*Table 6.10: CT: Cross-Tabulation*

RCT Data	Routine Data			
		CT Performed	CT Not Performed	Total
CT Performed		18 (19.8%)	15 (16.5%)	33 (36.3%)
CT Not Performed		9 (9.9%)	49 (53.8%)	58 (63.7%)
Total		27 (29.7%)	64 (70.3%)	91

### **6.15 Electroencephalogram (EEG)**

Twenty three participants in the SANAD II Dataset were included in the SAIL Primary Care Dataset. 18 participants underwent EEG and data were missing for one participant, excluded from this analysis. In the SAIL Primary Care Dataset eight records of EEG were identified. In seven participants EEG was also requested in the SANAD II Dataset. In one participant, there was no record of EEG in the SANAD II Dataset. Results were available in six participants in the Primary Care Dataset and were in agreement with the results recorded in the SANAD II Dataset.

Agreement has been assessed using cross tabulation and calculation of Cohens Kappa. The value of Kappa is 0.188 (P=0.131) indicating 'none to slight' agreement between the status of EEG determined from SANAD II and routine datasets. Results are presented in *Table 6.11 and 6.12.*

*Table 6.11: EEG: Cross-Tabulation*

<b>RCT Data</b>	<b>Routine Data</b>		
		<b>EEG Performed</b>	<b>EEG Not Performed</b>
<b>EEG Performed</b>	7 (31.8%)	11 (50%)	18 (81.8%)
<b>EEG Not Performed</b>	0	4 (18.2%)	4 (18.2%)
<b>Total</b>	7 (31.8%)	15 (68.2%)	22

Table 6.12: Clinical Investigations: Descriptive Statistics and Agreement

			<b>MRI</b>	<b>CT</b>	<b>EEG</b>
<b>RCT Data Patients</b>	Total Investigations		72 (73.5%)	33 (33.7%)	18 (78.2%)
	Total Eligible Patients		98	98	23
<b>Routine Data Patients</b>	Total Investigations		9 (9.2%)	27 (27.6%)	8 (34.8%)
	Total Eligible Patients		98	98	23
<b>Total Paired Patients</b>			91 (92.9%)	91 (92.9%)	22 (95.7%)
<b>Routine Datasets</b>	Total 'Investigations' in Dataset	HES: Accident and Emergency	0	17	N/A
		SAIL: Emergency	0	4	N/A
		SAIL: Primary Care	9	6	8
<b>Assessment of Agreement</b>		Cohens Kappa	-0.016	0.406	0.188
		Significance	P=0.602	P<0.001	P=0.131
		Interpretation	'None'	'Fair'	'None to Slight'

### **6.16 Conclusions: Clinical Investigations**

These results indicate that the occurrence and results of clinical investigations are poorly recorded in routine datasets. In 10 participants across the three investigations, results were recorded in only the Primary Care Dataset and were in agreement with results recorded in SANAD II. The occurrence of investigations could be identified in the Emergency and Primary Care Datasets. For MRI, CT and EEG there were instances where investigations were recorded in the routine datasets but not SANAD II. This may be explained by the investigations not being recorded during the trial assessments and therefore not being recorded in the SANAD II dataset. However, in the majority the anatomical site was not recorded and therefore the assumptions necessarily made result in such data having limited specificity and utility for clinical practice and research.

MRI Brain, EEG and to some extent CT Brain are unlikely to be performed in either the emergency department or primary care setting and are often performed in secondary care, either during an inpatient admission or on an outpatient basis. The ICD coding system used in secondary care includes relevant codes regarding investigations but investigations are not included in the Inpatient or Outpatient Datasets retrieved through NHS Digital or SAIL in this study. Expanding the Inpatient and Outpatient Datasets to include data regarding investigations would likely result in more complete data regarding the occurrence and results of investigations.

## **6.17 Conclusions**

In this chapter, the identification of seizure occurrence, diagnosis and classification of epilepsy and seizures and identification of the occurrence of clinical investigations in routinely recorded datasets compared to data collected using standard prospective methods have been examined, relevant to the identification and recruitment of individuals eligible for inclusion in SANAD II.

A first seizure occurrence could be identified in routinely recorded datasets in 23 of the 98 participants. For participants without a first seizure occurrence; approximately one third of participants had no relevant attendances, one third had a 'relevant attendance', defined as an attendance within 48 hours of the date of a definite seizure recorded in SANAD II, but with missing diagnostic information and one third had a relevant attendance within 48 hours but with inadequate or discrepant diagnostic codes not meeting the criteria for seizure occurrence. For the limited number of participants where first seizures were identified, the agreement for the date of occurrence compared to the date collected using standard prospective methods in SANAD II was poor. Similarly, a 'diagnosis of epilepsy' was present in less than half of the participants using the routinely recorded data and agreement for the date of diagnosis was poor. Furthermore, there was poor agreement for the classification of seizures, explained by the majority of participants being deemed 'unclassified' as a result of the record of codes with inadequate clinical detail. Finally, data regarding the occurrence of the clinical investigations MRI, CT and EEG were available only in the emergency and primary care datasets, with reasonably complete data and results available only in the primary care dataset. Additionally, there was evidence of investigations and results recorded in the primary care dataset that had not been recorded in the SANAD II dataset.

The poor quality and agreement of routinely recorded data compared to data collected using standard prospective methods has implications for the utility of routinely recorded data. In epilepsy research, routinely recorded data may be limited for the identification of eligible individuals for recruitment into prospective trials. In clinical practice, the missing routinely recorded data is perhaps of greater importance, with impacts on the incidence and prevalence rates if data are used for disease monitoring purposes. Explanations for these findings may include inaccurate recording of codes in routinely recorded datasets or inaccurate initial clinical diagnosis of seizures and epilepsy. Furthermore, the events may

not have been 'recordable', for example if participants did not seek medical attention following seizure occurrence or if relevant codes or detail are not included in the available routinely recorded datasets, as with the occurrence and results of clinical investigations.

The greatest potential of routinely recorded data may be in providing collateral clinical information, in addition to the primary dataset recorded using standard prospective methods. Considering the results thus far, the accuracy and reliability of such data must be questioned; however, knowledge would direct further assessment within the trial, either through source data verification or clarification with the individual participant.

This assessment has notable limitations. The comparator dataset was derived from the SANAD II data available at the time of assessment and a minority of data entries may have been subject to data checking and confirmation. The pre-specified clinical limits of agreement were defined following clinical discussion and although not the primary purpose, may account for this. For example, an uncertain day of seizure, but known month and year would be recorded as occurring on the 15th day of the month in the SANAD II dataset. The one month clinical limit of agreement, whilst defined to specify the acceptable limit of disagreement that would be clinically acceptable, would also account for this uncertainty. The variables and constructed outcomes derived from the routinely recorded datasets were defined and extracted using algorithms developed for each comparison. However, there is a risk that relevant clinical events may not have been identified, if the diagnostic code recorded is not included in the algorithm. To address this limitation and explore the data further, the routinely recorded data for each participant was examined individually and in detail. For example, diagnostic codes recorded within 48 hours of a seizure recorded in the SANAD II dataset were examined. This process was feasible as a result of the small sample size. The assessment of clinical investigations used only ICD 10 diagnostic codes. The omission of an assessment of OPCS-4 codes represents a methodological limitation, contributing to the poor quality and agreement identified. Finally, there were limitations with regards to the specific variables that could be compared, although this was more a limitation of the routinely recorded data rather than the methodological process.

In the following Chapter Seven, variables and outcome measures relevant to the follow-up of participants in SANAD II will be examined.

## **Chapter Seven**

### **Results: Data Variables and Outcomes Relevant to the Follow-Up in SANAD II, Healthcare Resource Use and Primary Care Data**

#### **7.1 Introduction**

In this chapter, an assessment of the quality of routinely recorded data and agreement between routinely recorded data and data collected using standard prospective methods for variables and outcome measures relevant to the follow-up of participants in SANAD II was completed. Relevant variables and outcome measures dependent on the assessment of seizure occurrence included 'date of first seizure', 'time to first seizure' and 'time to 12 month remission'. Subsequently, the record of antiepileptic drug prescriptions, adverse events and healthcare resource use were assessed.

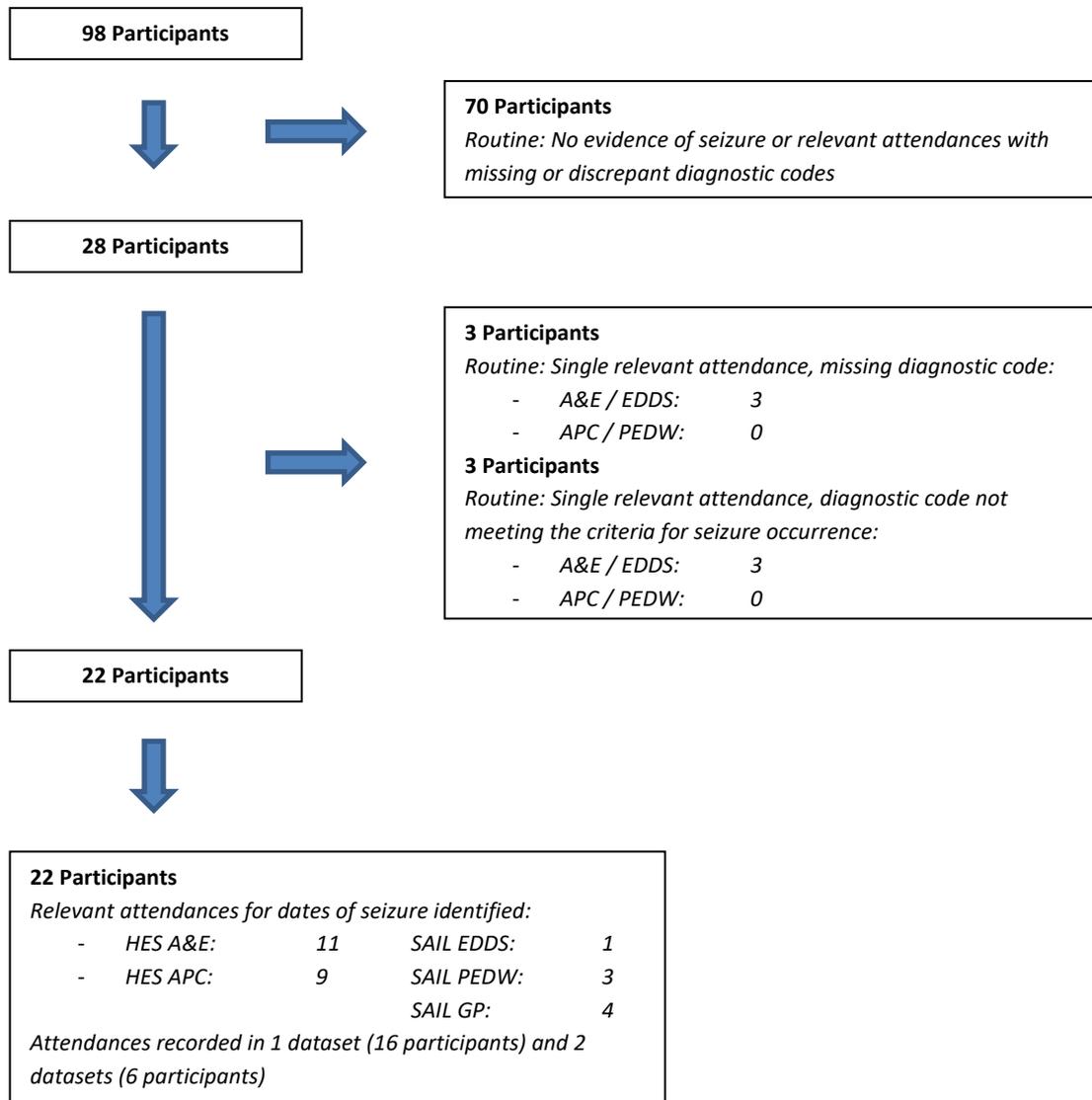
## **7.2 Date of First Follow-Up Seizure**

In the SANAD II dataset, 61 of the 98 participants experienced a seizure occurrence following the date of SANAD II randomisation.

In the routine datasets, a first follow-up seizure occurrence following the date of SANAD II randomisation was identified in 22 participants. The identification of relevant participants is presented in *Figure 7.1*. Of the 22 first follow-up seizure occurrences, 10 different diagnostic codes were recorded, although the most commonly recorded code remained the emergency code 'CNS Conditions, Epilepsy' occurring in six participants. The least specific emergency code 'CNS Disorder' occurred in three attendances, but in one a code from the inpatient dataset of greater quality was also recorded. In three participants without a first follow-up seizure in the SANAD II dataset, a first follow-up seizure occurrence was identified in routine datasets. In all participants, the seizures were recorded within two months of the baseline assessment date. All seizure occurrences identified in datasets using the ICD 10 coding system were classified as 'definite' using the developed algorithm.

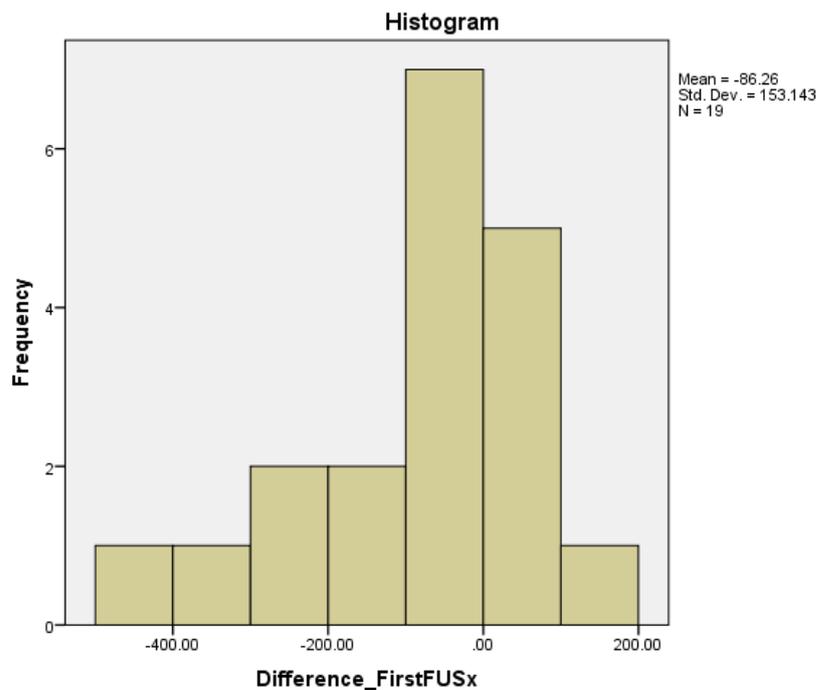
Three participants had a relevant attendance within 48 hours of the date of a definite seizure recorded in SANAD II but with diagnostic codes not meeting the criteria for seizure occurrence. All such participants had an attendance recorded in the emergency datasets. Codes included two records of 'CNS, Non-Epilepsy' and one record of 'Fracture' in the emergency datasets.

Figure 7.1: The Identification of the Date of First Follow-Up Seizure in Routine Datasets



The difference in days between the dates of first follow-up seizure from SANAD II subtracted from the dates of first follow-up seizure from routine datasets was calculated and the assumption of normal distribution around the mean assessed. The distribution is approximately normal on inspection, detailed in *Figure 7.2*.

*Figure 7.2: The Difference in Days Between the Date of First Follow-Up Seizure*

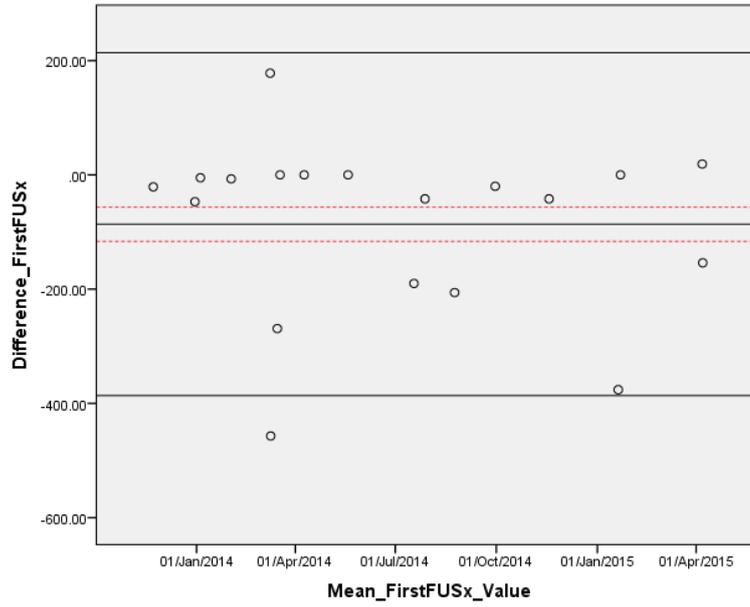


A Paired T Test has been performed in SPSS. The significant result ( $P=0.024$ ) indicates that the mean dates of first follow-up seizure calculated from SANAD II and routine datasets are significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 7.1* and *Figure 7.3*.

The Bland Altman Plot demonstrates that agreement is poor. The 95% confidence limits of agreement between the dates of first follow-up seizure are 214 and -386 days, clearly in excess of the specified 30 day clinically acceptable limit indicated by the red dashed lines. The mean of the difference between the dates is -86, indicating that the date of first follow-up seizure is identified in the SANAD II dataset a mean of 86 days earlier.

Figure 7.3: Date of First Follow-Up Seizure: Bland Altman Plot

Mean	-86.26
Upper 95% Confidence Limit of Agreement	213.89
Lower 95% Confidence Limit of Agreement	-386.41



### **7.3 Time to First Follow-Up Seizure**

Sixty one of the 98 participants experienced a seizure occurrence following the date of SANAD II randomisation in the SANAD II dataset. In the routine dataset, 22 participants experienced a seizure occurrence, of which three did not have a first follow-up seizure in the SANAD II dataset. Participants not experiencing a first follow-up seizure were censored using the final date routine data from all sources was available (31/12/15) or the date of last SANAD II follow-up assessment if the last follow-up assessment occurred prior to 31/12/15.

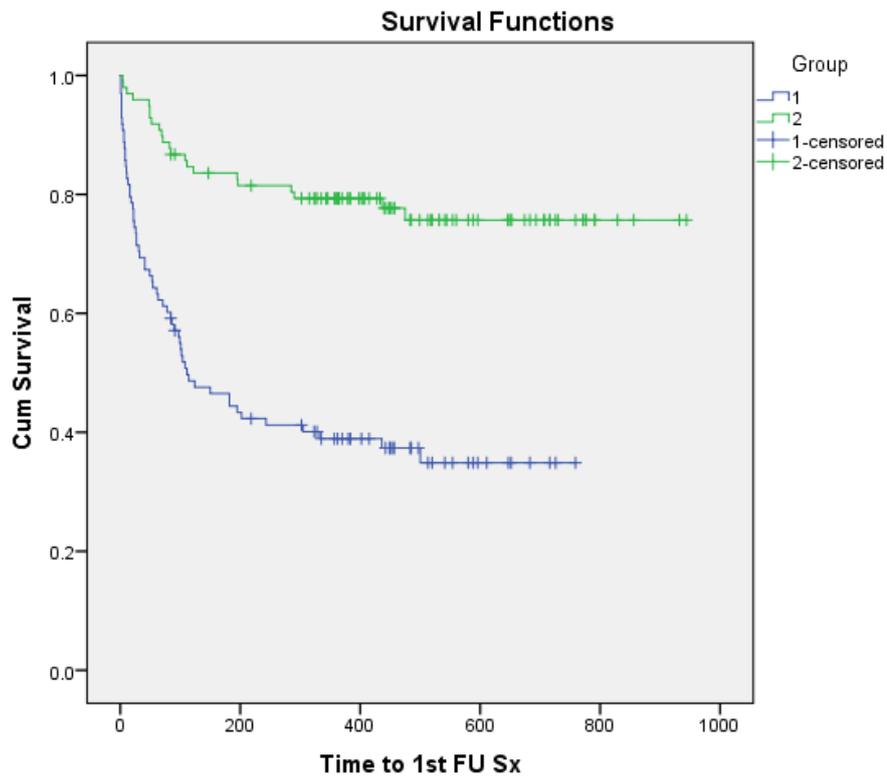
The 98 participants were included in a Kaplan Meier survival analysis, each with a time to first follow-up seizure calculated from the SANAD II and routine datasets. The mean time to first follow-up seizure was 325 days calculated using SANAD II data and 778 days calculated using routine data. The median was 111 days using the SANAD II data and could not be computed using the routine data. Descriptive statistics are presented in *Table 7.2*.

*Table 7.2: The Time to First Follow-Up Seizure: Descriptive Statistics*

	<b>Total: Included Patients</b>	<b>Total: Experiencing First Follow-Up Seizure</b>	<b>Total: Censored (%)</b>	<b>Mean (95% CI)</b>	<b>Median (95% CI)</b>
<b>RCT Data</b>	98	61	37 (37.8%)	325 (258-393)	111 (38-184)
<b>Routine Data</b>	98	22	76 (77.6%)	751 (680-822)	N/A

The difference in time to first follow-up seizure between datasets is statistically significant (Log Rank Test (Chi-Square 35.683),  $P < 0.001$ , *Figure 7.4*).

Figure 7.4: Kaplan Meier Curve: The Time to First Follow-Up Seizure



1 = SANAD II Dataset

2 = Routine Dataset

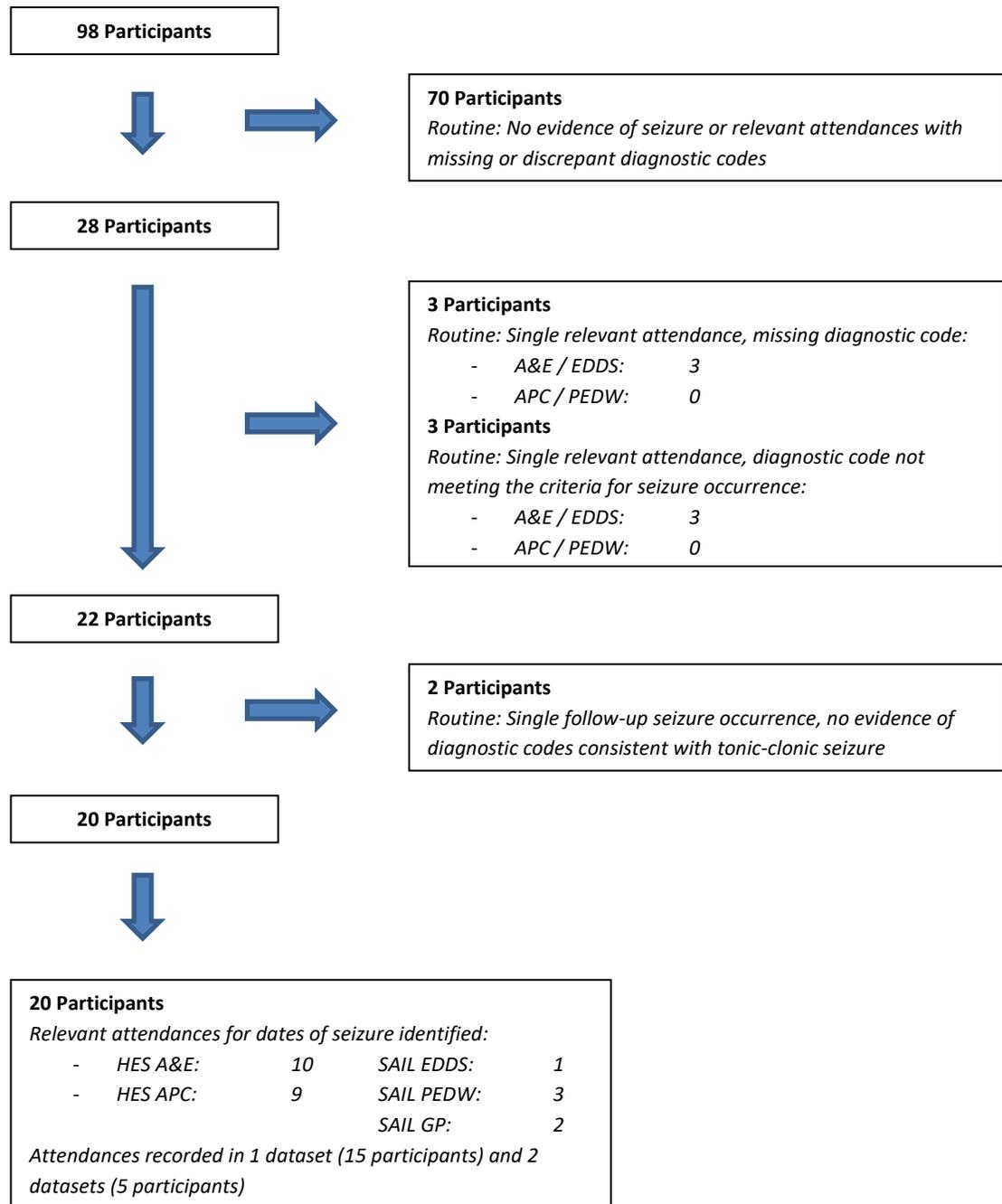
#### **7.4 Date of First Follow-Up Tonic-Clonic Seizure**

In the SANAD II dataset, 35 of the 98 participants experienced a tonic-clonic seizure occurrence following the date of SANAD II randomisation.

In the routine datasets a first follow-up tonic-clonic seizure occurrence following the date of SANAD II randomisation was identified in 20 participants. The identification of relevant participants is presented in *Figure 7.5*. Of the 20 first follow-up tonic-clonic seizure occurrences, seven different diagnostic codes were recorded, although the most commonly recorded codes remained the ICD code 'Epilepsy, Unspecified' occurring in seven participants and the emergency code 'CNS Conditions, Epilepsy' occurring in six participants. The least specific emergency code 'CNS Disorder' occurred in two attendances. In five participants without a first follow-up tonic-clonic seizure in the SANAD II dataset, a first follow-up tonic-clonic seizure occurrence was identified in routine datasets. In four cases, the ICD code 'Epilepsy, Unspecified' was recorded and in one case the READ code 'Fit in Known Epileptic'. All seizure occurrences identified in datasets using the ICD 10 coding system were classified as 'definite' using the developed algorithm.

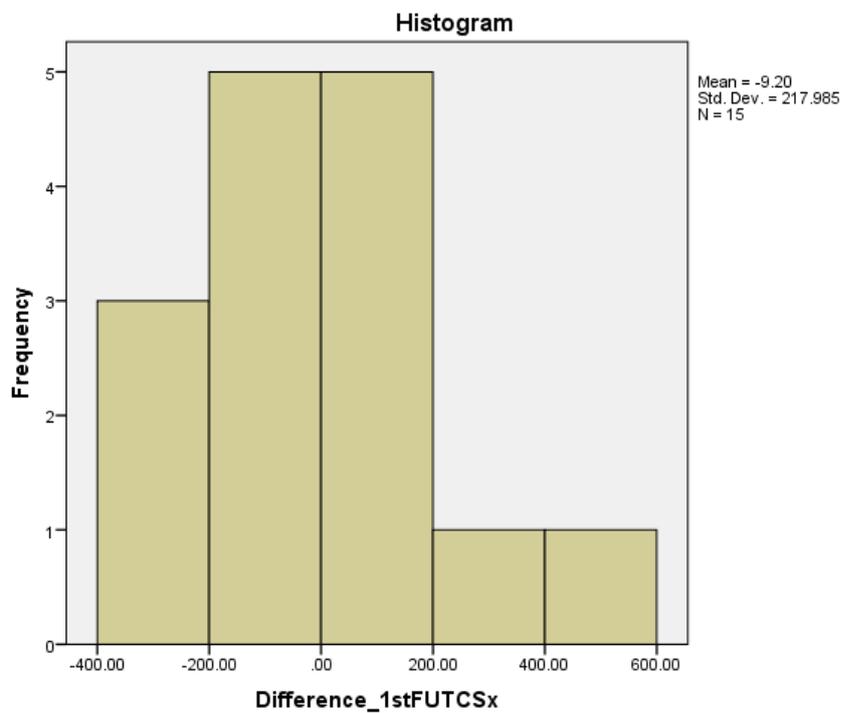
Three participants had a relevant attendance within 48 hours of the date of a definite seizure recorded in SANAD II but with diagnostic codes not meeting the criteria for seizure occurrence. All such participants had an attendance recorded in the emergency datasets. Codes included two records of 'CNS, Non-Epilepsy' and one record of 'Fracture' in the emergency datasets.

Figure 7.5: The Identification of the Date of First Follow-Up Tonic-Clonic Seizure in Routine Datasets



The difference in days between the dates of first follow-up tonic-clonic seizure from SANAD II subtracted from the dates of first follow-up tonic-clonic seizure from routine datasets was calculated and the assumption of normal distribution around the mean assessed. The distribution displays a positive skew inspection, detailed in *Figure 7.6*.

*Figure 7.6: The Difference in Days Between the Date of First Follow-Up Tonic-Clonic Seizure*



A Wilcoxon Signed Rank Test has been performed in SPSS. The non-significant result ( $P=0.374$ ) indicates that the mean dates of first follow-up tonic-clonic seizure calculated from SANAD II and routine datasets are not significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 7.1* and *Figure 7.7*.

The Bland Altman Plot demonstrates that agreement is poor. The 95% confidence limits of agreement between the dates of first follow-up seizure are 418 and -436 days, clearly in excess of the specified 30 day clinically acceptable limit indicated by the red dashed lines. The mean of the difference between the dates is -9, indicating that the date of first follow-up seizure is identified in the SANAD II dataset a mean of 9 days earlier.



### **7.5 Conclusions: Date of First Follow-Up and First Tonic-Clonic Follow-Up Seizures**

Follow-up seizures are poorly recorded in the routinely recorded datasets. The limitations identified in the assessment of 'first' seizures, prior to diagnosis of epilepsy are present but of greater severity in this assessment of follow-up seizures. Attendances as a result of follow-up seizure occurrence are either missing or present but including diagnostic codes not consistent with seizure.

For the limited number of participants where follow-up seizures were identified, the agreement compared to data collected using standard prospective methods in SANAD II is poor, with seizures identified in the SANAD II dataset earlier. This is despite in some cases there being no significant difference identified between the means calculated using routinely recorded data and data collected using standard prospective methods. However, the limited sample size explains the lack of power to detect a significant difference for these variables. The degree of missing data and delay in identification of first follow-up seizure occurrence has implications for the utility of routinely recorded data. In research, routinely recorded data is not suitable for measurement of outcomes in prospective studies such as SANAD II. In clinical practice, routinely recorded data is not suitable for monitoring treatment effectiveness and the missing data may impact on the incidence and prevalence rates if data are used for disease monitoring purposes on a population level.

Explanations for these results may include inaccurate recording of diagnostic codes in routinely recorded datasets or inaccurate clinical diagnosis of seizure. Furthermore, participants may not have sought medical attention following seizure occurrence. This may be more likely in participants with established epilepsy or those experiencing focal seizures.

Table 7.1: The First and First Tonic-Clonic Follow-Up Seizures: Descriptive Statistics and Agreement

			<b>First Follow-Up Seizure (All Types)</b>	<b>First Follow-Up Tonic-Clonic Seizure</b>
<b>RCT Data Patients</b>		Patients with Seizures	61 (62.2%)	35 (35.7%)
		Total Eligible Patients	98	98
<b>Routine Data Patients</b>		Patients with Seizures	22 (22.4%)	20 (20.4%)
		Total Eligible Patients	98	98
		<b>Total Paired Patients</b>	<b>19 (19.3%)</b>	<b>15 (15.3%)</b>
<b>Routine Datasets</b>	Total Seizures in Dataset (Total Seizures in 'Greatest Detail: Mutually Exclusive)	HES: Admitted Patient Care	9(9)	9(9)
		HES: Accident and Emergency	11(6)	10(6)
		HES: Outpatient	0(0)	0(0)
		HES: Adult Critical Care	0(0)	0(0)
		SAIL: Patient Episode Database for Wales	3(3)	3(3)
		SAIL: Emergency	1(0)	1(0)
		SAIL: Outpatient	0(0)	0(0)
		SAIL: Primary Care	4(4)	2(2)
<b>Assessment of Agreement</b>	<b>RCT Data</b>	Mean	18/05/14	10/07/14
		Range	24/07/13 – 16/04/15	01/11/13 – 02/07/15
	<b>Routine Data</b>	Mean	07/08/14	25/08/14
		Range	03/12/13 – 27/07/15	10/12/13 – 27/07/15
		Test for Significance	Paired T Test	Wilcoxon Signed Rank
		Significance	P=0.024	P=0.374

## **7.6 Date 12 Month Remission Achieved**

In the SANAD II dataset, 39 participants had episodes of seizure occurrence preventing the achievement of 12 month remission throughout the duration of follow-up. For participants without record of the occurrence of seizures, 11 participants had less than 12 months of SANAD II follow-up and therefore 12 month remission could not be achieved. Two participants had no evidence of the occurrence of seizures but with less than 12 months of available equivalent routine data and therefore 12 month remission could not be achieved. Forty six of the total 98 participants achieved 12 month remission from seizures in the SANAD II dataset.

In the routine datasets, three participants had episodes of seizure occurrence preventing the achievement of 12 month remission throughout the duration of follow-up. For participants without record of the occurrence of seizures, 14 participants had less than 12 months of equivalent SANAD II follow-up and therefore 12 month remission could not be achieved. Seven participants had no evidence of the occurrence of seizures but with less than 12 months of available routine data and therefore 12 month remission could not be achieved. Seventy four of the total 98 participants achieved 12 month remission from seizures in the routine datasets.

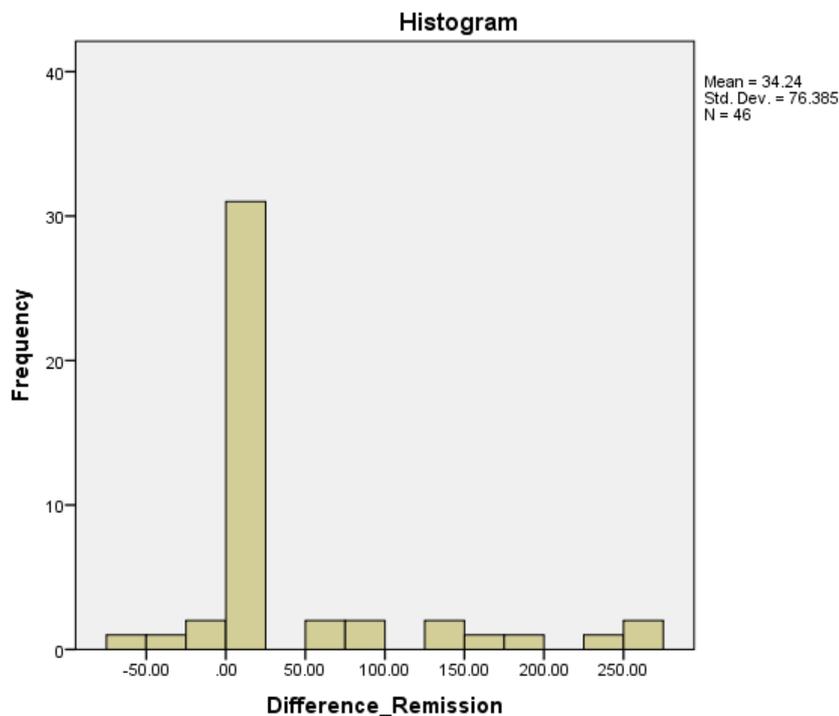
For participants achieving remission in the SANAD II dataset, remission was also achieved in the routine datasets. However, an additional 28 participants achieved remission in the routine datasets. *Table 7.3* summarises the identification of the date 12 month remission is achieved.

*Table 7.3: The Identification of the Date 12 Month Remission Achieved*

	<b>SANAD II Dataset</b>	<b>Routine Datasets</b>
<b>Total Participants</b>	98	98
<b>Participants Not Achieving Remission:</b>		
<b><i>Occurrence of Seizures</i></b>	39 (39.8%)	3 (3.1%)
<b><i>No Occurrence of Seizures:</i></b>		
- Insufficient SANAD II Follow-Up (<12 months SANAD II follow-up)	11 (11.2%)	14 (14.3%)
- Insufficient Routine Data ('Remission' occurring >31/12/15)	2 (2.0%)	7 (7.1%)
<b>Participants Achieving Remission</b>	46 (46.9%)	74 (75.5%)

The difference in days between the dates that 12 month remission is achieved identified from the SANAD II dataset was subtracted from the dates identified from the routine datasets and the assumption of normal distribution around the mean assessed. A spike-at-zero distribution is observed, detailed in *Figure 7.8*.

*Figure 7.8: The Difference in Days Between the Date 12 Month Remission Achieved*



A Wilcoxon Signed Rank Test has been performed in SPSS. The significant result ( $P=0.004$ ) indicates that the mean dates that 12 month remission is achieved calculated from the SANAD II and routine datasets are significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 7.4* and *Figure 7.9*.

The Bland Altman Plot demonstrates that agreement is poor. The 95% confidence limits of agreement between the dates 12 month remission is achieved are 184 and -115 days, clearly in excess of the specified 30 day clinically acceptable limit indicated by the red dashed lines. The mean of the difference between the dates is 34, indicating that the date 12 month remission is achieved is identified in the routine datasets a mean of 34 days earlier.

Figure 7.9: The Date 12 Month Remission Achieved: Bland Altman Plot

Mean	34.24
Upper 95% Confidence Limit of Agreement	183.96
Lower 95% Confidence Limit of Agreement	-115.48

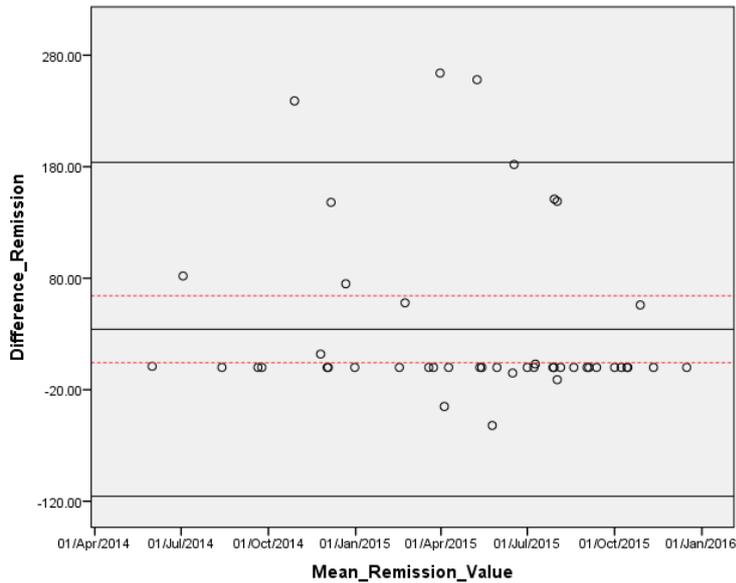


Table 7.4: The Date 12 Month Remission Achieved: Descriptive Statistics and Agreement

			Date 12 Month Remission Achieved
RCT Data Patients		Patients Achieving Remission	46 (46.9%)
		Total Eligible Patients	98
Routine Data Patients		Patients Achieving Remission	74 (75.5%)
		Total Eligible Patients	98
		Total Paired Patients	46 (46.9%)
Assessment of Agreement	RCT Data	Mean	17/05/15
		Range	01/06/14 – 16/12/15
	Routine Data	Mean	18/04/15
		Range	23/05/14 – 27/12/15
		Test for Significance	Wilcoxon Signed Rank
		Significance	P=0.004

### **7.7 Time to 12 Month Remission**

Participants not achieving 12 month remission were censored using the final date routine data from all sources was available (31/12/15) or the date of last follow-up assessment if the last follow-up assessment occurred prior to 31/12/15.

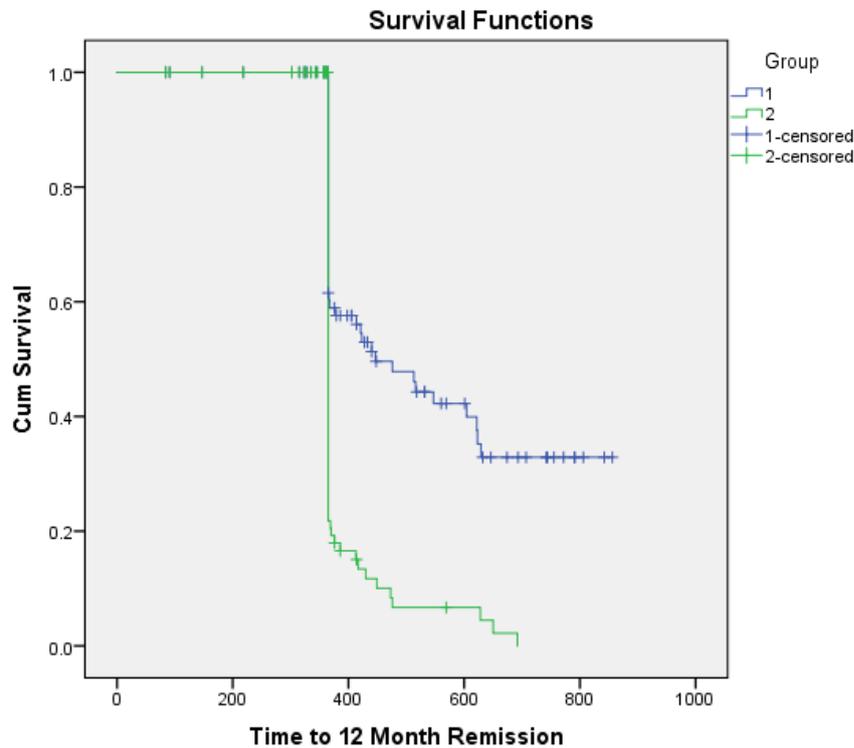
The 98 participants were included in a Kaplan Meier survival analysis, each with a time to 12 month remission calculated from the SANAD II and routine datasets. The mean time to 12 month remission was 567 days calculated using SANAD II data and 393 days calculated using routine data. Descriptive statistics are presented in *Table 7.5*.

*Table 7.5: The Time to 12 Month Remission: Descriptive Statistics*

	<b>Total: Included Patients</b>	<b>Total: Achieving 12 Month Remission</b>	<b>Total: Censored (%)</b>	<b>Mean (95% CI)</b>	<b>Median (95% CI)</b>
<b>RCT Data</b>	98	46	52 (26.5%)	567 (515-618)	447
<b>Routine Data</b>	98	74	24 (12.2%)	393 (375-410)	365

The difference in the time to 12 month remission between datasets is statistically significant (Log Rank Test (Chi-Square 38.466),  $P < 0.001$ , *Figure 7.10*).

Figure 7.10: Kaplan Meier Curve: The Time to 12 Month Remission



1 = SANAD II Dataset  
2 = Routine Dataset

### **7.8 Conclusions: Date 12 Month Remission is Achieved and Time to 12 Month Remission**

As identified, follow-up seizures are poorly recorded in the routinely recorded datasets with missing data and poor agreement observed. The results for these comparisons assessing 12 month remission reflect these earlier findings, with significantly more participants achieving 12 month remission using the routine datasets. These findings have significant implications. In research, assuming routinely recorded data alone were used to measure the SANAD II primary outcome of time to 12 month remission, the lack of recorded seizures would indicate improved treatment effectiveness for all antiepileptic drug interventions, constituting a type one error.

## **7.9 Total Number of Follow-Up Seizures**

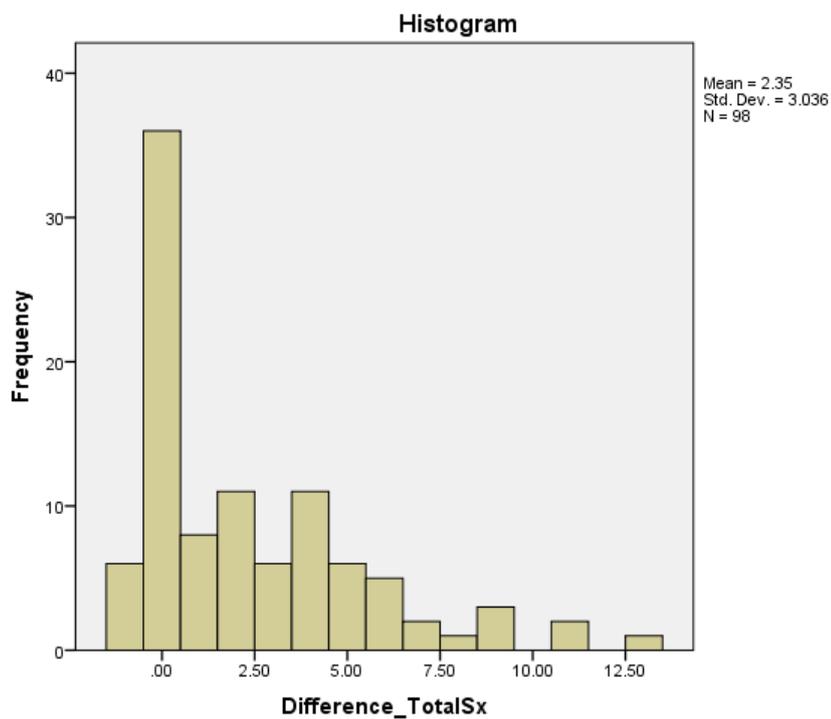
### ***7.9.1 Total Number of Follow-Up Seizures (All Seizure Types)***

In the SANAD II dataset, 61 participants experienced a total of 258 follow-up seizures (Range: 0-13).

In the routine datasets, 22 participants experienced a total of 28 follow-up seizures (Range: 0-5).

The difference between total number of seizures from SANAD II subtracted from total number from routine datasets was calculated and the assumption of normal distribution around the mean assessed. The distribution displays a positive skew on inspection, detailed in *Figure 7.11*.

*Figure 7.11: The Difference in Total Number of Follow-Up Seizures*

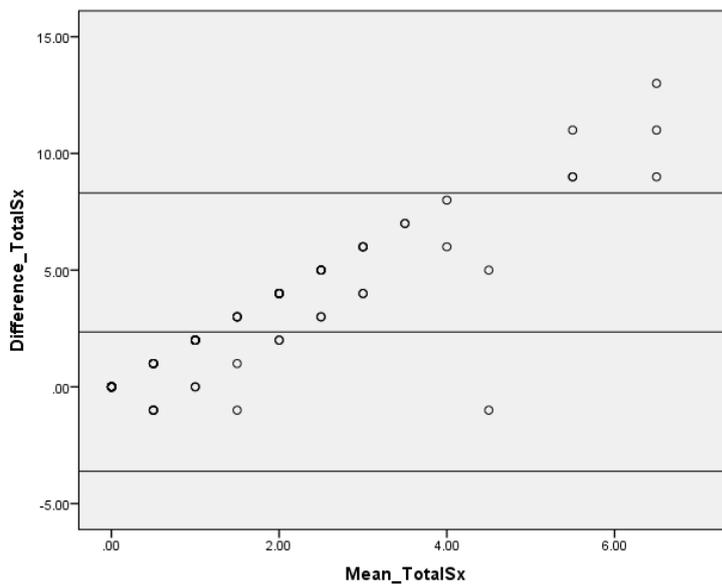


A Wilcoxon Signed Rank Test has been performed in SPSS. The significant result ( $P < 0.001$ ) indicates that the mean total number of follow-up seizures calculated from the SANAD II and routine datasets are significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 7.6* and *Figure 7.12*.

A positive trend is observed with an increasing mean associated with an increasing difference. This is explained as the majority of participants in the routinely recorded datasets had no evidence of seizure occurrence. Therefore, with increasing numbers of seizures identified in the SANAD II dataset, both the calculated mean and difference were greater. The 95% confidence limits of agreement between the total numbers of follow-up seizures are 8.3 and -3.6. The mean of the difference is 2.4, indicating that a mean 2.4 more follow-up seizures are identified in SANAD II.

*Figure 7.12: The Total Number of Follow-Up Seizures: Bland Altman Plot*

<i>Mean</i>	2.35
<i>Upper 95% Confidence Limit of Agreement</i>	8.31
<i>Lower 95% Confidence Limit of Agreement</i>	-3.61



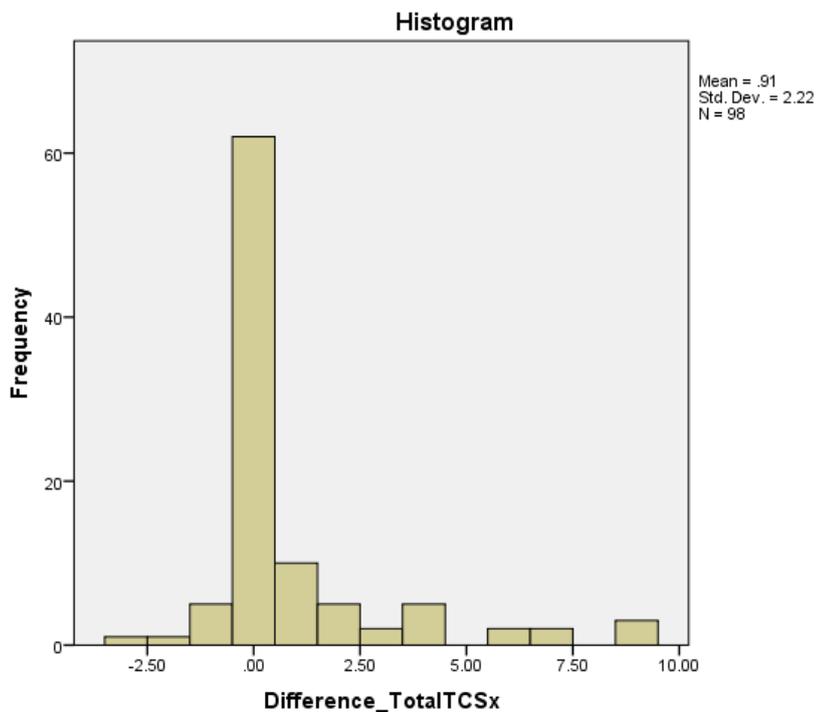
### 7.9.2 Total Number of Follow-Up Tonic-Clonic Seizures

In the SANAD II dataset, 35 participants experienced a total of 113 follow-up tonic-clonic seizures (Range: 0-11).

In the routine datasets, 20 participants experienced a total of 24 follow-up tonic-clonic seizures (Range: 0-3).

The difference between total number of seizures from SANAD II subtracted from total number from routine datasets was calculated and the assumption of normal distribution around the mean assessed. A spike-at-zero distribution is observed, detailed in *Figure 7.13*.

*Figure 7.13: The Difference in Total Number of Follow-Up Tonic-Clonic Seizures*



A Wilcoxon Signed Rank Test has been performed in SPSS. The significant result ( $P < 0.001$ ) indicates that the mean total number of follow-up tonic-clonic seizures calculated from the SANAD II and routine datasets are significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 7.6* and *Figure 7.14*.

A positive trend is again observed with an increasing mean associated with an increasing difference. The Bland Altman Plot demonstrates poor agreement although agreement is marginally improved compared to the assessment of all seizure types. The 95% confidence limits of agreement between the total numbers of follow-up seizures are 5.3 and -3.4. The mean of the difference is 0.9, indicating that a mean 0.9 more follow-up seizures are identified in SANAD II.

Figure 7.14: The Total Number of Follow-Up Tonic-Clonic Seizures: Bland Altman Plot

Mean	0.91
Upper 95% Confidence Limit of Agreement	5.26
Lower 95% Confidence Limit of Agreement	-3.44

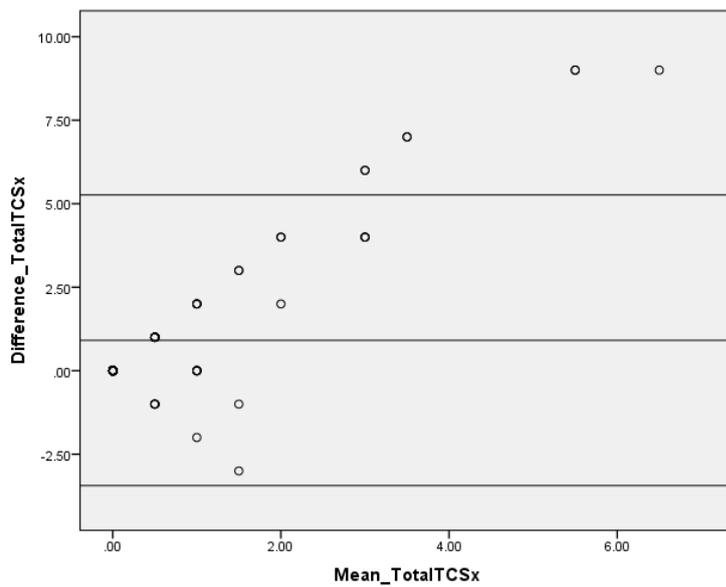


Table 7.6: The Total Number of Follow-Up Seizures: Descriptive Statistics and Agreement

			Follow-Up Seizures (All Types)	Follow-Up Seizures (Tonic-Clonic)
RCT Data Patients		Patients with Seizures	61 (62.2%)	35 (35.7%)
		Total Eligible Patients	98	98
Routine Data Patients		Patients with Seizures	22 (22.4%)	20 (20.4%)
		Total Eligible Patients	98	98
		Total Paired Patients	98 (100%)	98 (100%)
Assessment of Agreement	RCT Data	Mean Number of Seizures	2.63	1.15
		Range	0 - 13	0 - 11
	Routine Data	Mean Number of Seizures	0.29	0.24
		Range	0 - 5	0 - 3
		Test for Significance	Wilcoxon Signed Rank	Wilcoxon Signed Rank
		Significance	P<0.001	P<0.001

### **7.10 Conclusions: Total Number of Follow-Up Seizures**

As identified, follow-up seizures are poorly recorded in the routinely recorded datasets with missing data and poor agreement observed. Notably, marginally improved agreement was identified for tonic-clonic seizures compared to all seizure types. This would be expected as tonic-clonic seizures would be more likely to result in hospital attendance. However, the 'true' underlying difference is likely to be of even greater magnitude as only 'definite' dates of seizures recorded in SANAD II were included in this comparison.

These results have significant implications, reducing the utility of routinely recorded data for monitoring seizure occurrence in clinical practice and research. However, it is the identification of seizure freedom rather than the record of total number of seizures that is of greater importance in both clinical practice, for example influencing driving restrictions and research, where measures of seizure freedom represent more common prospective outcomes.

## **7.11 The Assessment of Antiepileptic Drug Prescribing in the Routinely Recorded Datasets**

### **7.11.1 The Date of AED First Prescription**

In the SANAD II dataset, 26 first AED prescriptions were recorded in 23 participants, three participants had a second AED prescribed following treatment failure of the initial AED.

In the SAIL Primary Care Dataset, 25 first AED prescriptions were identified in 22 participants. The prescribed AEDs in all cases were in agreement between the datasets and are summarised in *Table 7.7*. The first prescription of AED could not be identified in one participant. In this case it is evident from the subsequent SANAD II follow-up assessment that the AED was withdrawn shortly following the initial SANAD II prescription, which explains this anomaly.

*Table 7.7: Antiepileptic Drugs in the SANAD II and SAIL Primary Care Dataset*

<b>Antiepileptic Drug</b>	<b>SANAD II Records</b>	<b>SAIL Primary Care Records</b>
Lamotrigine	9	8
Levetiracetam	9	9
Sodium Valproate	1	1
Zonisamide	7	7

The difference in days between the dates of AED first prescription identified from SANAD II subtracted from the dates identified from the Primary Care Datasets was calculated and the assumption of normal distribution around the mean assessed. The distribution is approximately normal to inspection, detailed in *Figure 7.15*.

A Paired T Test has been performed in SPSS. The significant result ( $P < 0.001$ ) indicates that the mean dates of first prescription of AED calculated from SANAD II and routine datasets are significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 7.8* and *Figure 7.16*.

The Bland Altman Plot demonstrates that despite the significant difference between the mean dates of AED first prescription calculated using the Paired T Test, there is acceptable agreement between the calculated dates. The 95% confidence limits of agreement between the dates of AED first prescription are -68 and 28 days, within the specified acceptable clinical limit of 90 days, indicated by the red dashed lines. The mean of the difference between the dates is -20, indicating that the date of AED first prescription is identified in the SANAD II dataset a mean of 20 days earlier.

Figure 7.15: The Difference in Date of AED First Prescription

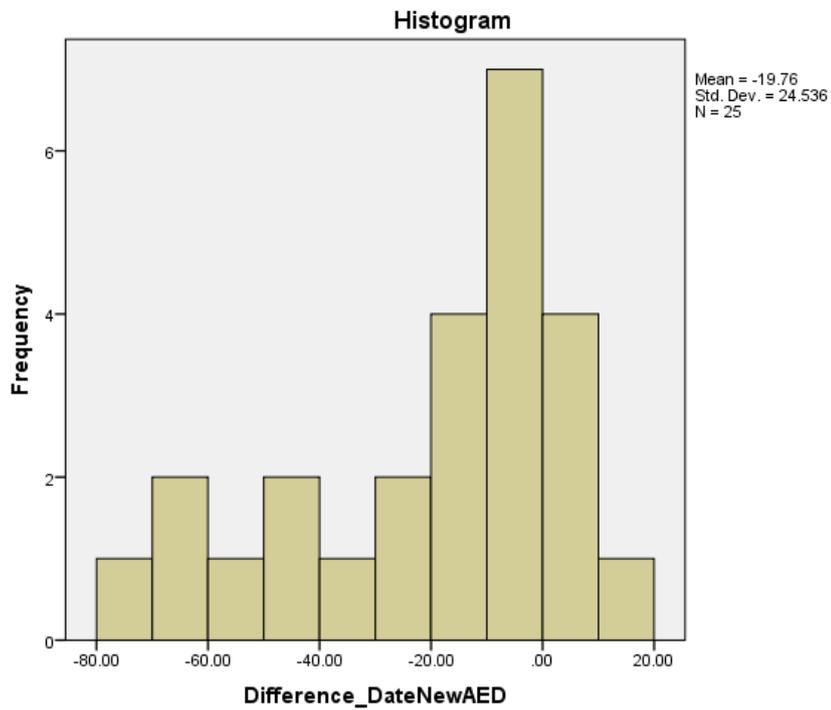
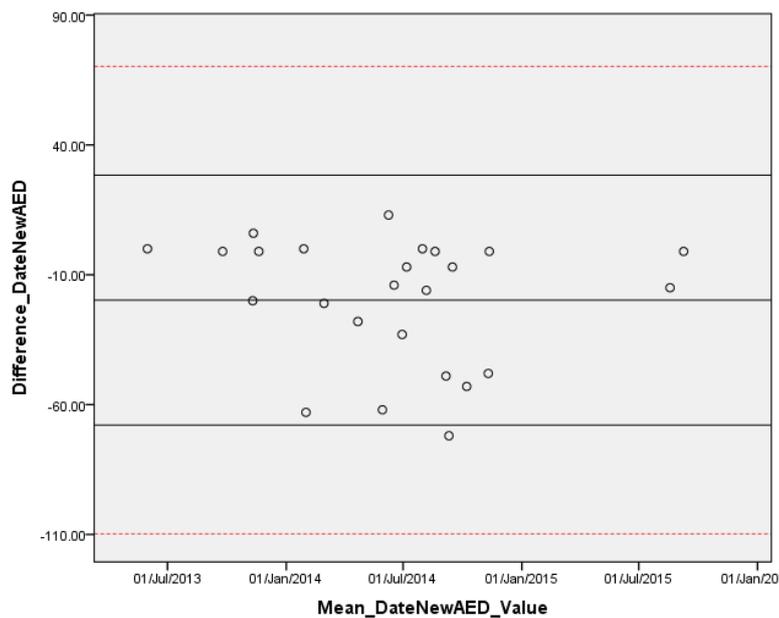


Figure 7.16: Dates of AED First Prescription: Bland Altman Plot

Mean	-19.76
Upper 95% Confidence Limit of Agreement	28.34
Lower 95% Confidence Limit of Agreement	-67.86



### 7.11.2 Compliance

Data regarding compliance are not recorded in SANAD II although can be inferred from the Primary Care Dataset based on regularity of repeat prescription (1, 2 or 3 monthly). In the SAIL Primary Care Dataset, 22 AED first prescriptions were followed by regular monthly repeat prescriptions. One AED first prescription was followed by regular monthly repeat prescriptions and subsequently, regular bi-monthly repeat prescriptions. Participants were 'not compliant' with two AED prescriptions. In both cases the AED was not issued for over 12 months yet in the SANAD II follow-up assessments the AED remained listed as prescribed. Explanations include poor compliance, an alternative source of the prescription of the AED or prescription of an alternative AED, although this was not identified in the available dataset. Finally, in one participant, an additional prescription of AED was identified in the Primary Care Dataset that was not recorded in the SANAD II dataset over a period of 12 months.

Table 7.8: First Prescriptions of Antiepileptic Drugs: Descriptive Statistics and Agreement

		<b>Antiepileptic Drugs</b>	
<b>RCT Data Patients</b>		Total First Prescriptions	26
		Total Patients	23
<b>Routine Data Patients</b>		Total First Prescriptions	25
		Total Patients	23
		Total Paired First Prescriptions	25
<b>Assessment of Agreement</b>	<b>RCT Data</b>	Mean	09/06/14
		Range	31/05/13 – 08/09/15
	<b>Routine Data</b>	Mean	25/06/14
		Range	31/05/13 – 09/09/15
		Test for Significance	Paired T Test
		Significance	P<0.001

### **7.12 Conclusions: The Assessment of Antiepileptic Drugs**

Data retrieved from the Primary Care Dataset regarding AEDs in this study were complete and had good agreement compared to data collected using standard prospective methods in SANAD II. The prescriptions of AEDs and dates of prescription could be identified in the Primary Care Dataset, although indication was absent. Additionally, data regarding compliance could be inferred and are potentially informative to clinical practice; for example informing the management of individuals with poor seizure control associated with poor compliance and research; for example informing the assessment of the treatment effectiveness of AEDs. Finally, perhaps the greatest benefit offered by routinely recorded prescribing data is the 'added information'. For example, one participant was prescribed a second AED for a period of over 12 months that was not recorded in SANAD II. Retrieving such data with potentially significant implications for confounding during the RCT may be informative to the analyses and interpretation of results.

### **7.13 The Assessment of Adverse Events in the Routinely Recorded Datasets**

In the SANAD II dataset, 44 participants reported 102 individual adverse events (Range: 1-9). Four participants reporting five adverse events during the time period following 31/12/15 were excluded from this comparison as a result of the lack of available equivalent routinely recorded data. The remaining 40 participants reported 97 adverse events (Range: 1-9). During the study period, there were no reported Serious Adverse Reactions (SAR's) or Suspected Unexpected Serious Adverse Reactions (SUSAR's).

In the routine datasets, two adverse events in two participants were identified. In the first participant, the adverse event 'Poorly Controlled Epilepsy' recorded in the SANAD II dataset correlated with a date of admission coded as 'Status Epilepticus' in the HES Inpatient Dataset. In the second participant, the adverse event 'Burning Sensation - Arms , Legs' correlated with a date of healthcare attendance coded as 'Pain in Leg' in the SAIL Primary Care Dataset. As a result of the limited sample size, further assessments of agreement were not performed. The adverse events identified and descriptive statistics are summarised in *Tables 7.9 and 7.10*.

Seven participants had 11 attendances within three months of a date of adverse event recorded in SANAD II. The dates of seven adverse events recorded in the SANAD II dataset were in agreement with an attendance recorded in the HES Accident and Emergency Dataset but with missing diagnostic information, while the dates of four adverse events were in agreement with an attendance recorded in the HES Accident and Emergency Dataset but with inadequate diagnostic information. For example, one participant with the adverse event 'Memory Problems' in the SANAD II dataset was in agreement with a date of attendance coded as 'CNS – other non-epilepsy' in the Accident and Emergency Dataset.

Potential alternative clinical explanations, other than adverse reaction to the AED, for three adverse events recorded in the SANAD II dataset were identified in routine datasets within three months of the recorded adverse event. The adverse event 'Sedation' was in agreement with a date of attendance coded as 'Alcohol' in the HES Accident and Emergency Dataset. The adverse event 'Nausea' was in agreement with a date of attendance coded as 'Gastritis' in the HES Inpatient Dataset. The adverse event 'Shortness of Breath' was in agreement with a date of attendance coded as 'Diaphragmatic Hernia' in the HES Inpatient Dataset.

Finally, there were no recorded codes consistent with 'adverse events' listed in the routinely recorded datasets.

#### **7.14 Conclusions: The Assessment of Adverse Events**

There is a significant magnitude of missing routinely recorded data regarding adverse events compared to data collected using standard prospective methods in SANAD II. Furthermore, diagnostic codes specifically indicating 'adverse events' were not recorded in any of the routinely recorded datasets. The utility of routinely recorded data for monitoring of adverse events in both clinical practice and research is therefore severely limited. Explanatory factors include the method of ascertainment in SANAD II – completion of self-reported questionnaires with options for free text in addition to suggested listed common adverse events. Such a method may encourage reporting of trivial adverse events that would otherwise not have resulted in a healthcare attendance. However, although trivial, such adverse events have the potential to exert a negative influence on quality of life. There were no serious adverse events recorded for the participants included in this study and in such cases the routinely recorded data could be expected to be more complete and informative.

These results indicate that routinely recorded data are inadequate for the identification of adverse events and have greater potential for providing collateral data, for example providing alternative explanations for clinical symptoms attributed to adverse events. Perhaps the most pragmatic suggestion for improvement would be the utilisation of the diagnostic codes already available, specifically indicating 'adverse event'.

Table 7.9: The Assessment of Adverse Events

Adverse Event	SANAD II Dataset: Total Events	Routine Dataset: Total Events
Weight Loss	4	0
Poorly Controlled Epilepsy	1	1
Dizziness	4	0
Skin Rash	3	0
Memory Problems	10	0
Irritability / Anger	13	0
Weight Gain	7	0
Sedation / Drowsiness	15	0
Gastro Intestinal Disturbance	10	0
Swelling of Face / Mouth	2	0
Chest Infection	2	0
Tinnitus	1	0
Deafness	1	0
Dry Mouth	1	0
Polydipsia	1	0
Polyuria	1	0
Low Mood	6	0
Insomnia	4	0
Headache	2	0
Vivid Dreams	1	0
Shortness of Breath	1	0
Low Platelet Count	1	0
Fever	1	0
Cramping / Tingling Hands	1	0
Hair Loss	1	0
Pruritus	1	0
Nose Bleed	1	0
Burning sensation / Arms, Legs	1	1
Tremor	1	0

Table 7.10: Adverse Events: Descriptive Statistics

		Adverse Events
<b>RCT Data Patients</b>	Total Adverse Events	97
	Total Patients	40
<b>Routine Data Patients</b>	Total Adverse Events	2
	Total Patients	40
Total Paired Adverse Events		2
<b>Routine Datasets</b>	Total Adverse Events in Dataset (Total Adverse Events in Greatest Detail: Mutually Exclusive)	
	HES: Admitted Patient Care	1(1)
	HES: Accident and Emergency	0(0)
	HES: Outpatient	0(0)
	HES: Adult Critical Care	0(0)
	SAIL: Patient Episode Database for Wales	0(0)
	SAIL: Emergency	0(0)
	SAIL: Outpatient	0(0)
SAIL: Primary Care	1(1)	

### **7.15 The Assessment of Healthcare Resource Use in the Routinely Recorded Datasets**

The SANAD II baseline and follow-up assessments, occurring in the majority during routine outpatient appointments are examples of ‘planned’ healthcare resource use. ‘Unplanned’ healthcare resource use includes emergency, inpatient and primary care attendances.

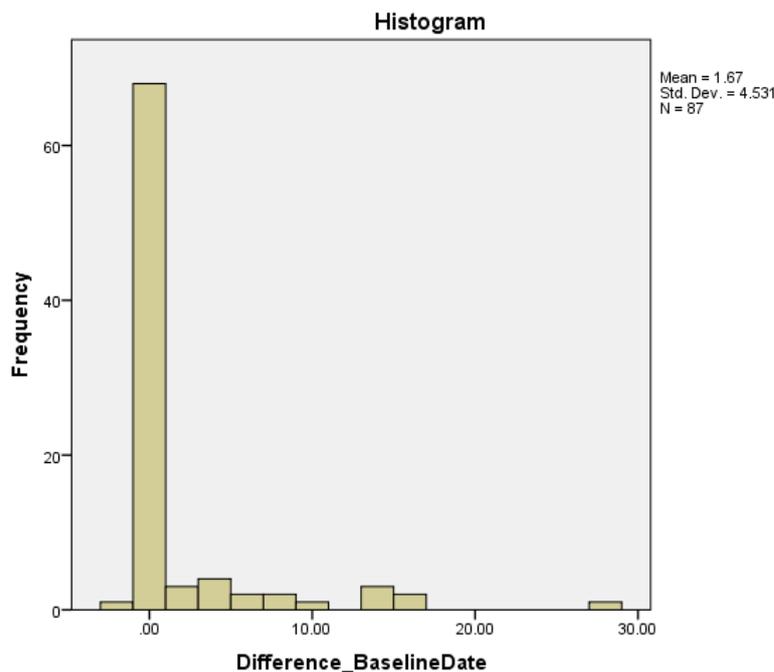
#### ***7.15.1 Planned Healthcare Attendances: The SANAD II Baseline Assessments***

All 98 participants enrolled in SANAD II have had the date of baseline assessment recorded.

In the routinely recorded datasets, the baseline assessments were identified in 87 participants. One baseline assessment was identified in the emergency care dataset, three assessments in the inpatient datasets and the remaining 83 assessments in the outpatient datasets. Eleven participants had no relevant healthcare attendance within the specified clinical limit of agreement of one month (30 days) in routine datasets. Five participants were resident in England and six in Wales. The missing data in these cases may be explained by the baseline assessment occurring out of the context of routine medical care, for example assessments may have taken place in dedicated research clinics that may not be included in routine datasets.

The difference in days between the dates of baseline assessment from SANAD II subtracted from the dates from routine datasets was calculated and the assumption of normal distribution around the mean assessed. A spike-at-zero distribution is observed, detailed in *Figure 7.17*.

Figure 7.17: The Difference in Days Between the Date of Baseline Assessment

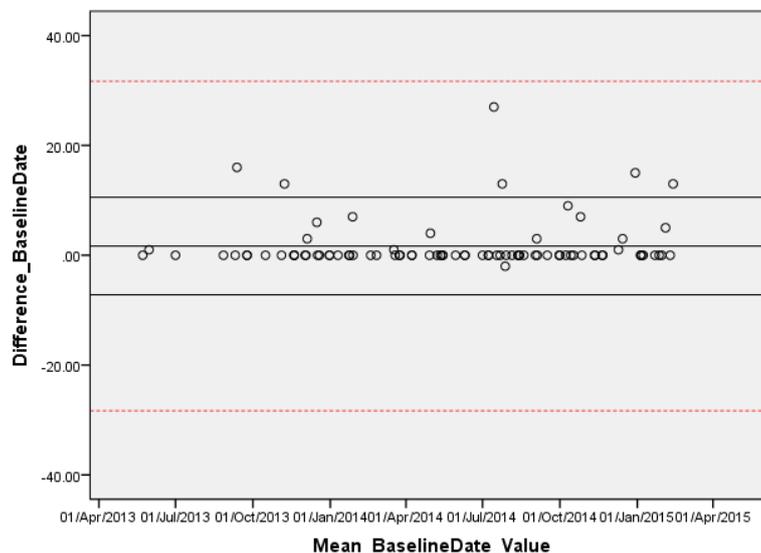


A Wilcoxon Signed Rank Test has been performed in SPSS. The significant result ( $P < 0.001$ ) indicates that the mean dates of baseline assessment calculated from SANAD II and routine datasets are significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 7.11* and *Figure 7.18*.

The Bland Altman Plot demonstrates that despite the significant difference between the mean dates of baseline assessment calculated using the Wilcoxon Signed Rank Test, there is acceptable agreement between the calculated dates. The 95% confidence limits of agreement between the dates of baseline assessment are -7 and 11 days, within the specified acceptable clinical limit of agreement of one month (30 days), indicated by the red dashed lines. The mean of the difference between the dates is 1.67, indicating that the date of baseline assessment is identified in routine datasets a mean 1.67 days earlier. A possible explanation is that the SANAD II Case Report Forms are completed and submitted in the days following the date of participant assessment as a result of the logistics of the assessment being completed in a busy outpatient clinic.

Figure 7.18: Date of Baseline Assessment: Bland Altman Plot

Mean	1.67
Upper 95% Confidence Limit of Agreement	10.55
Lower 95% Confidence Limit of Agreement	-7.21



### 7.15.2 Planned Healthcare Attendances: The SANAD II Follow-Up Assessments

In the SANAD II Dataset, 392 follow-up assessments were recorded in 98 participants. Forty two of the SANAD II follow-up assessments with no eligible attendance in routine datasets occurred following the date of 31/12/15. These assessments were excluded from the analysis as a result of lack of equivalent data coverage for this time period in the routine datasets. However, for participants residing in England, the Hospital Episode Statistics Outpatient Dataset extends into 2016 and follow-up assessments identified in the outpatient dataset in 2016 for this group of participants were included. Three hundred and fifty follow-up assessments recorded in the SANAD II dataset were included in this assessment.

In the routinely recorded datasets, 317 follow-up assessments were identified in 97 participants. Dates for 316 follow-up assessments were identified in the outpatient datasets with the remaining one assessment identified in the emergency care dataset. Thirty three SANAD II follow-up assessments in 19 participants had no relevant healthcare attendance within the specified clinical limit of agreement of one month (30 days) in routine datasets. Thirteen participants were residing in England and six in Wales. In one participant no follow-up assessments were identified in routine datasets, although a baseline assessment was identified. In the remaining 18 participants at least one of their SANAD II follow-up assessments could not be identified in routine datasets. The missing data in these cases may be explained by the follow-up assessments occurring out of the context of routine medical care, for example assessments may have taken place in dedicated research clinics that may not be included in routine datasets.

The difference in days between the dates of follow-up assessments from SANAD II subtracted from the dates from routine datasets was calculated and the assumption of normal distribution around the mean assessed. A spike-at-zero distribution is observed, detailed in *Figure 7.19*.

A Wilcoxon Signed Rank Test has been performed in SPSS. The non-significant result ( $P=0.14$ ) indicates that the mean dates of follow-up assessments calculated from SANAD II and routine datasets are not significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 7.11* and *Figure 7.20*.

The Bland Altman Plot demonstrates that there is acceptable agreement between the calculated dates. The 95% confidence limits of agreement between the dates of baseline assessment are -4 and 4 days, within the specified acceptable clinical limit of agreement of one month (30 days), indicated by the red dashed line. The mean of the difference between the dates is 0.07, indicating that there is essentially no difference in the date of completion of follow-up assessments.

Figure 7.19: The Difference in Days Between the Date of Follow-Up Assessments

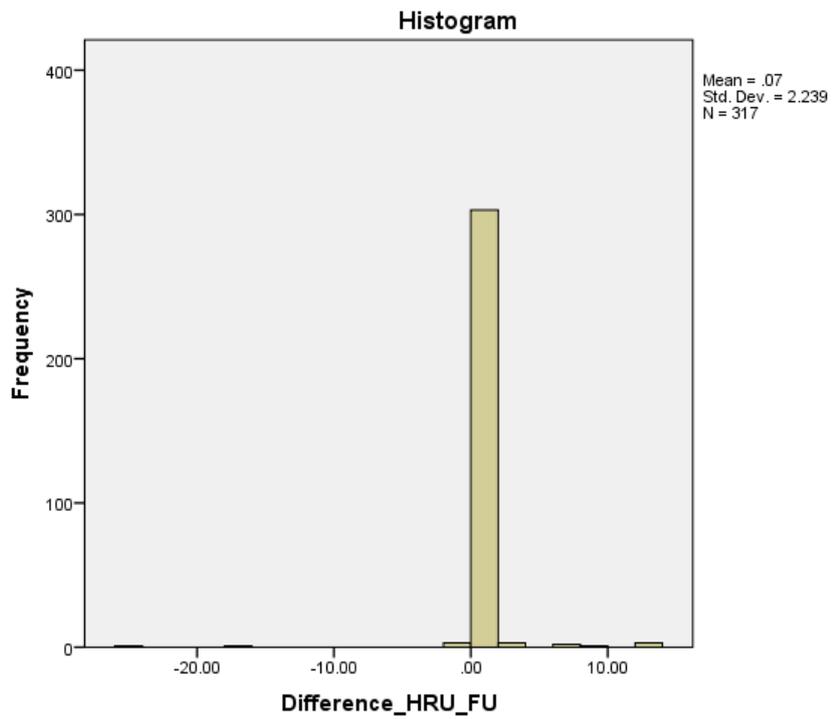


Figure 7.20: Date of Follow-Up Assessments: Bland Altman Plot

Mean	0.07
Upper 95% Confidence Limit of Agreement	4.47
Lower 95% Confidence Limit of Agreement	-4.33

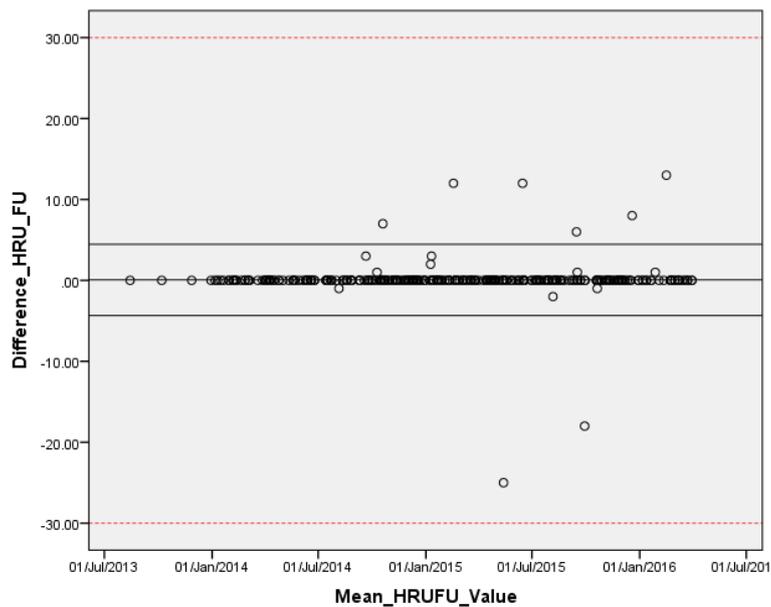


Table 7.11: Planned Healthcare Resource Use: Descriptive Statistics and Agreement

		Baseline Assessment	Follow-Up Assessments	
RCT Data	SANAD II Assessments	98 (100%)	350 (100%)	
	Total Eligible Patients	98	98	
Routine Data	SANAD II Assessments	87 (88.8%)	317 (90.6%)	
	Total Patients	87	97	
Total Paired Assessments		87 (88.8%)	317 (90.6%)	
Routine Datasets	Total SANAD II Assessments in Dataset	HES: Admitted Patient Care	2	0
		HES: Accident and Emergency	1	1
		HES: Outpatient	63	240
		HES: Adult Critical Care	0	0
		SAIL: Patient Episode Database for Wales	0	0
		SAIL: Emergency	3	0
		SAIL: Outpatient	17	76
		SAIL: Primary Care	0	0
Assessment of Agreement	RCT Data	Mean	25/05/14	10/02/15
		Range	23/05/13 – 19/02/15	02/08/13 – 30/03/16
	Routine Data	Mean	25/05/14	25/02/15
		Range	23/05/13 – 09/02/15	14/08/13 – 30/03/16
		Test for Significance	Wilcoxon Signed Rank	Wilcoxon Signed Rank
	Significance	P<0.001	P=0.14	

### 7.15.3 Conclusions: Planned Healthcare Attendances

Planned healthcare attendances are recorded in routine datasets and data are relatively complete and in agreement with data collected in SANAD II. This result is not surprising as the record of attendance is required for NHS remuneration, the primary purpose for routinely recording the data in the included datasets. There is therefore potential for routinely recorded data to contribute to health economic analyses within prospective research and provide additional data, for example identification of healthcare attendances not forwarded to the SANAD II team.

#### **7.15.4 Unplanned Healthcare Attendances**

In the SANAD II dataset, 94 participants completed 263 self-report questionnaires. Therefore 263 three-month time periods were examined in the routinely recorded datasets to identify unplanned emergency attendances and inpatient admissions.

##### *7.15.4.1 Emergency Department Attendances*

Thirty two participants reported 52 emergency department attendances in the SANAD II dataset. In the routinely recorded datasets, thirty seven emergency attendances were identified in 26 participants. Twelve SANAD II questionnaire responses were in agreement with the number of attendances identified in routine datasets. Sixteen SANAD II responses recorded a greater number of attendances than identified in routine datasets and eight SANAD II responses recorded fewer attendances than routine datasets. There is therefore a trend for participants to self-report a greater total number of emergency attendances than identified in routinely recorded datasets.

##### *7.15.4.2 Inpatient Admissions*

Seven participants reported 28 inpatient admissions in the SANAD II dataset. A single participant reported 16 admissions and this record is likely to be erroneous and has been excluded from the remainder of the analysis leaving six participants reporting 12 admissions. In the routinely recorded datasets, nineteen inpatient admissions were identified in 18 participants. Three SANAD II questionnaire responses were in agreement with the number of attendances identified in routine datasets. Three SANAD II responses recorded a greater number of attendances than identified in routine datasets and fourteen SANAD II responses recorded fewer attendances than routine datasets. There is therefore a trend for participants to self-report a fewer total number of inpatient attendances than identified in routinely recorded datasets. 'Admission' includes transfer to any ward following emergency department attendance, including transfer for brief periods, for example same-day discharge on medical and surgical assessment units. Limited participant understanding of the definition of 'admission' may partly explain the results.

*Table 7.12* presents the unplanned healthcare attendances for each participant self-reported in SANAD II and identified in routine datasets.

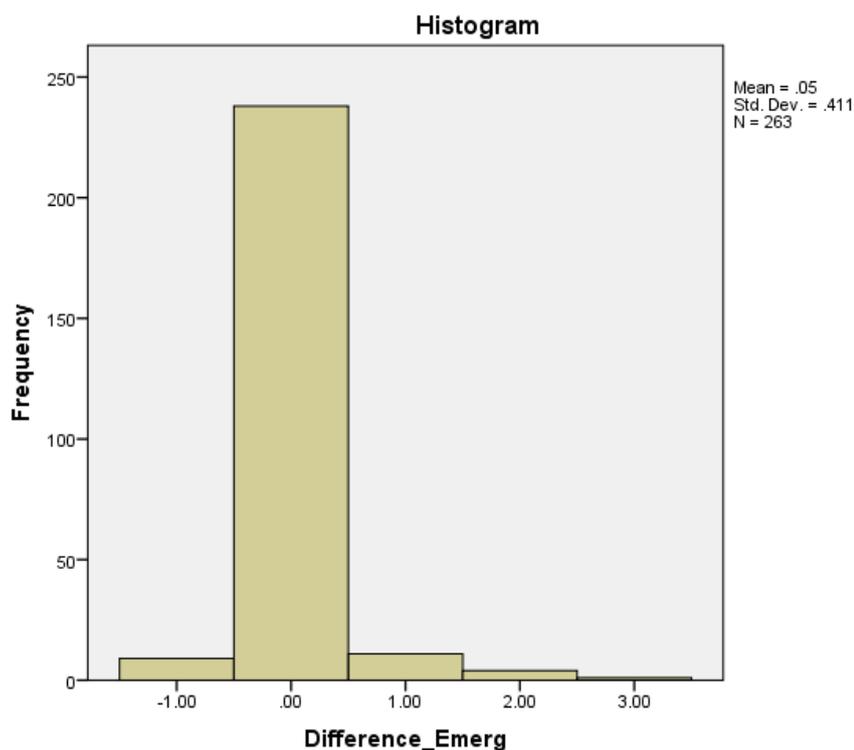
Table 7.12: Identified Unplanned Healthcare Attendances

Participant	SANAD II Emergency Attendances	Routine Emergency Attendances	Participant	SANAD II Inpatient Admissions	Routine Inpatient Admissions
012	1	0	013	3	1
013	1	1	013	4	0
013	1	0	059	1	0
043	1	1	059	1	1
048	3	1	072	0	1
059	1	1	084	0	1
059	1	1	097	0	1
059	1	1	111	1	1
062	1	0	149	0	1
072	3	1	170	0	1
097	1	0	174	0	1
111	1	0	174	0	1
111	1	1	212	0	1
134	2	1	231	0	1
137	2	2	286	0	1
137	3	1	289	0	1
149	1	1	414	16	1
170	3	0	478	0	1
212	1	1	229	2	2
227	0	1	231	0	1
231	8	7			
286	1	0	<b>Total</b>	<b>28</b>	<b>19</b>
311	1	1			
314	1	1			
351	0	1			
383	1	2			
383	1	2			
383	1	2			
384	1	1			
384	0	1			
414	1	2			
478	4	2			
227	1	0			
231	1	0			
460	1	0			
478	0	1			
<b>Total</b>	<b>52</b>	<b>37</b>			

#### 7.15.4.3 Emergency Department Attendances: Agreement

The difference between the number of emergency department attendances recorded in SANAD II subtracted from the number recorded in routine datasets was calculated and the assumption of normal distribution around the mean assessed. A spike-at-zero distribution is observed, detailed in *Figure 7.21*.

*Figure 7.21: The Difference in Number of Emergency Department Attendances*

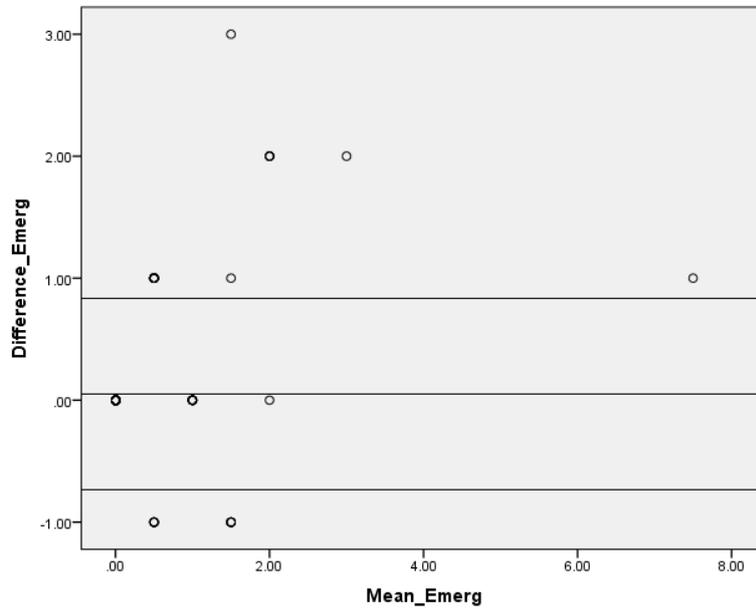


A Wilcoxon Signed Rank Test has been performed in SPSS. The non-significant result ( $P=0.051$ ) indicates that the mean number of emergency department attendances reported in SANAD II and identified from routine datasets are not significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 7.13* and *Figure 7.22*.

The Bland Altman Plot demonstrates that there is acceptable agreement. The 95% confidence limits of agreement between the numbers of emergency department attendances are -0.7 and 0.8. The mean of the difference is 0.05, indicating that there is essentially no difference between the number of emergency department attendances.

Figure 7.22: Number of Emergency Department Attendances: Bland Altman Plot

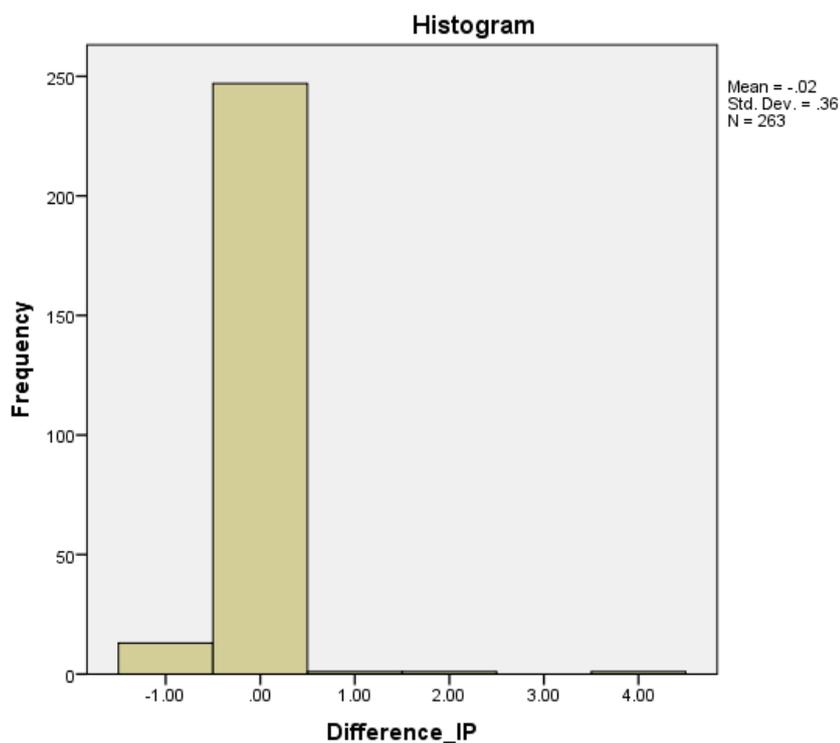
Mean	0.05
Upper 95% Confidence Limit of Agreement	0.834
Lower 95% Confidence Limit of Agreement	-0.734



#### 7.15.4.4 Inpatient Admissions: Agreement

The difference between the number of inpatient admissions recorded in SANAD II subtracted from the number recorded in routine datasets was calculated and the assumption of normal distribution around the mean assessed. A spike-at-zero distribution is observed, detailed in *Figure 7.23*.

*Figure 7.23: The Difference in Number of Inpatient Admissions*

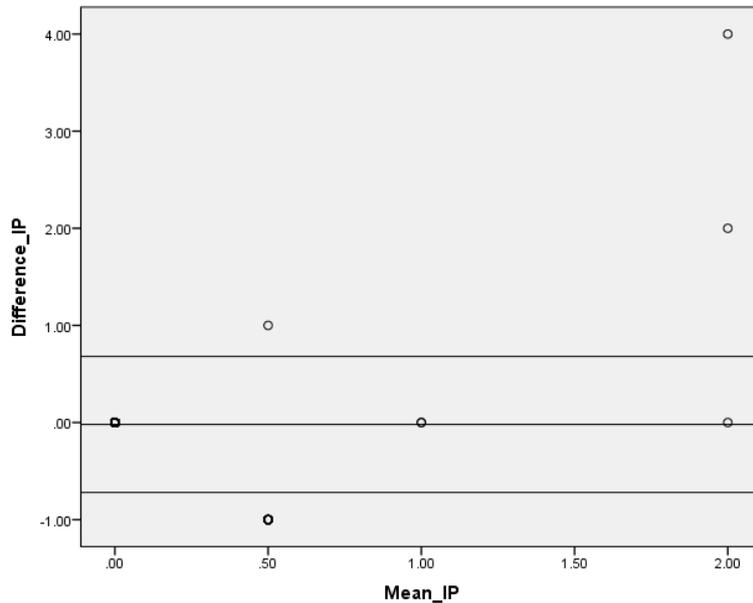


A Wilcoxon Signed Rank Test has been performed in SPSS. The non-significant result ( $P=0.098$ ) indicates that the mean number of inpatient admissions reported in SANAD II and identified from routine datasets are not significantly different. Subsequently, agreement was assessed through the construction of a Bland Altman Plot. Results are presented in *Table 7.13* and *Figure 7.24*.

The Bland Altman Plot demonstrates that there is acceptable agreement. The 95% confidence limits of agreement between the numbers of inpatient admissions are -0.7 and 0.7. The mean of the difference is -0.02, indicating that there is essentially no difference between the number of inpatient admissions.

Figure 7.24: Number of Inpatient Admissions: Bland Altman Plot

Mean	-0.02
Upper 95% Confidence Limit of Agreement	0.68
Lower 95% Confidence Limit of Agreement	-0.72



### 7.15.5 Conclusions: Unplanned Healthcare Attendances

In the majority of the 263 comparison three-month periods, participants reported no unplanned attendances in SANAD II and no attendances were identified in routine datasets. However, in the small number of cases where attendances were reported, participants in SANAD II self-reported a greater number of emergency attendances and fewer inpatient admissions, compared to routinely recorded datasets. For emergency attendances, 23% of SANAD II responses were in agreement with routine data and for inpatient admissions, 10%. Despite these differences, there was overall acceptable agreement, although this was as a result of the large number of responses indicating zero unplanned attendances. Routinely recorded data demonstrate potential to inform both routine clinical practice and research. The differences in reporting require further assessment and it is likely the greatest utility of routinely recorded data in research would be in complementing data collected using standard prospective methods, in line with the current status quo for the collection of health economic data in RCTs.

Potential explanatory factors for the differences identified may include recall bias and participant misunderstanding of the constitution of 'inpatient admission'.

Table 7.13: Unplanned Healthcare Resource Use: Descriptive Statistics and Agreement

			<b>Emergency Attendances</b>	<b>Inpatient Admissions</b>
<b>RCT Data</b>		Total Eligible Patients	94	94
		Total Self-Report Responses	263 (100%)	263 (100%)
		Total Healthcare Use	52 (19.8%)	12 (4.6%)
<b>Routine Data</b>		Total Eligible Patients	94	94
		Total Healthcare Use	37 (14.1%)	19 (7.2%)
		Total Paired	263 (100%)	263 (100%)
<b>Assessment of Agreement</b>	<b>RCT Data</b>	Mean	0.20	0.05
		Range	0 - 8	0 - 4
	<b>Routine Data</b>	Mean	0.15	0.07
		Range	0 - 7	0 - 2
		Test for Significance	Wilcoxon Signed Rank	Wilcoxon Signed Rank
		Significance	P=0.051	P=0.098

### **7.16 Primary Care Data, North West England**

Primary care data were extracted directly from the General Practices for two participants resident in North West England. The electronic medical records including UK READ Codes, free text entries, investigation results and clinical correspondence were reviewed.

As a result of the sample size a descriptive analysis was performed, summarised in *Table 7.14*. The data retrieved from the primary care electronic medical records were in general, complete. For example, dates of seizures and prescriptions and the occurrence of adverse events were identified. Furthermore, additional data were available such as the result of an EEG that was not included in the SANAD II dataset. However, there is some variation in the dates recorded and this may be explained by variations in the dates recorded when entering information onto the electronic medical record system. For example, for outpatient attendances, dates may be recorded as the date of attendance or the date of receipt of correspondence following the outpatient attendance.

The primary care records were informative for the assessment of prescriptions and unplanned healthcare resource use, for example as a result of the occurrence of adverse events. However, the majority of the relevant data retrieved was extracted from the clinical correspondence recorded on the primary care electronic medical record. Such correspondence included letters following emergency care attendances and Neurology outpatient clinic attendances.

Table 7.14: Primary Care Data, North West England: Descriptive Analysis

Variable	Participant 1	Participant 2
<b>Date of First Seizure</b>	<b>SANAD II:</b> 17/12/13 <b>Primary Care:</b> 17/12/13, A&E letter	<b>SANAD II:</b> 01/11/12 <b>Primary Care:</b> N/A, pre-2013
<b>Diagnosis and Classification of Epilepsy</b>	<b>SANAD II:</b> 31/01/14, Focal <b>Primary Care:</b> 24/01/14, Focal, Neurology OP letter	<b>SANAD II:</b> 13/08/13, Focal <b>Primary Care:</b> 15/08/13, Focal, Neurology OP letter
<b>Clinical Investigations</b>	<b>SANAD II:</b> EEG (26/03/14, abnormal), MRI (30/04/14, abnormal) <b>Primary Care:</b> EEG (15/02/14, abnormal), MRI (08/05/14, abnormal)	<b>SANAD II:</b> MRI (17/07/13, abnormal) <b>Primary Care:</b> EEG (23/09/13, abnormal)
<b>Date of First Follow-Up Seizure</b>	<b>SANAD II:</b> 19/05/14 <b>Primary Care:</b> 29/01/14, A&E letter, attendance following 2 convulsive seizures	<b>SANAD II:</b> 23/10/14 <b>Primary Care:</b> 24/10/14, GP attendance following seizure
<b>Date of Last Follow-Up Seizure</b>	<b>SANAD II:</b> 15/10/14 <b>Primary Care:</b> Neurology letter, 'last partial seizure October 2014'	<b>SANAD II:</b> No further seizures <b>Primary Care:</b> No further seizures
<b>Date 12 Month Remission Achieved</b>	<b>SANAD II:</b> 15/10/15 <b>Primary Care:</b> October 2015, Neurology letter, 'last partial seizure October 2014'	<b>SANAD II:</b> Insufficient follow-up <b>Primary Care:</b> Insufficient follow-up
<b>Total Number of Follow-Up Seizures</b>	<b>SANAD II:</b> 2 <b>Primary Care:</b> 4	<b>SANAD II:</b> 1 <b>Primary Care:</b> 3 (within 24 hours)
<b>Date of First Prescription of Antiepileptic Drugs</b>	<b>SANAD II:</b> 31/01/14 <b>Primary Care:</b> 31/01/14	<b>SANAD II:</b> 13/08/13 <b>Primary Care:</b> 15/08/13
<b>Dates and Nature of Adverse Events</b>	<b>SANAD II:</b> 28/01/14 (Tiredness) <b>Primary Care:</b> 25/04/14 (Tiredness) Neurology OP letter	<b>SANAD II:</b> 27/08/13 (Memory impairment), 13/09/13 (Reduced appetite, weight loss), 21/01/14 (Rash), 15/04/14 (Poor sleep), 15/06/14 (Hair loss), 15/09/14 (Anger, aggression) <b>Primary Care:</b> 04/12/13 (Unpleasant taste, weight loss, memory impairment) Neurology OP letter, 11/03/14 (Rash) Neurology OP letter, 09/10/14 (Anger, hair loss, poor sleep) Neurology OP letter, 30/10/14 (Blistered tongue) GP attendance
<b>Dates of SANAD II Assessments</b>	<b>SANAD II:</b> 31/01/14, 25/04/2014, 23/05/2014, 01/08/2014, 25/03/2015, 17/09/2015, 23/03/2016 <b>Primary Care:</b> 24/01/14, 25/04/14, 23/05/14, 01/08/14, 25/03/15, 16/09/15, 22/03/16	<b>SANAD II:</b> 13/08/13, 26/11/2013, 03/03/2014, 10/11/2014, 19/12/2014, 18/08/2015 <b>Primary Care:</b> 15/08/13, 18/11/13, 11/03/14, 10/11/14, 19/12/14, 20/08/15
<b>Dates of Unplanned Healthcare Attendance</b>	<b>SANAD II:</b> 02/08/14 ('In the last 3 months' 1 A&E attendance) <b>Primary Care:</b> 1 A&E attendance, A&E letter	<b>SANAD II:</b> N/A <b>Primary Care:</b> 3 GP attendances as a result of adverse events, 1 GP attendance as a result of seizure occurrence

### **7.17 Conclusions**

In this chapter, routinely recorded data have been compared to data collected using standard prospective methods for variables and outcome measures relevant to the follow-up of participants in SANAD II.

Follow-up seizures were poorly recorded in the routinely recorded datasets, with missing data and poor agreement compared to data collected using standard methods. Resultantly, a significantly greater proportion of participants achieved the outcome measure 'time to 12 month remission' calculated using routine data, with significantly fewer total follow-up seizures recorded. Routinely recorded data regarding AEDs were complete and had good agreement compared to data collected using standard methods. Additionally, data regarding compliance could be inferred and further data regarding prescriptions were available, including prescribed AEDs not recorded in SANAD II. However, adverse events were not recorded, either through specific codes indicating 'adverse events' or through healthcare attendances correlating with the dates of adverse events recorded in SANAD II. Episodes of planned healthcare resource use could be identified, although participants in SANAD II self-reported a greater number of unplanned emergency attendances and fewer unplanned inpatient admissions, compared to routinely recorded datasets.

The poor quality and agreement of routinely recorded data compared to data collected using standard methods has implications for the utility of routinely recorded data. In epilepsy research, routinely recorded data may be limited for the measurement of prospective seizure outcomes, adverse events and episodes of unplanned healthcare attendances. In clinical practice, routinely recorded data are not suitable for monitoring treatment effectiveness or disease monitoring including incidence and prevalence rates.

Explanations for the poor record of seizures include inaccurate recording of codes in routinely recorded datasets, inaccurate clinical diagnosis of seizures or participants not seeking medical attention, which may also explain the poor record of adverse events. The poor record of unplanned healthcare resource use is likely explained by recall bias and limited participant understanding of the context of 'admission'. As above, the greatest potential of routinely recorded data may be in providing collateral clinical information. For example, a first follow-up seizure was identified in three participants in the routinely recorded datasets, but not in the SANAD II datasets.

Considering the results thus far, the accuracy and reliability of such 'seizure occurrences' must be questioned, however, knowledge of such data would direct further assessment within the trial, either through source data verification or clarification with the individual participant.

This assessment has notable limitations, previously discussed in Chapter Six, 6.17.

In the following Chapter Eight, the feasibility and efficiency of accessing and using data routinely recorded for participants in SANAD II is discussed and recommendations for improvement proposed.



## **Chapter Eight**

### **Results: The Feasibility and Efficiency of Accessing and Using Routinely Recorded Data**

#### **8.1 Introduction**

In Chapter Four, routinely recorded data sources ‘accessible’ for this study were identified. In Chapters Six and Seven the attributes of routinely recorded data compared to data collected using standard prospective methods in SANAD II were assessed. In this chapter, the feasibility and efficiency of accessing and using data routinely recorded for participants in SANAD II is discussed and recommendations for improvement proposed.

#### **8.2 Accessing and Using Routinely Recorded Data**

The assessment of ‘accessibility’ considered in Chapter Four is an important component of this assessment of feasibility and efficiency of accessing and using routinely recorded data. In addition, the resources, including financial and time required for applications, the outcome of applications, resources required for data preparation and the attributes of the data provided are factors also included in this assessment, as detailed in Chapter Five, Methods. *Table D.1, Appendix D* summarises the key events and timeline in the scoping and application processes. This assessment, together with the review of accessibility presented in Chapter Four has been published and is also included in *Appendix D* [186].

### **8.2.1 'Accessible' Routinely Recorded Data Sources**

Routinely recorded secondary care clinical data sources were accessible in this study. Data for participants resident in England and Wales were requested from NHS Digital and The Secure Anonymised Information Linkage Databank (SAIL) respectively. Data from NHS Services Scotland; Information Services Division (ISD) were accessible, but there was only limited number of eligible individuals resident in Scotland and providing consent and an application in this study would not have been worthwhile. Routinely recorded primary care clinical data was accessible from North West eHealth (NWEH) and participants General Practitioners directly, for individuals resident in the North West of England. Finally, individual-level mortality data from ONS were accessible but for this study requiring participant consent and assessing a retrospective period, were not required.

#### *8.2.1.1 NHS Digital [55]*

NHS Digital were first requested to review the Participant Information Sheet (PIS) and Consent Form in August 2015. Following two further requests, a discussion in person with a member of the Data Access Request Service (DARS) resulted in a full application being submitted in November 2016. Feedback from the DARS team was received in December 2016. Following this, feedback was also received from the Data Access and Information Sharing (DAIS) Team, responding to the initial requests. Aware of forthcoming alterations to the application and approval process, NHS Digital were contacted to advise on the method of full application. A full application was submitted in February 2016, using the existing system as advised. Multiple contacts were then attempted and reassurance received that the application had been accepted. The first formal acknowledgement of the application in April 2016 advised that all applications must now be submitted via the newly introduced DARS Online Portal. The application was re-submitted. In May, the pre-Data Access Advisory Group (DAAG) and DAAG meetings recommended the application should be approved once minor amendments had been addressed. The Data Sharing Agreement was signed in June 2016 and Hospital Episode Statistics (HES) made available for download on 13<sup>th</sup> July 2016. The total cost of the data, based on a cost-recovery system was £10,200 including VAT.

### *8.2.1.2 The Secure Anonymised Information Linkage Databank (SAIL) [100]*

SAIL were first contacted in April 2015 during the assessment of accessibility. SAIL confirmed they were affiliated with The Administrative Data Research Network (ADRN) and a single application to the ADRN could include access to SAIL datasets. However, subsequently data access through the ADRN was not possible. In June 2015, information regarding the application process was requested and a teleconference arranged. The SAIL Scoping Document was completed in July 2015. In August 2015, review of the PIS and Consent Form was requested and promptly received. In January 2016, further review was requested, following alterations at the request of the other organisations including NHS Digital. The full application was submitted in February 2016. Minor amendments were requested in March 2016 and the application was submitted to the Information Governance Review Panel (IGRP) in April 2016. The IGRP approved the application in July 2016 and data were available for download in August 2016. The total cost of the data, based on a cost-recovery system was £3390 including VAT.

### *8.2.1.3 North West eHealth (NWEH) [97]*

NWEH were first contacted in October 2015 to discuss the study outline and confirm feasibility in principle. In November 2015 the study protocol, PIS and Consent Form were reviewed. Methodological details were discussed in person, including the requirements for data access. Discussions with both NWEH and Apollo determined the specific requirements. General Practitioners already consented and involved with NWEH for the Salford Lung Study would be included. Such GP's had the Apollo data extraction software already installed. The existing data extraction query would be used within the Apollo software. The total cost of the data, based on a cost-recovery system was £16,800, funding both NWEH and Apollo. This cost did not include third party enrolment of the GP's as it was decided the researcher would be responsible for this as a cost saving measure. In May 2016, following recruitment of the study sample, less than five individuals were registered in GP's enrolled in the Salford Lung Study. Therefore, accessing primary care data through NWEH was not worth the cost.

#### *8.2.1.4 General Practices', North West England*

As a result of the prohibitively expensive costs for access to primary care data through NWEH, an amendment was approved by the Research Ethics Committee and Health Research Authority for the researcher to approach the registered GP's for participants in this study resident in the North West of England. The researcher requested to attend the GP on one occasion and transcribe the relevant data onto a de-identified Data Extraction Form. As detailed in the Chapter Five, Methods, a Letter of Access from the National Institute of Health Research Clinical Research Network was required and an invitation pack mailed to the GP's of 18 participants providing consent. Two GP's provided consent and data was extracted for the participants in accordance with the protocol in July 2016. Three GP's refused participation and the remaining 13 GP's provided no response despite repeated attempts at contact via phone call and email.

#### **8.2.2 'Non-Accessible' Routinely Recorded Data Sources**

The majority of routinely recorded primary care clinical data sources were not accessible in this study. Additionally, The Driver and Vehicle Licensing Agency, HM Revenue and Customs and The Department for Work and Pensions were not accessible, either directly or through the Administrative Data Research Network.

##### *8.2.2.1 Routinely Recorded Primary Care Clinical Data Sources*

The Clinical Practice Research Datalink (CPRD) [90] was first contacted in November 2014 and the feasibility of the study was broadly confirmed. In August 2015, CPRD were re-contacted and reported that the CPRD Confidentiality Advisory Group, ethical and governance approvals needed to be updated to permit identifiable, linked data release. At the time of this scoping procedure, this process was in development; NHS Digital are also CPRD's Trusted Third Party. However, the timelines to resolve these barriers were unclear resulting in CPRD not being accessible for this study. Furthermore, an estimated quote was received and was expensive; the total cost of the data based on a cost recovery system was £17,000.

The Health Improvement Network (THIN) [95], ResearchOne [91] and QResearch [93] were contacted in September 2015.

All organisations operate on a de-identified basis, with no facility to re-identify individuals. Such data sources were therefore not accessible for this study.

#### *8.2.2.2 The Driver and Vehicle Licensing Agency (DVLA) [18]*

The DVLA was initially contacted in October 2014. Despite multiple phone calls and emails regarding the broad feasibility of accessing DVLA data for this study, no response was received. In February 2015, following discussion with a member of a DVLA expert committee, a DVLA medical advisor was contacted. The study was discussed with the DVLA Data Sharing Team and the response indicated that the DVLA would not have the capacity to assist with the study and the DVLA data security requirements are 'over and above those in the NHS or University'.

#### *8.2.2.3 Department for Work and Pensions (DWP) and HM Revenue and Customs (HMRC) [103, 104]*

The DWP and HMRC were first contacted in November 2014. No response was received from the HMRC; however, the DWP directed the enquiry to the External Data Sharing and Advice Centre. In December 2014, the External Data Sharing Advice Centre responded and advised that data access directly with the DWP or HMRC would not be possible and the enquiry should be discussed with the Administrative Data Research Network.

#### *8.2.2.4 The Administrative Data Research Network (ADRN) [102]*

The ADRN were first contacted in December 2014 to discuss feasibility for this study. Despite multiple attempts at contact, no response was received until February 2015. General guidance was provided via email. In March 2015 a teleconference was arranged to discuss the study. ADRN confirmed that the study was eligible for their services and they can request access to DWP and HMRC data linked to routinely recorded clinical data such as HES provided by NHS Digital. ADRN agreed to begin contacting the relevant data sources to request access to data. In April 2015 a further teleconference revealed no significant progress. In May 2015 HMRC declined participation although no clear reasoning was provided. The discussions with DWP were on-going. Furthermore, ADRN reported that if the DWP do not permit access to their data, the study would not be eligible for application through the ADRN solely for routinely recorded clinical data and independent applications must be submitted to the relevant organisations. In July 2015 the ADRN informed the researcher via email that the DWP have not been forthcoming but negotiations were on-going. No further feedback was received.

## **8.3 The Feasibility and Efficiency of Accessing and Using Routinely Recorded Data**

### ***8.3.1 Clinical Routine Data Sources***

#### ***8.3.1.1 Secondary Care Data***

Routinely recorded secondary care data were successfully requested on an individual-level, identifiable basis for participants resident in England and Wales through NHS Digital and SAIL. However, although access to data was achieved there were notable limitations.

In England, NHS Digital has a target time to data access of 60 working days following submission for complex applications including bespoke data linkage from multiple datasets. From the date of submission of the application via the DARS Online Portal access to HES data was granted within this timeframe. However, this positive experience following submission of the application is countered by limitations in the pre-application period. Acknowledging the significant update to the online application and approval procedures that occurred during the period of application, there remained a considerable period of time required in the development of the application. The nature of the request for identifiable data necessitated participant consent as the valid legal basis. NHS Digital require ethical and governance approvals to be in place prior to DAAG review and to prevent future amendments and delays, it was rational to ensure the participant consent materials had been reviewed by the NHS Digital Information Governance Team, prior to submitting the documents for ethical and governance approval. NHS Digital publish written guidance regarding consent and advise that documents for individual projects should be reviewed. However, in our experience there is no formalised process for providing this review. Following significant correspondence the consent materials were reviewed by the Data Access and Information Sharing Team. However, this feedback was provided following a full application submission and review by the Data Access Request Service. This process was inefficient for both NHS Digital and the researcher. Furthermore, at £10,200 the data were expensive despite being calculated using a cost-recovery system.

For participants in Wales, data were retrieved through SAIL Databank. SAIL provided a streamlined pre-application service, including promptly engaging in multiple discussions and completion of a scoping document, outlining the study methods and costs involved at an early stage. Consent materials were also promptly reviewed by a member of the Information Governance Team with each request.

Although the pre-application time period was similar to NHS Digital, it was NHS Digital who were primarily responsible for the prolonged duration. In contrast to NHS Digital however, data were provided by SAIL eight months following submission of the application, compared to three months for NHS Digital. The cost of SAIL data was £3390, noticeably less than the £10,200 required by NHS Digital despite both organisations using a cost-recovery system.

Data retrieved from both NHS Digital and SAIL required a period of 'data cleaning'. This involved reviewing each individual data item for each participant and extracting relevant data according to the developed algorithms – the total included number of ninety eight participants permitting this process, requiring four weeks to complete. The data from both sources were in similar formats although there were slight differences with naming conventions and the availability of specific variables. Data dictionaries were available from both sources, although SAIL supplied data dictionaries together with the data. Finally, as presented in Chapters Six and Seven, the quality and agreement of the routinely recorded data are poor compared to data collected using standard prospective methods.

There are stringent Information Governance requirements that must be in place prior to application for both NHS Digital and SAIL. These include Information Security assessments, specific inclusion regarding the 'processing of healthcare data for the subjects of research' in the institutional Data Protection Act registration and in the case of NHS Digital, an institutional Data Sharing Framework Contract. Adequate guidance is provided by both NHS Digital and SAIL and if not addressed early in the pre-application process may cause delay, although such delay was avoided in this study.

#### *8.3.1.2 Primary Care Data*

Routinely recorded primary care data were successfully requested on an individual-level, identifiable basis for participants resident in North West England, although is noticeably less accessible than secondary care data.

The majority of primary care routinely recorded data sources including The Health Improvement Network, ResearchOne and QResearch are collaborations between academic or private institutions and the developers of the Information Technology systems used in General Practices. Each data source records data for GP's who have provided consent and who have the relevant software installed.

This introduces an initial limitation as individuals in a RCT may be registered to different GP's using a number of different software providers and primary care data for this group of individuals would therefore be recorded in a number of different routine data sources. Furthermore, such sources record data on a de-identified basis with no facility to re-identify individuals. Therefore, where specific individuals need to be identified as for this study, these sources are not accessible.

The Clinical Practice Research Datalink is a similar data source recording data on a de-identified basis and as a result of the required ethical and governance approvals not being in place, individuals could not be re-identified and data were not accessible for this study. Furthermore, the £17,000 quote provided for access to data was prohibitively expensive, when only a small proportion of the participants in SANAD II are likely to be registered in CPRD enrolled GP's.

NorthWest eHealth has an established infrastructure for accessing linked primary and secondary care data and prescribing data. NWEH do not routinely provide a bespoke primary care data extraction service for research, however following a period of discussion agreed to participate in this study. In order to ensure the most time and cost-effective methodology, GP's already enrolled with NWEH would be included and the existing data extraction query used. Additionally, the usual process of using a third party to visit, recruit and consent the GP's would be replaced with the researcher performing these duties. Despite these measures aiming to maximise cost efficiency, the quote included £11,027 for 'data handling', £1575 for 'data check', £1326 for Project Management and £7200 for Apollo data query testing, totalling £21,128. If recruitment was included in addition using *CK Aspire*, the company contracted by NWEH for the Salford Lung Study an additional £6800 would be charged. Following recruitment of the participants in this study, just three were registered to GP's enrolled with NWEH, making the cost per participant of £7042 prohibitively expensive.

Routinely recorded primary care data were extracted directly from participants' electronic medical records at their registered GP. With the exception of research time, there was no additional cost and for the two participants where data were available, the attributes were encouraging as discussed in Chapter Seven. However, the time intensive approach for data extraction and preparation limited the feasibility and efficiency. Perhaps of greater importance, despite participant consent only two of 18 GP's consented.

This may indicate a systematic limitation with this approach and explanatory factors may include the current demands placed on GP's together with the lack of financial incentive to participate in this research.

### ***8.3.2 Non-Clinical Routine Data Sources***

Routinely recorded non-clinical data on an individual-level were not accessible in this study. Following the literature review in Chapter Two and systematic review in Chapter Three, this result was expected.

The Department of Work and Pensions (DWP) and HM Revenue and Customs (HMRC) directed the request for data to The Administrative Data Research Network (ADRN). The ADRN have been unable to negotiate access to data during this study, without reasoning being provided. Clearly accessing and using data from HMRC and DWP during prospective research is not feasible, despite the potential benefits in informing health economic analyses.

The Driver and Vehicle Licensing Agency (DVLA) declined the request for data access, citing insufficient internal resources to process the request and more stringent data protection requirements than those employed in the NHS or academic institutions. However, explicit details regarding these requirements were not provided. Again, accessing and using DVLA data during prospective research is not feasible, despite the potential benefits for assessing clinical outcomes.

## **8.4 Discussion**

Routinely recorded data have potential use in prospective research including measuring the outcomes of RCTs [64] and conducting pragmatic RCTs including the stages of recruitment, randomisation, administration of interventions and follow-up assessments [67]. Academic, political [70] and health service [71] interest in UK sources of routinely recorded data have resulted in expansion and improvements, notably in the access to linked datasets. However the experience in this study with accessing individual-level data for specific participants providing written consent highlights persisting limitations.

The access and use of routinely recorded data for individuals enrolled in SANAD II was assessed. Including the scoping assessments of accessibility, protocol development, research ethics and governance approvals and submission of the applications for data access, a period of 18 months was required before data was provided to the researcher. Notably, of the total 11 identified routinely recorded data sources, excluding those not suitable such as the ONS and NHS ISD for individuals resident in Scotland for which there were insufficient participants, data was successfully retrieved from just three sources of routinely recorded clinical data. The data retrieved covered a retrospective period with just an 18 month time period common to all routine datasets. At the time of receipt of the data (August 2016), the most recently available data common to all datasets was recorded eight months previously (31/12/15). This delay in data availability potentially limits the utility of such sources in prospective clinical research, such as drug trials where prompt reporting is clinically important and a regulatory requirement. Ninety eight participants were included and broadly the attributes of routinely recorded data compared to data collected using standard prospective methods was poor, with missing data and poor agreement the frequent results. The total cost for the data required by NHS Digital and SAIL was £13,590. However, the total underlying financial cost would also include researcher salary. Furthermore, the 18 months of full-time equivalent time required to access the data indicates a significant human resource requirement.

Routinely recorded clinical data sources are numerous and there was comprehensive national coverage of emergency, inpatient, outpatient and critical care, under the umbrella of secondary care. The experience in this study suggests that accessing individual level data is possible but the feasibility is limited.

There were notable inefficiencies in the application process, particularly during the pre-application phase, for example requesting feedback on the Patient Information Sheet and Consent Form prior to Research Ethics and Governance review. The resulting inefficiencies and duplication of tasks negatively impacting both the data holders and the researcher.

Access to routinely recorded individual - level primary care data has not been feasible in this study. The majority of primary care data sources have limited geographical coverage based on the Information Technology software installed and usually record data on a de-identified basis with no facility to re-identify individuals. Data is accessible by approaching GP's directly using an established infrastructure developed by NWEH, but is prohibitively expensive. There was poor engagement with the study when GP's were approached directly, resulting in a poor response rate. The inception of the NHS Digital *General Practice Extraction Service* which routinely records primary care data nationally for England, represents the most optimistic national source, however access is currently restricted to Department of Health initiatives, such as research involving screening procedures [139].

Access to non-clinical data has not been possible. The ADRN has been established to act on behalf of the researcher in negotiating access to de-identified, linked routinely recorded data from a number of organisations and the study proposal was promptly re-directed to the ADRN. However, the decision whether to release data remains with the data holder. This process is therefore inefficient; an application must be submitted and approved by the ADRN who then subsequently approach each organisation individually. Ideologically, the next step would be the storage of de-identified linked data from participating organisations in a single repository, similar to those established for RCT data [201]. This would create a single point of access and remove the burden for each organisation to consider each study individually. This would however require significant information governance and security barriers to be cleared and in light of recent developments within the research climate, individual consent. Including the public as stakeholders in the development of such a data repository would be essential [82].

Although there are examples of pragmatic RCTs being coordinated through routine data sources [67], in this study the process of accessing and using routinely recorded data for participants of SANAD II was not feasible. Perhaps the most important factor in this assessment is that of the attributes of the routinely recorded data.

The degree of missing data and results of poor agreement compared to data collected using standard prospective methods in SANAD II results in an inability to recommend the sole use of routinely recorded data in an RCT such as SANAD II. Consideration of the differences in cost and resource-use required to obtain routinely recorded data compared to data collected using standard prospective methods become less relevant. Therefore, if routinely recorded data is not feasible to measure the outcomes of SANAD II and recommendation is made for data in a future RCT to be collected using standard prospective methods, the relevant question then is what is the added value of retrieving routinely recorded data? For example, routinely recorded data may provide information regarding AED compliance or healthcare resource use that is not recorded or incompletely recorded using standard methods. The relative value of developing the application and retrieving such routinely recorded data in addition to data collected using standard prospective methods would then require assessment. Value of Information (VOI) analysis is a method used to assess the return on investment in research and can be defined as the amount an individual would be willing to pay before making a decision [202]. VOI analysis is appropriate for application in this study to determine the 'optimal mix' of data, or the value of information derived from routinely recorded data compared to data collected using standard methods. VOI analysis however, involves calculation of the objective function, Incremental Net Benefit (INB) derived from the Incremental Cost Effectiveness Ratio (ICER). In this study we could not calculate the ICER as a result of the small sample size and the ongoing status of SANAD II – routinely recorded data from all included participants would be required for this assessment. In a future analysis using the complete SANAD II dataset and data retrieved from routine sources, a VOI analysis could be performed to quantitatively define the value of retrieving routinely recorded data in addition to collecting data using standard methods. However, as discussed previously, the degree of missing data and poor agreement allows prediction that the relative value of additionally retrieving routinely recorded data will be limited.

## **8.5 Conclusions**

The major challenge in accessing and using routinely recorded data for a purpose such as this study with clear secondary benefit to the public and health services seems inappropriate when the 'public purse' funds the research, the researcher and many of the organisations recording the data. Perhaps a significant cause or contributor to the current limitations is the controversy and bad publicity following the *Care.Data* initiative in 2014. The proposal to extract primary care records from all individuals in the UK was opposed publicly by a number of groups and for example, resulted in an internal inquiry within the Health and Social Care Information Centre (HSCIC). Data applications were suspended during this period and our current experience may be explained partly by the concurrent revision of the HSCIC application and approval procedures. Indeed HSCIC has also rebranded itself NHS Digital. However, in the medium term, of more concern is the harm in public perception that has resulted. Currently, more than 1.2 million individuals in the UK have submitted a 'Type 2 objection', meaning that their data will not be shared for purposes other than direct care [203]. Although the application procedures may improve and in time we may be able to access data more efficiently, the loss of 2.2% of the population will have implications for the routinely recorded data that will then be made available for research. Involving the public as important stakeholders and re-gaining their trust will be an essential factor in realising the individual and population healthcare benefits of routinely recorded data [83].

## **8.6 Recommendations**

Recommendations are proposed to improve the access and use of routinely recorded data in prospective research, presented in *Table 8.2*.

Table 8.2: Recommendations to Improve the Access and Use of Routinely Recorded Data in Research

<b>General</b>
<p>Routinely recorded data are being used to measure RCT outcomes with the agreement, additional benefits and cost-efficiency of such data compared to data collected through standard RCT methods being unknown.  <b>Further research in epilepsy and alternative clinical settings should be performed to assess the agreement, additional benefits and cost-efficiency of accessing routinely recorded data to measure RCT outcomes compared to data collected through standard RCT methods.</b></p>
<p>The costs required for data access from routine data sources vary widely, although all reportedly operate on a cost recovery, not-for-profit basis.  <b>Costs should be standardised and rationalised between routine data sources.</b></p>
<p>The time lag before data is available in routine data sources represents a significant limitation to the access of routinely recorded data for prospective research, including RCTs.  <b>The infrastructure and procedures should be developed to reduce the time lag seen in routinely recorded data sources.</b></p>
<p>The requirement for linkage between sources of routinely recorded data has been observed and improvements are on-going, for example with the establishment of the ADRN.  <b>A standardised set of identifying variables could be recorded by all (clinical and non-clinical) data sources to improve the accuracy of data linkage, similar to a Core Outcome Set for clinical trials [204].</b></p>
<p>The public mistrust in the sharing and linking of routinely recorded data will hamper future efforts to develop routinely recorded databases, despite the likely benefits to individual patients and the population.  <b>Further research and public engagement should be undertaken to define the issues of most importance to the public and develop strategies to address these.</b></p>
<b>Clinical Routine Data Sources</b>
<p>There are numerous requirements prior to application and criteria to fulfil on submission of an application, yet the guidance and support during development of an application remains limited.  <b>Formalise and improve access to guidance and review of study materials during the 'pre-application stage'.</b></p>
<p>There is national coverage of routinely recorded secondary care data, yet primary care coverage remains patchy, based on geographical area or GP IT system.  <b>Develop the primary care data sources to provide national coverage, either through collaboration of existing sources and data linkage or development of national data sources, such as the General Practice Extraction Service.</b></p>
<b>Non-Clinical Routine Data Sources</b>
<p>Access to non-clinical data sources to inform clinical research was not possible during this study, despite the significant potential to inform Health Technology Assessment and the increasing importance of such assessments in a healthcare system where resources are increasingly limited.  <b>To assist with Health Technology Assessment and particularly the analysis of health economic outcomes, urgent research is required to consider facilitating access to individual-level identifiable data from non-clinical sources. This would include:</b></p> <ol style="list-style-type: none"> <li>a. Research regarding the public perception and acceptability of using their personal economic data for clinical research.</li> <li>b. Internal review within non clinical sources such as the DWP and HMRC to assess the feasibility and limitations of permitting access to data for clinical research.</li> <li>c. Formalisation of the approval processes through the independent party, the ADRN for access to non – clinical administrative data – currently, following internal approval the ADRN then negotiate access to administrative data on a project by project basis.</li> </ol>

## **8.7 General Conclusions**

In this chapter, the feasibility and efficiency of accessing and using data routinely recorded for participants in SANAD II have been discussed and recommendations for improvements proposed. Limitations include the narrative assessment of feasibility and efficiency, although based upon pre-specified criteria. Quantitative data were presented where possible, such as the costs of data from each source. Methods for the assessment of the 'optimal mix' of data from routine sources and data collected using standard prospective methods were discussed, although as a result of the sample size, inability to calculate Incremental Cost Effectiveness Ratios and the on-going status of SANAD II, it was not possible to perform these analyses.

In the following Chapter Nine, the research in this thesis will be discussed including key results, implications, and recommendations for the improved implementation and use of routinely recorded data in both research and clinical practice.



## **Chapter Nine**

# **Discussion and Conclusions: An Assessment of the Use of Routinely Recorded Data in the UK in a Randomised Controlled Trial**

### **9.1 Introduction**

In Chapter One, routinely recorded data in the UK and the potential for use in clinical research were introduced. Epilepsy and the case study RCT SANAD II were subsequently introduced before finally the objectives of this research were presented. In Chapter Two the use of routinely recorded data in RCTs in the UK was reviewed and in Chapter Three the agreement between UK routinely recorded data compared to data collected using standard methods in prospective studies was assessed in a systematic review. In Chapter Four, sources of routinely recorded data in the UK relevant to the outcomes of SANAD II were reviewed and sources where routinely recorded data were accessible for individuals recruited into SANAD II were identified.

In Chapter Five, the methods for the assessment of the attributes of routinely recorded data retrieved from electronic medical records compared to data collected using standard prospective methods in a randomised controlled trial were presented. The assessment of seizure occurrence, diagnosis and classification of epilepsy in routinely recorded datasets were assessed in Chapter Six and variables and outcome measures relevant to the follow-up of participants in SANAD II were assessed in Chapter Seven. The feasibility and efficiency of accessing and using routinely recorded data for participants in SANAD II were assessed in Chapter Eight.

In this chapter, the headline results and conclusions will be discussed. Recommendations for improving the use of routinely recorded data and suggested areas for further research will be proposed.

## **9.2 Discussion**

The potential use of routinely recorded data in prospective research has been recognised and the accuracy of routinely recorded data compared to medical records has been assessed for a multitude of diagnoses, including epilepsy. However, there is minimal evidence of the assessment of agreement between routinely recorded data and the standard methods of data collection employed in prospective research. Acknowledging the rapidly increasing use of routinely recorded data in prospective research including RCTs and the status of the RCT in remaining the standard for approval of novel treatments in healthcare [110] and to inform everyday treatment decisions, the need for an assessment of the feasibility and agreement of routinely recorded data compared to data collected using standard prospective methods is pressing.

This research reviewed the use of routinely recorded data in prospective research and agreement compared to standard prospective data collection methods, reviewed accessible sources of routinely recorded data in the UK and assessed the attributes including agreement of using routinely recorded data compared to data collected using standard prospective methods in a RCT assessing treatments for epilepsy:

### ***9.2.1 The Use of Routinely Recorded Data in the UK to Assess Outcomes in RCTs***

The use of individual-level routinely recorded data from specified data sources in the UK to inform the assessment of outcomes of RCTs was reviewed.

Routinely recorded data have potential advantages when used in prospective research but the overall experience of accessing data for this purpose remains limited. Registry mortality and secondary care routinely recorded data were the most commonly accessed.

Explanations may include the legal requirement regarding death notification and national recording of secondary care data. Primary care routinely recorded data were infrequently accessed but there was evidence for the feasibility of completing cluster RCTs with simple pragmatic interventions. Despite the potential for non-clinical HMRC, DWP and DVLA data to measure outcomes beyond the standard clinical assessments, there was no evidence of use of data for this purpose in a RCT. Furthermore, a data release register could only be identified for HMRC and on review there was no evidence of data use for clinical research of any methodology.

Explanations for the limited use of routinely recorded data in prospective clinical research may include the unclear accuracy of data and agreement to data collected using standard prospective methods or limitations with the feasibility and efficiency of accessing and using routinely recorded data. Furthermore, the discrepancy in this review between the results of the electronic database search and manual review of published data releases was notable, despite the search strategy being developed to be both sensitive and non-specific. This highlights the likely poor indexing of methodological information in electronic databases.

### ***9.2.2 The Agreement of Routinely Recorded Data with Data Collected Using Standard Prospective Methods in UK Studies***

A systematic review was undertaken to assess the agreement between specified routinely recorded data in the UK and data collected using standard prospective methods to measure the outcomes in prospective clinical studies.

Routinely recorded data have a generally poor pattern of agreement for both clinical data and healthcare economic data. In general, the level of agreement identified in this review is not sufficient to recommend use in place of data collected using standard methods in prospective studies including RCTs. 'All-cause mortality' identified from UK Mortality Registers is the exception where acceptable agreement was observed. This is consistent with the finding of more common use in RCTs, identified in the previous objective. However, cause of death, which relies on coded data recorded on the medical death certificate, was found to be in poor agreement in some studies. This is consistent with the general pattern of poor agreement for clinical outcomes from primary and secondary care medical records.

In addition to the heterogeneity of identified studies and limitations stated in Section 9.2.1, this systematic review is limited by the lack of evidence available, particularly the lack of RCTs with outcomes relevant to the treatment of epilepsy.

### ***9.2.3 The Identification and Accessibility of UK Routinely Recorded Data Sources***

Sources of UK routinely recorded data relevant to the outcomes of SANAD II were presented and the accessibility for individuals recruited into SANAD II assessed.

Routinely recorded secondary care data and mortality data were accessible. Primary care data from the majority of sources were not accessible as a result of the de-identified record of the data, resulting in an inability to identify the specific individuals recruited into SANAD II. Non-clinical data sources were not accessible, the ADRN being unsuccessful in negotiating access to HMRC and DWP data and the DVLA citing insufficient resources and stringent security measures.

There are limitations associated with this assessment of accessibility. It is possible that relevant sources were not identified. Furthermore, the focus of the research was on large regional or national sources of routinely recorded data to ensure generalisability of the results, relevant to the UK-wide SANAD II RCT. The accessibility of smaller sources, such as disease specific registers is likely to differ.

### ***9.2.4 The Attributes of Routinely Recorded Data Extracted from Electronic Medical Records Compared Against Data Collected Using Standard Prospective Methods in the RCT SANAD II***

The quality of routinely recorded data and agreement between routinely recorded data and data collected using standard prospective methods was assessed for the 98 included participants of SANAD II.

A first seizure occurrence could be identified in routinely recorded datasets in 23 of the 98 participants. For participants without a first seizure occurrence; approximately one third of participants had no relevant attendances, one third had a 'relevant attendance', defined as an attendance within 48 hours of the date of a definite seizure recorded in SANAD II, but with missing diagnostic information and one third had a relevant attendance within 48 hours but with inadequate or discrepant diagnostic codes not meeting the criteria for seizure occurrence. For the limited number of participants where first seizures were identified, the agreement for the date of occurrence compared to the date collected using standard prospective methods in SANAD II was poor.

Similarly, a 'diagnosis of epilepsy' was present in less than half of the participants using the routinely recorded data and agreement for the date of diagnosis was poor. Furthermore, there was poor agreement for the classification of seizures, explained by the majority of participants being deemed 'unclassified' as a result of the record of codes with inadequate clinical detail. Similarly, follow-up seizures were poorly recorded in the routinely recorded datasets, with missing data and poor agreement compared to data collected using standard methods. Resultantly, a significantly greater proportion of participants achieved the outcome measure 'time to 12 month remission' calculated using routine data, with significantly fewer total follow-up seizures recorded.

Data regarding the occurrence of the clinical investigations MRI, CT and EEG were available only in the emergency and primary care datasets, with reasonably complete data and results available only in the primary care dataset. Data regarding AEDs recorded in the primary care dataset were complete and had good agreement compared to data collected using standard methods. However, adverse events were not recorded, either through specific codes indicating 'adverse events' or through healthcare attendances correlating with the dates of adverse events recorded in SANAD II. Episodes of planned healthcare resource use could be identified, although participants in SANAD II self-reported a greater number of unplanned emergency attendances and fewer unplanned inpatient admissions, compared to routinely recorded datasets.

For a number of the assessed variables, data were recorded in the routinely recorded datasets and not recorded in SANAD II, with potentially significant implications. For example, seizures were identified in the routine datasets for individuals' seizure free in the SANAD II dataset. Investigations such as EEG were identified in the primary care dataset, not recorded in the SANAD II dataset and data regarding AED prescriptions were available, including prescribed AEDs not recorded in the SANAD II dataset. Finally, data regarding compliance could be inferred from the routine datasets, although with a number of notable assumptions.

Explanations for these findings may include inaccurate recording of codes in routinely recorded datasets or inaccurate initial clinical diagnosis of seizures and epilepsy. Furthermore, the events may not have been 'recordable', for example if participants did not seek medical attention following seizure occurrence or the occurrence of adverse events or if relevant codes or detail are not included in the available routinely recorded datasets.

The poor record of unplanned healthcare resource use is likely explained by recall bias and limited participant understanding of the context of 'admission'.

This assessment has notable limitations. The comparator dataset was derived from the SANAD II data available at the time of assessment and a minority of data entries may have been subject to data checking and confirmation. The variables and constructed outcomes derived from the routinely recorded datasets were defined and extracted using algorithms developed for each comparison. However, there is a risk that relevant clinical events may not have been identified, if the diagnostic code recorded is not included in the algorithm. To address this limitation and explore the data further, the routinely recorded data for each participant was examined individually and in detail. For example, diagnostic codes recorded within 48 hours of a seizure recorded in the SANAD II dataset were examined. This process was feasible as a result of the small sample size. This research has assessed 'agreement' and the discussion has largely considered the RCT data as the 'gold standard'. However, for selected outcomes, specifically episodes of healthcare resource use, it is likely that the accuracy of the routinely recorded data, recorded with the primary function of NHS reimbursement, is greater than the RCT data, collected using methods such as self-report questionnaires. There were limitations with regards to the specific variables that could be assessed, informed by the availability of comparable data in both sources. Furthermore, there were limitations in certain variables that were assessed. For example, outcomes such as 'date of first tonic-clonic seizure' or an assessment of the occurrence of serious adverse events would be reasonably expected to result in healthcare attendance or admission and such events would be recorded in routinely recorded datasets. However, a number of the comparisons in this study have limited validity, such as the assessments of 'dates of follow-up seizures' where participants may reasonably not seek healthcare attendance following each seizure episode. This limitation results from the post-hoc and retrospective design of the study and in a subsequent, prospective study, variables should be selected informed by the known content, strengths and limitations of the routinely recorded data.

### ***9.2.5 The Feasibility and Efficiency of Accessing and Using Data from Electronic Medical Records in the Randomised Controlled Trial SANAD II***

The feasibility and efficiency of the access and use of routinely recorded data for individuals enrolled in SANAD II was assessed.

Following the 18 month period of scoping discussions and protocol development, data were retrieved from three of the total 11 identified routinely recorded data sources. The data retrieved covered a retrospective period of 18 months and the most recently available data were recorded eight months previously, limiting the potential utility in prospective clinical research. As discussed, the quality and agreement of routinely recorded data compared to data collected using standard prospective methods was poor and the total cost £13,590, excluding researcher time and remuneration. The feasibility and efficiency of the use of routinely recorded data in prospective clinical research including RCTs are therefore limited.

The descriptive nature of the assessment, potential omission of relevant criteria and subjective interpretation are limitations in this assessment of feasibility and efficiency.

### ***9.2.6 Implications for Clinical Research and Practice***

The literature and systematic reviews highlight the limited experience of using clinical routinely recorded data in RCTs, which is more pronounced for primary care compared to secondary care data. Non-clinical sources of routinely recorded data had no precedent of use in prospective clinical research and in the cases of HMRC and DWP, no use in clinical research of any methodology.

This research identified a persistent issue with missing routinely recorded data compared to data collected using standard prospective methods in the RCT SANAD II. Furthermore, poor agreement was noted for the majority of variables and outcome measures. Finally, the limitations with accessibility together with the poor quality and agreement result in the process of retrieving routinely recorded data during a RCT being of limited feasibility. In addition, factoring in the application time and financial resources required for data access, the feasibility is further limited.

### *9.2.6.1 Implications for Clinical Research*

The results of this research have potential implications for the utility of routinely recorded data in prospective clinical research. In the SANAD II RCT assessing antiepileptic drug treatments for individuals newly diagnosed with epilepsy, routinely recorded data were broadly unsuitable for the identification of eligible individuals for recruitment and measurement of prospective outcomes and adverse events, compared to data collected using standard prospective methods. These results were not unexpected, acknowledging the identified concerns regarding accuracy of routinely recorded data when used for clinical research [47, 72] and the results of the systematic review presented in Chapter Three.

The limited feasibility, quality and agreement suggest that use of routinely recorded data as the primary data source or as a means of validating data collected using standard methods in prospective clinical research, would be limited for both the identification of eligible individuals for trial recruitment and the measurement of the trial outcomes. However, exceptions include data regarding prescribing and aspects of healthcare resource use, in both cases the routinely recorded data likely being of increased accuracy. Notably, the outcome most commonly measured using routinely recorded data in previous published research including RCTs and an exception; with acceptable evidence for agreement in the systematic review completed in Chapter Three, was mortality.

As a result of the nature of this study and the requirement for participant consent, mortality could not be assessed. The results of this research raise concerns regarding the current and future use of routinely recorded data in RCTs. Excluding mortality, which has demonstrated acceptable agreement and healthcare resource use, which is commonly identified using routinely recorded data together with patient assessment and recall, the acceptable inclusion of routinely recorded data in RCTs seems limited. In the systematic review presented in Chapter Three, acceptable agreement between patient recall and medical records data was observed for the record of selected common, chronic medical diagnoses such as diabetes mellitus [180]. This would be expected, although one study did not find acceptable agreement between self-reported stroke and diagnosis of stroke in medical records [178]. In addition to common medical diagnoses, the alternative scenario where acceptable agreement may be expected is in the identification of complex medical events. For example, 'cancer progression' was successfully identified in medical records compared to physician assessment. In such cases, there may be a multitude of 'events'

including admissions, urgent specialist referrals, investigations and treatments that may be included in an algorithm to identify the complex outcome.

However, the 'middle ground' between pragmatic, common diagnoses and complex events may be where the utility of routinely recorded data is most limited. The systematic review, for example in the identification of poor agreement for minor post-operative complications [155] or pressure ulcers [181] and the results of this research, for example in the identification of seizure occurrence, support this conclusion.

Routinely recorded data in the context of prospective clinical research may represent an important source of collateral data in addition to the primary data collected using standard prospective methods, for example to identify additional events such as seizures not recorded using standard methods. Acknowledging the identified limitations, the accuracy and reliability of such data must be questioned. However, this additional information may direct further interrogation within the trial, including source data verification or clarification with the individual participant. Furthermore, routinely recorded data may also be valuable in providing additional data, for example data regarding longer-term follow-up, beyond the normal lifespan of a RCT and following measurement of the primary outcomes [205]. Lastly, informed by knowledge of the content, strengths and limitations, routinely recorded data alone may reasonably be used to measure specific outcomes such as mortality, episodes of healthcare resource use or clinical events resulting in healthcare attendance.

This research also has potential implications for the use of routinely recorded data in retrospective research. Routinely recorded data held in administrative datasets are seen as a valuable resource for retrospective research and their use is established. Using the diagnosis of epilepsy as an example, there are studies worldwide assessing the accuracy of algorithms applied to routinely recorded data to identify individuals with a diagnosis of epilepsy. ICD-10 codes consistent with a diagnosis of epilepsy together with  $\geq 1$  AED recorded in the Australian National Hospital Morbidity Database resulted in a PPV of 81.4% [73]. A similar study using multiple linked Canadian administrative healthcare databases found a PPV of 91.9% [74] and there are further studies with equivalent findings indicating algorithmic approaches applied to routinely recorded data are sensitive for identifying individuals with diagnoses of epilepsy [75-78]. However, such studies providing verification of the use of routinely recorded data for diagnosis frequently use written medical records as the comparator, themselves routinely recorded.

This research has assessed the agreement compared to standard prospective data collection methods used in RCTs, which remain the gold standard for the approval of novel treatments in healthcare [110]. The poor agreement identified raises questions regarding the accuracy and methods for assessment of accuracy of routinely recorded data for use in retrospective, in addition to prospective research. A notable limitation of this conclusion is the inclusion of the prescription of AEDs in the diagnostic algorithms in these published studies. The inclusion of AEDs in the diagnostic algorithm was not possible in this research as a result of the lack of suitable data in the datasets. Furthermore, this research involved the identification of new diagnoses and the inclusion of AEDs in a diagnostic algorithm would not be appropriate.

#### *9.2.6.2 Implications for Clinical Practice*

Finally, the attributes of routinely recorded data, extracted from electronic medical records have potential implications for clinical practice both at the individual and population level. Missing or inaccurate clinical data may have negative implications on an individual patient's clinical care, particularly with the increasingly central role of electronic medical records. For example, inaccurate diagnostic data may result in eligible individuals being missed as part of screening investigations and inaccurate data regarding the occurrence of seizures and adverse events may result in the limited utility of electronic medical records in monitoring treatment effectiveness.

Furthermore, on a population level, routinely recorded data are frequently used to monitor the incidence and prevalence of diagnoses or associated risk factors. If such data are inaccurate decisions regarding the commissioning and allocation of treatment resources may result in inequitable clinical services.

### **9.3 Recommendations and Further Research**

Recommendations and suggestions for further research can be proposed with the aim of improving the quality of UK routinely recorded data and feasibility of use in clinical research. *Table 8.2* in Chapter Eight has previously summarised a number of recommendations.

#### **9.3.1 General Recommendations**

##### ***Further Research to Assess and Improve the Use of Routinely Recorded Data in Clinical Research***

There is evidence of the inclusion of routinely recorded data in prospective clinical research, including all stages of a RCT being conducted using an administrative healthcare database [124], despite the notable limitations identified in this research. Further research is urgently required to assess the quality, additional benefits, feasibility and cost-efficiency of accessing routinely recorded data to assist with the identification of eligible individuals and recruitment into trials and measure RCT outcomes. Such research should include patients and outcomes in epilepsy together with other disease areas. To inform future prospective assessments, systematic reviews in individual disease areas should aim to identify the use of routinely recorded data in research and review assessments of the accuracy (compared to medical records) and agreement (compared to standard prospective research methods). Such reviews with a focus on methodology should include relevant electronic database searches and as a result of the poor recording of methodological information, also be complemented with a manual review of relevant resources.

A proposed method to complete such methodological research assessing the quality, additional benefits, feasibility and cost-efficiency of accessing routinely recorded data to assist with the identification of eligible individuals and recruitment into trials and measure RCT outcomes compared to standard prospective methods is to use the 'Studies Within A Trial' (SWAT) approach. The SWAT initiative has been developed by the Medical Research Council Hubs for Trials Methodology Research (MRC HTMR) and proposes embedding methodological research within an existing prospective trial [206].

In the majority of current publically funded UK RCTs such as SANAD II, the retrieval of data from administrative healthcare databases is frequently standard, in order to inform the healthcare economic analyses. Using the SWAT approach, a methodological sub-study can be embedded within the trial and prospectively completed, for example with calculation of the trial outcomes using routinely recorded data and comparison to outcomes calculated using the data collected using standard prospective methods. This process would be procedurally efficient as the data would frequently be requested and examined as part of the trial routinely and ethically sound as no additional or subsequent request or retrieval of data would be necessary. It may be reasonable for such an assessment to be included in all future UK RCTs and including this requirement within the stipulations for funding from major national funders such as the National Institute for Health Research Health Technology Assessment (NIHR HTA) programme would ensure the inclusion of such methodological assessments are completed.

As an extension to the above prospective assessments, the relative value of routinely recorded data and 'optimal mix' of routinely recorded data and data collected using standard methods requires assessment for specific disease areas and outcomes. Value of Information (VOI) analysis is a method used to assess the return on investment in research and can be defined as the amount an individual would be willing to pay before making a decision [202]. VOI analysis would be appropriate for application in future prospective studies to determine the 'optimal mix' of data, or the value of information derived from routinely recorded data compared to data collected using standard methods. The inclusion of VOI analysis prospectively and retrieval of routinely recorded data for the complete study sample would permit calculation of the objective function, Incremental Net Benefit (INB) derived from the Incremental Cost Effectiveness Ratio (ICER). As a direct extension of this research, using the complete SANAD II dataset and data retrieved from routine sources, a VOI analysis could be performed to quantitatively define the value of retrieving routinely recorded data in addition to collecting data using standard methods.

Lastly, to maximise the validity of future prospective assessments, the content, strengths and limitations of the routinely recorded datasets must be considered and inform the selection of the specific variables to be assessed.

### ***Standardisation and Rationalisation of Costs***

The costs required for data access from routine data sources where data were available in this research varied widely, although all reportedly operated on a cost recovery, not-for-profit basis. For both clinical and non-clinical sources of routinely recorded data, costs for data access for research should be standardised and rationalised between data sources.

The feasibility of achieving costs standardisation will be greater for publically funded data sources, although similar approaches to achieving standardisation could also be taken for privately funded sources. Proposed approaches would include the establishment of a task force or working group with representatives from each routine data source. Prospective analysis of costs should follow and would inform the definition of 'standardised' costs that each routine data source would aim to adhere to. Such an initiative may be more successful if government driven and medical researchers and funders lobbying relevant governmental departments, including the Department of Health could also be suggested. Where there are reasonable variations in costs, government subsidies could assist in ensuring the cost to the end users, including the researchers, is standardised.

### ***Reduction of the Time to Data Availability***

There is a time delay before data are available in all routine data sources resulting from the requirement to collect, transfer, clean and process data. Although less of an issue for retrospective research, this represents a significant limitation to the access of routinely recorded data for prospective research, including RCTs. This limitation will, for many RCTs, rule out the sole use of routinely recorded data to measure RCT outcomes, where prompt reporting is ethically important and frequently a regulatory requirement. The development of 'real-time' routine data recording is an unrealistic short-term aim as a result of the need to process, clean and format the data. However, development of the infrastructure and data extraction and processing procedures and increasing resources dedicated to such tasks should aim to reduce the time to data availability.

### ***Standardise and Improve Data Linkage***

Improvements in the access to linked data are on-going, for example with the establishment of the ADRN. However, further improvements could be suggested. To improve the accuracy of data linkage, a standardised set of identifying variables could be recorded by all (clinical and non-clinical) data sources. Suggested variables would include name, date of birth, National Insurance Number and NHS Number.

Furthermore, standardised demographic variables could also be recorded, with the potential to inform research. Definition of specific variables would require further research involving researchers, data holders and members of the public as important stakeholders, similar to process involved in the development of a Core Outcome Set for RCTs [204].

The experience during this research is that the process for retrieving data through the ADRN is inefficient; an application must be submitted and approved by the ADRN who then subsequently approach each organisation individually. A recommendation for future improvement could include the storage of de-identified linked data from participating clinical and non-clinical organisations in a single repository, similar to those established for RCT data [201]. This would create a single point of access and remove the burden for each organisation to consider each study individually. Development of this repository would however require significant information governance and security barriers to be cleared and in light of recent developments in data protection regulation [51] applicable to research, individual consent. Including the public as stakeholders in the development of such a data repository would be essential [82].

#### ***Assess Public Perceptions and Improve Public Engagement***

Finally, in the current climate the public mistrust regarding the sharing and linking of routinely recorded data will hamper future efforts to develop linked routinely recorded administrative databases, despite the likely benefits to individuals and the population. This issue is perhaps the most important hurdle to overcome as, despite an improved quality of data, feasibility of use and potential benefits, without public trust and consent, improved implementation and utility of routinely recorded data will not be possible. Further research is required with public engagement to define the issues of most importance to members of the public and assess perspectives with regards the routine recording of data and subsequent use for secondary purposes including research. Such an assessment could include qualitative methods including interviews and focus groups with relevant stakeholders and quantitative methods such as discrete choice experiments for preference elicitation [207].

### **9.3.2 Routinely Recorded Clinical Data**

#### ***Streamline the Process for Application Development and Provisional Review***

Access to routinely recorded secondary care data was achieved in this study, although the process for application development and provisional review could be improved. A number of requirements must be fulfilled prior to submission of the application to NHS Digital and SAIL, including research ethics and governance approvals. General guidance has been published by each source regarding the development of the consent materials.

However guidance regarding, for example the specific phrasing required in the consent form was not published, and discussion with the Information Governance teams was suggested. However the process for requesting review of the relevant study documents and obtaining specific feedback was inefficient, with delays and duplication of work for both the researcher and data holders. To improve this period of protocol development and application, data sources should improve the guidance and formalise the pathway for requesting review of study documents prior to formal submission of the application for research ethics and governance approvals.

#### ***Standardise Data Recording Between Similar Datasets***

'Missing data' was identified throughout this research and a contributing factor may be the requirements for data recording within each dataset. For example, diagnostic data in the inpatient datasets is mandatory. However, the record of diagnosis is not mandatory in the emergency care or outpatient datasets. Consequently, a high proportion of attendances were observed with no diagnostic information recorded. Individuals attending the emergency department without subsequent inpatient admission therefore, were unlikely to have diagnosis recorded. Significant improvements could be expected if the record of diagnostic data is mandatory in all datasets. This is particularly relevant for research involving epilepsy as patients are frequently discharged from the emergency department, without inpatient admission following the occurrence of a seizure. Furthermore, to improve the accuracy of clinical coding, clinicians could have greater involvement in the recording of diagnostic codes. In primary care, GP's are frequently involved in the direct recording of relevant READ codes. The advent of electronic medical records and patient administration systems should improve the ease of this process in secondary care and therefore the acceptability for clinicians. For example, in the outpatient dataset, diagnostic codes permitting classification of epilepsy were recorded for all participants in one centre

and rarely recorded in all other centres. This single centre requires the clinician select a diagnostic code from a standardised electronic form following the clinic attendance and at the point of confirming follow-up arrangements, resulting in all patients having a diagnostic code recorded.

### ***Development of National Primary Care Datasets***

Access to primary care routinely recorded data was not feasible in this study. The 'de-identified' nature of the data, cost and poor geographical coverage are notable limitations. Development of the infrastructure to record national primary care data coverage could be suggested. This may be achieved either through improved collaboration between existing routine data sources with individual-level data linkage or the development of national data sources, such as the NHS Digital General Practice Extraction Service [139].

### ***Development of an Integrated Electronic Healthcare Record***

A number of countries have integrated healthcare systems allowing for national administrative healthcare databases, such as the Swedish Hospital Discharge Register, the Danish National Hospital Register and the Canadian Chronic Disease Surveillance System. In these examples it is possible to retrieve routinely recorded data from electronic medical records for individuals across hospital inpatient admissions and emergency care, outpatient clinic and primary care attendances.

An integrated electronic health record would perhaps have the greatest potential for the improved record of data and improved use in clinical practice and research. Such a proposed system would be entirely electronic and would involve the clinician recording free text 'medical record' entries, in addition to involvement in the direct selection of relevant clinical codes, such as diagnoses. This may result in improving the accuracy of routinely recorded coded data, reducing the potential for transcription errors and improving the efficiency of data record. In the optimal scenario, medical records from primary and secondary care settings, together with additional data such as pharmacy prescribing and dispensing data would be included in the single electronic record. The electronic health record would then contribute directly to individual clinical care and directly provide data to the administrative databases for secondary uses, including disease monitoring and clinical research. The utility of such a dataset for retrospective and prospective research, compared to the datasets currently available in the UK is likely to be significantly improved.

For example, the most sensitive and specific algorithms used in retrospective research to identify individuals with a diagnosis of epilepsy include ICD 10 codes from all clinical settings together with data regarding the prescription of AEDs, with Positive Predictive Values of up to 91.9% [74]. Application of a similar algorithm in the UK would require linking between a number of individual datasets, which is logistically more difficult and introduces the potential for error. Integration of research functions, such as including RCT electronic CRFs [208], into the electronic health record would also improve their utility in prospective clinical research.

Finally, a single integrated electronic health record may result in improved accessibility for individual patients. This may permit patients to view their medical records but also provide the facility for their contribution of certain data. For example, permitting patients to record their dates of seizure occurrence would inform their clinical care, such as alerting the clinical team to unexpected increased seizure frequency, in addition to the obvious benefits for clinical research.

### ***9.3.3 Routinely Recorded Non-Clinical Data***

#### ***Further Research to Assess and Improve the Use of Non-Clinical Data in Research***

Further to the relevant general recommendations presented above, urgent research is required in the short to medium term to improve access to individual-level data from non-clinical sources for research:

##### ***Public Perceptions and Acceptability***

Research regarding the public perceptions and acceptability of using personal non-clinical data, including economic data for clinical research is needed, prior to widespread investment in development of the infrastructure to permit this. Such research may include qualitative methods such as interviews and focus groups and quantitative methods such as discrete choice experiments for preference elicitation.

##### ***Formalisation of the Approval Processes through the ADRN***

In the short-term, the greatest potential for improved data access is likely to involve the ADRN. Formalisation of the approval processes through the independent ADRN and the storage of linked data in a single repository, as previously discussed, would improve efficiency and reduce the burden placed upon the individual data holders. Furthermore, utilising the ADRN who retrieve data on a de-identified basis before providing to the researcher could be proposed as a method of improving data security, alleviating the concerns raised by the DVLA during this research.

##### ***Internal Review of Feasibility, Processes and Resource Implications***

The feasibility, methodological process and resource implications of permitting access to data for research requires review internally within relevant data sources such as the HMRC and DWP. The initiative would be best and likely only led from government and therefore lobbying government, once the public perceptions and acceptability are known, may be a suitable initial option.

#### **9.4 Concluding Remarks**

In Chapter One, routinely recorded data in the UK and the potential for use in clinical research were introduced. Epilepsy and the case study RCT SANAD II were subsequently introduced before finally the objectives of this research were presented. In Chapter Two the use of routinely recorded data in RCTs in the UK was reviewed and in Chapter Three the agreement of UK routinely recorded data compared to data collected using standard methods in prospective studies was assessed in a systematic review. In Chapter Four, sources of routinely recorded data in the UK relevant to the outcomes of SANAD II were reviewed and sources where routinely recorded data were accessible for individuals recruited into SANAD II were identified.

In Chapter Five, the methods for the assessment of the quality and agreement of routinely recorded data retrieved from electronic medical records compared to data collected using standard prospective methods in a randomised controlled trial were presented. The assessment of seizure occurrence, diagnosis and classification of epilepsy in routinely recorded datasets was assessed in Chapter Six and variables and outcome measures relevant to the follow-up of participants in SANAD II were assessed in Chapter Seven. The feasibility and efficiency of accessing and using routinely recorded data for participants in SANAD II was assessed in Chapter Eight.

In this chapter, the significant results for each objective of this research and the implications for clinical practice and research have been discussed. Recommendations for improving the use of routinely recorded data have been proposed and avenues for further research suggested.

This research has identified limited previous experience of using routinely recorded data in UK RCTs. The accessibility and feasibility of use were limited and degree of missing data and agreement compared to data collected using standard methods unsatisfactory. The results of this research suggest routinely recorded data in the context of prospective clinical research may be an important source of additional data, for example to identify additional events such as seizures not recorded using standard methods. The results suggest that use of routinely recorded data as the primary data source or as a means of validating data collected using standard methods for the variables assessed in this study, would be limited.

Although these results are potentially disappointing for immediate improvements in the use of routinely recorded data in RCTs, such as improvements in resource and cost-efficiency, they enable recommendations to be proposed.

Substantial further development is now required to improve the utility of routinely recorded data in clinical practice and research. Recommendations include suggestions for improving the access to routinely recorded data for research, development of an integrated electronic health record for use in both clinical practice and research, further assessment of the attributes and 'optimal mix' of routinely recorded data compared to data collected using standard methods. To improve the likelihood of significant progress, initiatives for development should be led from the government. Additionally and perhaps most importantly acknowledging recent controversies, involving patients and the public as important stakeholders and re-gaining their trust will be essential in realising the individual and population healthcare benefits of routinely recorded data.

## References

1. Bone, I., *Neurology and Neurosurgery Illustrated*. 4th ed. 2004.
2. Guberman, A., *Essentials of Clinical Epilepsy*. 2nd ed. 1999: Butterworth Heinemann: Boston.
3. Hauser, W.A., J.F. Annegers, and W.A. Rocca, *Descriptive epidemiology of epilepsy: Contributions of population-based studies from Rochester, Minnesota*. Mayo Clinic Proceedings, 1996. **71**(6): p. 576-586.
4. De La Court, A., et al., *Prevalence of epilepsy in the elderly: The Rotterdam study*. Epilepsia, 1996. **37**(2): p. 141-147.
5. *Proposal for Revised Clinical and Electroencephalographic Classification of Epileptic Seizures: From the Commission on Classification and Terminology of the International League Against Epilepsy*. Epilepsia, 1981. **22**(4): p. 489-501.
6. Fisher, R.S., et al., *Operational classification of seizure types by the International League Against Epilepsy: Position Paper of the ILAE Commission for Classification and Terminology*. Epilepsia, 2017. **58**(4): p. 522-530.
7. Fisher, R.S., et al., *ILAE Official Report: A practical clinical definition of epilepsy*. Epilepsia, 2014. **55**(4): p. 475-482.
8. Fisher, R.S., et al., *Epileptic seizures and epilepsy: Definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE)*. Epilepsia, 2005. **46**(4): p. 470-472.
9. *Proposal for Revised Classification of Epilepsies and Epileptic Syndromes: Commission on Classification and Terminology of the International League Against Epilepsy*. Epilepsia, 1989. **30**(4): p. 389-399.
10. Scheffer, I.E., et al., *ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology*. Epilepsia, 2017. **58**(4): p. 512-521.
11. Mrcp, M.M., et al., *National general practice study of epilepsy (ngpse): Partial seizure patterns in a general population*. Neurology, 1992. **42**(10): p. 1911-1917.
12. Walker, L.E., et al., *Personalized medicine approaches in epilepsy*. Journal of Internal Medicine, 2015. **277**(2): p. 218-234.
13. Kjeldsen, M.J., et al., *Epileptic seizures and syndromes in twins: The importance of genetic factors*. Epilepsy Research, 2003. **55**(1-2): p. 137-146.
14. Saxena, S. and J. Orley, *Quality of life assessment: The World Health Organization perspective*. European Psychiatry, 1997. **12**(SUPPL. 3): p. 263S-266S.
15. Cramer, J.A., et al., *The impact of seizures and adverse effects on global health ratings*. Epilepsy and Behavior, 2007. **11**(2): p. 179-184.
16. Szaflarski, M., et al., *Quality of life in medication-resistant epilepsy: The effects of patient's age, age at seizure onset, and disease duration*. Epilepsy and Behavior, 2006. **8**(3): p. 547-551.
17. Suurmeijer, T.P.B.M., M.F. Reuvekamp, and B.P. Aldenkamp, *Social functioning, psychological functioning, and quality of life in epilepsy*. Epilepsia, 2001. **42**(9): p. 1160-1168.
18. UK Government. *Driver and Vehicle Licensing Agency*. 2016 [cited 2016 02/05/2016]; Available from: <https://www.gov.uk/government/organisations/driver-and-vehicle-licensing-agency>.
19. Shorvon, S.D. and D.M.G. Goodridge, *Longitudinal cohort studies of the prognosis of epilepsy: Contribution of the National General Practice Study of Epilepsy and other studies*. Brain, 2013. **136**(11): p. 3497-3510.

20. Cockerell, O.C., et al., *Remission of epilepsy: results from the National General Practice Study of Epilepsy*. The Lancet, 1995. **346**(8968): p. 140-144.
21. Kwan, P. and M.J. Brodie, *Early identification of refractory epilepsy*. New England Journal of Medicine, 2000. **342**(5): p. 314-319.
22. Stephen, L.J. and M.J. Brodie, *Pharmacotherapy of epilepsy: Newly approved and developmental agents*. CNS Drugs, 2011. **25**(2): p. 89-107.
23. Marson, A.G., et al., *The SANAD study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial*. Lancet, 2007. **369**(9566): p. 1016-1026.
24. Marson, A.G., et al., *The SANAD study of effectiveness of carbamazepine, gabapentin, lamotrigine, oxcarbazepine, or topiramate for treatment of partial epilepsy: an unblinded randomised controlled trial*. Lancet, 2007. **369**(9566): p. 1000-1015.
25. Nunes, V.D., et al., *Diagnosis and management of the epilepsies in adults and children: Summary of updated NICE guidance*. BMJ (Online), 2012. **344**(7842).
26. Glauser, T., et al., *Updated ILAE evidence review of antiepileptic drug efficacy and effectiveness as initial monotherapy for epileptic seizures and syndromes*. Epilepsia, 2013. **54**(3): p. 551-563.
27. Tomson, T., et al., *Valproate in the treatment of epilepsy in girls and women of childbearing potential*. Epilepsia, 2015. **56**(7): p. 1006-1019.
28. Voinescu, P.E. and P.B. Pennell, *Management of epilepsy during pregnancy*. Expert Review of Neurotherapeutics, 2015. **15**(10): p. 1171-1187.
29. Greenwood, R.S., *Adverse effects of antiepileptic drugs*. Epilepsia, 2000. **41**(6 SUPPL. 2): p. S42-S52.
30. Mei, P.A., et al., *Pharmacovigilance in epileptic patients using antiepileptic drugs*. Arquivos de Neuro-Psiquiatria, 2006. **64**(2 A): p. 198-201.
31. Baker, G.A., A. Jacoby, and D.W. Chadwick, *The associations of psychopathology in epilepsy: A community study*. Epilepsy Research, 1996. **25**(1): p. 29-39.
32. Karlsson, L., B. Wettermark, and T. Tomson, *Drug treatment in patients with newly diagnosed unprovoked seizures/epilepsy*. Epilepsy Research, 2014. **108**(5): p. 902-908.
33. Bonnett, L.J., et al., *Time to 12-month remission and treatment failure for generalised and unclassified epilepsy*. Journal of Neurology, Neurosurgery and Psychiatry, 2014. **85**(6): p. 603-610.
34. Bonnett, L.J., et al., *Treatment outcome after failure of a first antiepileptic drug*. Neurology, 2014. **83**(6): p. 552-560.
35. Kwan, P., et al., *Definition of drug resistant epilepsy: Consensus proposal by the ad hoc Task Force of the ILAE commission on therapeutic strategies*. Aktuelle Neurologie, 2010. **37**(8): p. 372-381.
36. NICE. *NICE Glossary: Randomised Controlled Trial*. 2017 [cited 2017 3/11/17]; Available from: <https://www.nice.org.uk/glossary?letter=r>.
37. Maguire, M.J., et al., *Reporting and analysis of open-label extension studies of anti-epileptic drugs*. Epilepsy Research, 2008. **81**(1): p. 24-29.
38. Rheims, S., et al., *Factors determining response to antiepileptic drugs in randomized controlled trials. A systematic review and meta-analysis*. Epilepsia, 2011. **52**(2): p. 219-233.
39. Hemming, K., et al., *Vigabatrin for refractory partial epilepsy*. Cochrane database of systematic reviews (Online), 2013. **1**.
40. Beacher, N.G., M.J. Brodie, and C. Goodall, *A case report: Retigabine induced oral mucosal dyspigmentation of the hard palate*. BMC Oral Health, 2015.

41. Tomson, T., et al., *Dose-dependent risk of malformations with antiepileptic drugs: An analysis of data from the EURAP epilepsy and pregnancy registry*. The Lancet Neurology, 2011. **10**(7): p. 609-617.
42. Bothwell, L.E., et al., *Assessing the gold standard - Lessons from the history of RCTs*. New England Journal of Medicine, 2016. **374**(22): p. 2175-2181.
43. Sertkaya, A., *Examination of clinical trial costs and barrier for drug development: report to the Assistant Secretary of Planning and Evaluation, D.o.H.a.H. Services, Editor*. 2014.
44. Bodenheimer, T., *Uneasy alliance - clinical investigators and the pharmaceutical industry*. New England Journal of Medicine, 2000. **342**(20): p. 1539-1544.
45. Bourgeois, F.T., S. Murthy, and K.D. Mandl, *Outcome reporting among drug trials registered in ClinicalTrials.gov*. Annals of Internal Medicine, 2010. **153**(3): p. 158-166.
46. Lewsey, J.D., et al., *Using routine data to complement and enhance the results of randomised controlled trials*. Health Technology Assessment, 2000. **4**(22): p. i+iii-iv+1-45.
47. Loke, Y.K., *Use of databases for clinical research*. Archives of Disease in Childhood, 2014. **99**(6): p. 587-589.
48. Garrett E, B.H., Dibbon C *Health Administrative Data: Exploring the potential for academic research*. 2010.
49. McKee, M., *Routine data: a resource for clinical audit?* Quality in health care : QHC, 1993. **2**(2): p. 104-111.
50. UKGovernment. *The Data Protection Act*. 2016 Accessed March 2016 [cited 2016 03/01/2016]; Available from: (<https://www.gov.uk/data-protection/the-data-protection-act>)
51. UKGovernment. *The General Data Protection Regulation (GDPR), Information Comissioners Office (ICO)*. 2018 [cited 2018 Feb 2018]; Available from: <https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/>.
52. UKGovernment. *The Freedom of Information Act*. 2016 [cited 2016 04/03/2016]; Available from: (<http://www.legislation.gov.uk/ukpga/2000/36/contents>)
53. Health, D.o. *Your Data: Better Security, Better Choice, Better Care*. 2017 [cited 2017 23/10/2017]; Available from: [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/627493/Your\\_data\\_better\\_security\\_better\\_choice\\_better\\_care\\_government\\_response.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/627493/Your_data_better_security_better_choice_better_care_government_response.pdf).
54. NHEngland. *NHS Payment System*. 2016 [cited 2016 18/07/2016]; Available from: (<http://www.england.nhs.uk/resources/pay-syst>). .
55. NHSDigital. *The Health and Social Care Information Centre*. 2016 [cited 2016 16th July]; Available from: <http://www.hscic.gov.uk>.
56. World-Health-Organisation. *International Statistical Classification of Diseases and Related Health Problems (ICD) 10*. 2016 [cited 2016 18/07/2016]; Available from: <http://apps.who.int/classifications/icd10/browse/2016/en>.
57. NHSDigital. *UK READ Codes*. 2017 [cited 2017 21/03/17]; Available from: <https://data.gov.uk/dataset/uk-read-code>.
58. Raftery, J., P. Roderick, and A. Stevens, *Potential use of routine databases in health technology assessment*. Health technology assessment (Winchester, England), 2005. **9**(20): p. 1-92, iii-iv.
59. Clarson, L.E., et al., *Increased risk of vascular disease associated with gout: A retrospective, matched cohort study in the UK Clinical Practice Research Datalink*. Annals of the Rheumatic Diseases, 2015. **74**(4): p. 642-647.

60. van Staa, T.P., et al., *The opportunities and challenges of pragmatic point-of-care randomised trials using routinely collected electronic records: Evaluations of two exemplar trials*. Health Technology Assessment, 2014. **18**(43): p. 1-146.
61. IMSHealth. *TrialViz*. 2016 [cited 2016 04/04/2016]; Available from: <http://www.dataline.co.uk/portfolio/cprd-trialviz>.
62. McGregor, J., et al., *The Health Informatics Trial Enhancement Project (HITE): Using routinely collected primary care data to identify potential participants for a depression trial*. Trials [Electronic Resource], 2010. **11**: p. 39.
63. McCowan, C., et al., *Using Electronic Health Records to Support Clinical Trials: A Report on Stakeholder Engagement for EHR4CR*. BioMed Research International, 2015. **2015**.
64. Williams, J.G., et al., *Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment*. Health technology assessment (Winchester, England), 2003. **7**(26): p. iii, v-x, 1-117.
65. Barry, S.J.E., et al., *Are Routinely Collected NHS Administrative Records Suitable for Endpoint Identification in Clinical Trials? Evidence from the West of Scotland Coronary Prevention Study*. PLoS ONE, 2013. **8**(9).
66. Thorn, J.C., et al., *Validating the use of hospital episode statistics data and comparison of costing methodologies for economic evaluation: An end-of-life case study from the cluster randomised trial of PSA testing for prostate cancer (CAP)*. BMJ Open, 2016. **6**(4).
67. Gulliford, M.C., et al., *Cluster randomized trials utilizing primary care electronic health records: Methodological issues in design, conduct, and analysis (eCRT Study)*. Trials, 2014. **15**(1).
68. NWEH. *The Salford Lung Study*. 2016 [cited 2016 18/07/2016]; Available from: <http://www.salfordlungstudy.co.uk>
69. New, J.P., et al., *Obtaining real-world evidence: The Salford Lung Study*. Thorax, 2014. **69**(12): p. 1152-1154.
70. *The Plan for Growth*. 2011.
71. *The NHS Constitution for England*. 2013.
72. Fairweather, N.B. and S. Rogerson, *A moral approach to electronic patient records*. Medical Informatics and the Internet in Medicine, 2001. **26**(3): p. 219-234.
73. Tan, M., et al., *Development and validation of an epidemiologic case definition of epilepsy for use with routinely collected Australian health data*. Epilepsy and Behavior, 2015. **51**: p. 65-72.
74. Reid, A.Y., et al., *Development and validation of a case definition for epilepsy for use with administrative health data*. Epilepsy Research, 2012. **102**(3): p. 173-179.
75. Tu, K., et al., *Assessing the validity of using administrative data to identify patients with epilepsy*. Epilepsia, 2014. **55**(2): p. 335-343.
76. Franchi, C., et al., *Validation of healthcare administrative data for the diagnosis of epilepsy*. Journal of Epidemiology and Community Health, 2013. **67**(12): p. 1019-1024.
77. Jetté, N., et al., *How accurate is ICD coding for epilepsy?* Epilepsia, 2010. **51**(1): p. 62-69.
78. Christensen, J., et al., *Validation of epilepsy diagnoses in the Danish National Hospital Register*. Epilepsy Research, 2007. **75**(2-3): p. 162-170.
79. Weitzman, E.R., L. Kaci, and K.D. Mandl, *Sharing medical data for health research: The early personal health record experience*. Journal of Medical Internet Research, 2010. **12**(2).

80. Haddow, G., et al., *'Nothing is really safe': A focus group study on the processes of anonymizing and sharing of health data for research purposes*. *Journal of Evaluation in Clinical Practice*, 2011. **17**(6): p. 1140-1146.
81. Stevenson, F., *The use of electronic patient records for medical research: Conflicts and contradictions*. *BMC Health Services Research*, 2015. **15**(1).
82. Eugene C Nelson, M.D.-W., Paul B Batalden, Karen, A.D.V.C. Homa, Tamara S Morgan, Elena, and E.S.F. Eftimovska, John Ovretveit, Wade Harrison, Cristin Lind Staffan Lindblad, *Patient focused registries can improve health, care and science*. *BMJ*, 2016. **354**.
83. Tjeerd-Pieter Van Staa, I.B., Ben Goldacre, Liam Smeeth, *Big health data: the need to earn public trust after past management*. *BMJ*, 2016. **354**: p. 95-97.
84. Bouras, G., et al., *Linked hospital and primary care database analysis of the incidence and impact of psychiatric morbidity following gastrointestinal cancer surgery in England*. *Annals of Surgery*, 2016. **264**(1): p. 93-99.
85. Turner, E.L., et al., *Design and preliminary recruitment results of the Cluster randomised triAl of PSA testing for Prostate cancer (CAP)*. *British Journal of Cancer*, 2014. **110**(12): p. 2829-36.
86. NWIS. *NHS Wales Informatics Service*. 2016 [cited 2016 04/05/2016]; Available from: <http://www.wales.nhs.uk/nwis/page/52490>.
87. Ismail, S.I. and B. Puyk, *The rise of obstetric anal sphincter injuries (OASIS): 11-year trend analysis using Patient Episode Database for Wales (PEDW) data*. *Journal of Obstetrics & Gynaecology*, 2014. **34**(6): p. 495-8.
88. ISDSScotland. *The Information Services Division* 2016 [cited 2016 05/05/2016]; Available from: <http://www.isdscotland.org/Products-and-Services/index.asp>.
89. Ahmed, A., et al., *Upper gastrointestinal bleeding in Scotland 2000-2010: Improved outcomes but a significant weekend effect*. *World Journal of Gastroenterology*, 2015. **21**(38): p. 10890-7.
90. CPRD. *The Clinical Practice Research Datalink*. 2016 [cited 2016 04/05/2016]; Available from: <http://www.cprd.com/intro.asp>.
91. TPP. *ResearchOne*. 2016 [cited 2016 04/05/2016]; Available from: <http://www.tpp-uk.com/products/systemone>.
92. Herrett, E., et al., *Text messaging reminders for influenza vaccine in primary care: protocol for a cluster randomised controlled trial (TXT4FLUJAB)*. *BMJ Open*, 2014. **4**(5): p. e004633.
93. QResearch. *QResearch*. 2016 [cited 2016 06/05/2016]; Available from: <http://www.qresearch.org/SitePages/Home.aspx>.
94. Hill, T., et al., *Antidepressant use and risk of epilepsy and seizures in people aged 20 to 64 years: cohort study using a primary care database*. *BMC Psychiatry*, 2015. **15**: p. 315.
95. THIN. *The Health Improvement Network*. 2016 [cited 2016 05/04/2016]; Available from: <http://www.thin-uk.net>.
96. González-Pérez, A., et al., *Incidence and Predictors of Hemorrhagic Stroke in Users of Low-Dose Acetylsalicylic Acid*. *Journal of Stroke and Cerebrovascular Diseases*, 2015. **24**(10): p. 2321-2328.
97. NWEH. *North West eHealth*. 2016 [cited 2016 05/05/2016]; Available from: <http://nweh.org.uk>.
98. ApolloMedical. *Apollo Data Extraction*. 2016 [cited 2016 06/04/2016]; Available from: <http://www.apollo-medical.com/>.
99. GraphnetHealth. *Graphnet*. 2016 [cited 2016 06/04/2016]; Available from: <http://www.graphnethealth.com/what-we-do/overview/what-we-do>.

100. SAIL. *The Secure Anonymised Information Linkage Databank*. 2016 [cited 2016 01/06/2016]; Available from: <http://www.saildatabank.com>.
101. Sayers, A., et al., *Evidence for a persistent, major excess in all cause admissions to hospital in children with type-1 diabetes: results from a large Welsh national matched community cohort study*. *BMJ Open*, 2015. **5**(4): p. e005644.
102. ADRN. *The Administrative Data Research Network*. 2016 [cited 2016 16/06/2016]; Available from: <http://adrn.ac.uk>.
103. UKGovernment. *HM Revenue and Customs*. 2016 [cited 2016 04/04/2016]; Available from: <https://www.gov.uk/government/organisations/hm-revenue-customs>.
104. UKGovernment. *The Department for Work and Pensions*. 2016 [cited 2016 01/05/2016]; Available from: <https://www.gov.uk/government/organisations/department-for-work-pensions>.
105. (US), N.-R.-C., *The Prevention and Treatment of Missing Data in Clinical Trials.*, in *National Academies Press (US)*. 2010.
106. UKGovernment. *HMRC Research Approvals*. 2016 [cited 2016 22/09/2016]; Available from: [https://www.gov.uk/government/uploads/system/uploads/attachment\\_data/file/367874/DatalabProposalsApproved\\_October2014.pdf](https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/367874/DatalabProposalsApproved_October2014.pdf).
107. ONS. *The Office for National Statistics*. 2016 [cited 2016 04/04/2016]; Available from: <https://www.ons.gov.uk>.
108. ONS. *Official Labour Market Statistics*. 2016 [cited 2016 04/04/2016]; Available from: <http://www.nomisweb.co.uk>.
109. OCSI. *Data 4 Neighbourhoods and Regeneration 2016* [cited 2016 07/04/2016]; Available from: <http://www.data4nr.net/introduction>.
110. Friedman, C.P., et al., *Toward a science of learning systems: A research agenda for the high-functioning Learning Health System*. *Journal of the American Medical Informatics Association*, 2015. **22**(1): p. 43-50.
111. Gargon, E., P.R. Williamson, and M. Clarke, *Collating the knowledge base for core outcome set development: Developing and appraising the search strategy for a systematic review*. *BMC Medical Research Methodology*, 2015. **15**(1).
112. Dregan, A., et al., *Cluster randomized trial in the general practice research database: 2. Secondary prevention after first stroke (eCRT study): study protocol for a randomized controlled trial*. *Trials [Electronic Resource]*, 2012. **13**: p. 181.
113. Gulliford, M.C., et al., *Cluster randomised trial in the General Practice Research Database: 1. Electronic decision support to reduce antibiotic prescribing in primary care (eCRT study)*. *Trials [Electronic Resource]*, 2011. **12**: p. 115.
114. Reckless, J., et al., *Projected cost-effectiveness of ezetimibe/simvastatin compared with doubling the statin dose in the United Kingdom: findings from the INFORCE study*. *Value in Health*, 2010. **13**(6): p. 726-34.
115. Subramonia, S. and T. Lees, *Radiofrequency ablation vs conventional surgery for varicose veins - a comparison of treatment costs in a randomised trial*. *European Journal of Vascular & Endovascular Surgery*, 2010. **39**(1): p. 104-11.
116. Moore, G.F., et al., *Impacts of the Primary School Free Breakfast Initiative on socio-economic inequalities in breakfast consumption among 9-11-year-old schoolchildren in Wales*. *Public Health Nutrition*, 2014. **17**(6): p. 1280-9.
117. Mosis, G., et al., *A randomized database study in general practice yielded quality data but patient recruitment in routine consultation was not practical*. *Journal of Clinical Epidemiology*, 2006. **59**(5): p. 497-502.
118. Scholefield, J.H., et al., *Nottingham trial of faecal occult blood testing for colorectal cancer: a 20-year follow-up*. *Gut*, 2012. **61**(7): p. 1036-40.

119. Ashton, H.A., et al., *The Multicentre Aneurysm Screening Study (MASS) into the effect of abdominal aortic aneurysm screening on mortality in men: a randomised controlled trial*. *Lancet*, 2002. **360**(9345): p. 1531-9.
120. Brown, L.C., et al., *The UK EndoVascular Aneurysm Repair (EVAR) trials: randomised trials of EVAR versus standard therapy*. Health Technology Assessment (Winchester, England), 2012. **16**(9): p. 1-218.
121. Bale, G., et al., *Long-term mortality follow-up of the ISOLDE participants: causes of death during 13 years after trial completion*. *Respiratory Medicine*, 2008. **102**(10): p. 1468-72.
122. Brooks, C.J., et al., *Use of a patient linked data warehouse to facilitate diabetes trial recruitment from primary care*. *Primary care diabetes*, 2009. **3**(4): p. 245-8.
123. Dregan, A., et al., *Point-of-care cluster randomized trial in stroke secondary prevention using electronic health records*. *Stroke*, 2014. **45**(7): p. 2066-71.
124. Gulliford, M.C., et al., *Electronic health records for intervention research: a cluster randomized trial to reduce antibiotic prescribing in primary care (eCRT study)*. *Annals of Family Medicine*, 2014. **12**(4): p. 344-51.
125. Horspool, M.J., et al., *Preventing and lessening exacerbations of asthma in school-age children associated with a new term (PLEASANT): study protocol for a cluster randomised control trial*. *Trials [Electronic Resource]*, 2013. **14**: p. 297.
126. Henderson, R.A., et al., *10-Year Mortality Outcome of a Routine Invasive Strategy Versus a Selective Invasive Strategy in Non-ST-Segment Elevation Acute Coronary Syndrome: The British Heart Foundation RITA-3 Randomized Trial*. *Journal of the American College of Cardiology*, 2015. **66**(5): p. 511-20.
127. Molyneux, A.J., et al., *The durability of endovascular coiling versus neurosurgical clipping of ruptured cerebral aneurysms: 18 year follow-up of the UK cohort of the International Subarachnoid Aneurysm Trial (ISAT)*. [Erratum appears in *Lancet*. 2015 Mar 14;385(9972):946]. *Lancet*, 2015. **385**(9969): p. 691-7.
128. Perera, D., et al., *Long-term mortality data from the balloon pump-assisted coronary intervention study (BCIS-1): a randomized, controlled trial of elective balloon counterpulsation during high-risk percutaneous coronary intervention*. *Circulation*, 2013. **127**(2): p. 207-12.
129. Simmons, R.K., et al., *Screening for type 2 diabetes and population mortality over 10 years (ADDITION-Cambridge): a cluster-randomised controlled trial*. *Lancet*, 2012. **380**(9855): p. 1741-8.
130. NHS Digital. *HSCIC Data Release Registers*. 2016 [cited 2016 21/09/16]; Available from: <http://digital.nhs.uk/dataregister>.
131. CPRD. *Approved Studies*. 2016 [cited 2016 21/09/16]; Available from: <https://www.cprd.com/ISAC/datause.asp>.
132. TPP. *ResearchOne Current Projects*. 2016 [cited 2016 22/09/16]; Available from: <http://www.researchone.org/category/current-projects/>.
133. QResearch. *QResearch Publications*. 2016 [cited 2016 22/09/2016]; Available from: <http://www.qresearch.org/SitePages/publications.aspx>.
134. THIN. *THIN Bibliography*. 2016 [cited 2016 22/09/2016]; Available from: <http://csdmruk.cegedim.com/THINBibliography.pdf>.
135. NWEH. *NWEH Linked Database System*. 2016 [cited 2016 22/09/2016]; Available from: <http://nweh.co.uk/products/linked-database-system>.
136. SAIL. *SAIL Publications*. 2016 [cited 2016 22/09/2016]; Available from: <http://www.saildatabank.com/data-dictionary/publications>.
137. Snooks, H., et al., *Support and assessment for fall emergency referrals (SAFER 2) research protocol: Cluster randomised trial of the clinical and cost effectiveness of*

- new protocols for emergency ambulance paramedics to assess and refer to appropriate community-based care.* BMJ Open, 2012. **2**(6).
138. ADRN. *ADRN Approved Projects*. 2016 [cited 2016 22/09/2016]; Available from: <https://adrn.ac.uk/research-projects/approved-projects/>.
  139. NHS Digital. *General Practice Extraction Service*. 2016 [cited 2016 05/04/2016]; Available from: <http://www.hscic.gov.uk/gpes>.
  140. Soriano, J.B., et al., *Validation of general practitioner-diagnosed COPD in the UK General Practice Research Database*. European Journal of Epidemiology, 2001. **17**(12): p. 1075-80.
  141. Tudur Smith, C., et al., *The Value of Source Data Verification in a Cancer Clinical Trial*. PLoS ONE, 2012. **7**(12).
  142. Thorn, J.C., et al., *Validation of the Hospital Episode Statistics Outpatient Dataset in England*. Pharmacoeconomics, 2016. **34**(2): p. 161-168.
  143. Lewsey, J.D., et al., *Comparing outcomes of percutaneous transluminal coronary angioplasty with coronary artery bypass grafting; can routine health service data complement and enhance randomized controlled trials?* European Heart Journal, 1999. **20**(23): p. 1731-5.
  144. Tannen, R.L., M.G. Weiner, and S.M. Marcus, *Simulation of the Syst-Eur randomized control trial using a primary care electronic medical record was feasible*. Journal of Clinical Epidemiology, 2006. **59**(3): p. 254-64.
  145. Tannen, R.L., M.G. Weiner, and D. Xie, *Replicated studies of two randomized trials of angiotensin-converting enzyme inhibitors: further empiric validation of the 'prior event rate ratio' to adjust for unmeasured confounding by indication*. Pharmacoepidemiology & Drug Safety, 2008. **17**(7): p. 671-85.
  146. Tannen, R.L., M.G. Weiner, and D. Xie, *Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings*. BMJ, 2009. **338**: p. b81.
  147. Tannen, R.L., et al., *A simulation using data from a primary care practice database closely replicated the women's health initiative trial*. Journal of Clinical Epidemiology, 2007. **60**(7): p. 686-95.
  148. Tannen, R.L., et al., *Estrogen affects post-menopausal women differently than estrogen plus progestin replacement therapy*. Human Reproduction, 2007. **22**(6): p. 1769-77.
  149. Bryant, M., et al., *Agreement between routine and research measurement of infant height and weight*. Archives of Disease in Childhood, 2015. **100**(1): p. 24-9.
  150. Cleland, J.A., et al., *An exploratory, pragmatic, cluster randomised trial of practice nurse training in the use of asthma action plans*. Primary Care Respiratory Journal, 2007. **16**(5): p. 311-8.
  151. Dixon, S., et al., *Is it cost effective to introduce paramedic practitioners for older people to the ambulance service? Results of a cluster randomised controlled trial*. Emergency Medicine Journal, 2009. **26**(6): p. 446-51.
  152. Doshi, J., et al., *Post-tonsillectomy morbidity statistics: are they underestimated?* Journal of Laryngology & Otology, 2008. **122**(4): p. 374-7.
  153. Kingston, N., et al., *Assessing the amount of unscheduled screening ("contamination") in the control arm of the UK "Age" Trial*. Cancer Epidemiology, Biomarkers & Prevention, 2010. **19**(4): p. 1132-6.
  154. Hutchings, H.A., et al. *Can electronic routine data act as a surrogate for patient-assessed outcome measures?* International Journal of Technology Assessment in Health Care, 2005. **21**, 138-43.

155. Iyer, R., et al., *Patient-reporting improves estimates of postoperative complication rates: a prospective cohort study in gynaecological oncology*. British Journal of Cancer, 2013. **109**(3): p. 623-32.
156. Steward, W.P., et al., *Chemotherapy administration and data collection in an EORTC collaborative group--can we trust the results?* European Journal of Cancer, 1993. **29A**(7): p. 943-7.
157. Weiner, M.G., D. Xie, and R.L. Tannen, *Replication of the Scandinavian Simvastatin Survival Study using a primary care medical record database prompted exploration of a new method to address unmeasured confounding*. Pharmacoepidemiology & Drug Safety, 2008. **17**(7): p. 661-70.
158. Byford, S., et al., *Comparison of alternative methods of collection of service use data for the economic evaluation health care interventions*. Health Economics, 2007. **16**(5): p. 531-536.
159. Petrou, S., et al., *The accuracy of self-reported healthcare resource utilization in health economic studies*. International Journal of Technology Assessment in Health Care, 2002. **18**(3): p. 705-710.
160. Kennedy, A.D.M., et al., *Resource use data by patient report or hospital records: Do they agree?* BMC Health Services Research, 2002. **2**: p. 1-5.
161. Delaney, J.A., E.E. Moodie, and S. Suissa, *Validating the effects of drug treatment on blood pressure in the General Practice Research Database*. Pharmacoepidemiology & Drug Safety, 2008. **17**(6): p. 535-45.
162. Dobbie, F., et al., *Evaluating Long-term Outcomes of NHS Stop Smoking Services (ELONS): a prospective cohort study*. Health Technology Assessment (Winchester, England), 2015. **19**(95): p. 1-156.
163. Ishihara-Paul, L., et al., *Prospective association between emotional health and clinical evidence of Parkinson's disease*. European Journal of Neurology, 2008. **15**(11): p. 1148-54.
164. Cornish, R.P., et al., *Using linked educational attainment data to reduce bias due to missing outcome data in estimates of the association between the duration of breastfeeding and IQ at 15 years*. International Journal of Epidemiology, 2015. **44**(3): p. 937-45.
165. Andersen, J.R., et al., *Impact of source data verification on data quality in clinical trials: an empirical post hoc analysis of three phase 3 randomized clinical trials*. British Journal of Clinical Pharmacology, 2015. **79**(4): p. 660-8.
166. Mitchell, A.J., et al., *Is there a difference between hospital-verified and self-reported self-harm? Implications for repetition*. General Hospital Psychiatry, 2016. **43**: p. 12-16.
167. Richards, S.H., J. Coast, and T.J. Peters, *Patient-reported use of health service resources compared with information from health providers*. Health and Social Care in the Community, 2003. **11**(6): p. 510-518.
168. Shaw, A., et al., *Can we trust the quality of routine hospital outpatient information in the UK? Validating outpatient data from the patient administration system (PAS)*. Journal of Health Services Research and Policy, 1998. **3**(4): p. 203-206.
169. Barbara, A.M., et al., *Agreement between self-report and medical records on signs and symptoms of respiratory illness*. Primary Care Respiratory Journal, 2012. **21**(2): p. 145-152.
170. Mukerjee, A.K., *Ascertainment of non-respiratory tuberculosis in five boroughs by comparison of multiple data sources*. Communicable disease and public health / PHLS, 1999. **2**(2): p. 143-144.
171. O'Brien, J. and C. Bowie, *A methodology for collecting outcome measures for common hospital conditions*. Journal of Public Health, 1992. **14**(4): p. 380-384.

172. Breeman, S., et al. *Patient reported clinical outcomes: the challenges and implications for randomised controlled trials [abstract]*. *Trials*, 2011. **12**, A72.
173. DIRUM. *Database of Instruments for Resource Use Measurement*. 2012 [cited 2016 October 2016]; Available from: <http://www.dirum.org/>.
174. Chishti, T., et al., *How reliable are stroke patients' reports of their numbers of general practice consultations over 12 months?* *Family Practice*, 2013. **30**(1): p. 119-122.
175. Ford, T., et al., *The children's services interview: Validity and reliability*. *Social Psychiatry and Psychiatric Epidemiology*, 2007. **42**(1): p. 36-49.
176. Mistry, H., et al., *Comparison of general practitioner records and patient self-report questionnaires for estimation of costs*. *European Journal of Health Economics*, 2005. **6**(3): p. 261-266.
177. Morrell, C.J., et al., *Costs and benefits of community postnatal support workers: A randomised controlled trial*. *Health Technology Assessment*, 2000. **4**(6): p. i-iii+1-77.
178. Britton, A., et al., *Validating self-reported strokes in a longitudinal UK cohort study (Whitehall II): Extracting information from hospital medical records versus the Hospital Episode Statistics database*. *BMC Medical Research Methodology*, 2012. **12**: p. 83.
179. Husain, M.J., et al., *HERALD (health economics using routine anonymised linked data)*. *BMC Medical Informatics & Decision Making*, 2012. **12**: p. 24.
180. Pastorino, S., et al., *Validation of self-reported diagnosis of diabetes in the 1946 British birth cohort*. *Primary care diabetes*, 2015. **9**(5): p. 397-400.
181. Smith, I., S. Brown, and S. Coleman, *Assessing the accuracy of routinely collected data and their potential in pressure ulcer trials*. 2015: Poster Presentation: International Clinical Trials Methodology Conference 2015.
182. Herrington, W., K. Wallendszus, and L. Bowman, *Can vascular mortality be reliably ascertained from the underlying cause of death recorded on a medical death certificate? Evidence from the 2800 adjudicated heart protection study deaths*. 2015: Poster Presentation: International Clinical Trials Methodology Conference 2015.
183. Embleton, A., E. Clark, and S. Townsend, *Impact of retrospective data verification on the results of the academic led ICON6 trial*. 2015: Poster Presentation: International Clinical Trials Methodology Conference 2015.
184. Wright-Hughes, A., L. Graham, and A. Farrin, *Can the use of routine data enhance collection of the primary outcome in the SHIFT trial?* . 2013: Poster Presentation: Scottish Health Informatics Programme International Conference 2013.
185. Li, L. and P.M. Rothwell, *Biases in detection of apparent "weekend effect" on outcome with administrative coding data: Population based study of stroke*. *BMJ (Online)*, 2016. **353**.
186. Powell, G.A., et al., *Using routinely recorded data in the UK to assess outcomes in a randomised controlled trial: The Trials of Access*. *Trials*, 2017. **18**(1).
187. NHS Digital. *Secure File Transfer*. 2016 [cited 2016 18/07/2016]; Available from: <http://systems.hscic.gov.uk/infogov/security/infrasec/sft>.
188. University-of-Liverpool. *Research Data Management*. 2016 [cited 2016 30/11/2015]; Available from: (<http://www.liv.ac.uk/csd/research-data-management/storage>). .
189. NHEngland. *Data Dictionary*. 2016 [cited 2016 18/07/2016]; Available from: [http://www.datadictionary.nhs.uk/web\\_site\\_content/navigation/supporting\\_information\\_menu.asp?shownav=1](http://www.datadictionary.nhs.uk/web_site_content/navigation/supporting_information_menu.asp?shownav=1).
190. NHS Digital. *HES Data Dictionary*. 2016 [cited 2016 18/07/2016]; Available from: <http://www.hscic.gov.uk/hesdatadictionary>.

191. SAIL. *Data Dictionary*. 2016 [cited 2016 18/07/2016]; Available from: <http://www.datadictionary.wales.nhs.uk/default.htm?turl=WordDocuments%2Fnhswalesdatadictionaryversion30.htm>.
192. NHS Digital. *International Statistical Classification of Diseases and Related Health Problems (ICD) 10 and Office of Population and Census Surveys (OPCS) Version 4*. 2016 [cited 2016 18/07/2016]; Available from: <https://isd.hscic.gov.uk/trud3/user/authenticated/group/0/pack/37>.
193. NHS Digital. *NHS READ Codes Clinical Terms Version 3 Browser*. 2016 [cited 2016 18/07/2016]; Available from: <https://isd.hscic.gov.uk/trud3/user/authenticated/group/0/pack/9>.
194. IHDSTO. *Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT)*. 2016 [cited 2016 18/07/2016]; Available from: <http://browser.ihtsdotools.org/?perspective=full&conceptId1=404684003&edition=uk-edition&release=v20151001&server=https://browser-aws-1.ihtsdotools.org/api/snomed&langRefset=9000000000000508004>.
195. NHS Digital. *Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT)*. 2016 [cited 2016 18/07/2016]; Available from: <https://isd.hscic.gov.uk/trud3/user/authenticated/group/0/pack/26>
196. Altman, D., *Practical Statistics for Medical Research*. 1990: Chapman and Hall. 624.
197. Grainger, R., et al., *Referral patterns after a seizure admission in an English region: An opportunity for effective intervention? An observational study of routine hospital data*. *BMJ Open*, 2016. **6**(1).
198. Walker, P.P., et al., *Use of mortality within 30 days of a COPD hospitalisation as a measure of COPD care in UK hospitals*. *Thorax*, 2013. **68**(10): p. 968-970.
199. Shalihdi, M., et al., *Variation in gastroscopy rate in English general practice and outcome for oesophagogastric cancer: Retrospective analysis of Hospital Episode Statistics*. *Gut*, 2014. **63**(2): p. 250-261.
200. Abraham, K.A., et al., *Inequalities in outcomes of acute kidney injury in England*. *QJM*, 2012. **105**(8): p. 729-740.
201. Clinical-Study-Data-Request. *Clinical Study Data Request*. 2016 [cited 2016 05/04/2016]; Available from: <https://www.clinicalstudydatarequest.com>.
202. Wilson, E.C.F., et al., *Efficient Research Design: Using Value-of-Information Analysis to Estimate the Optimal Mix of Top-down and Bottom-up Costing Approaches in an Economic Evaluation alongside a Clinical Trial*. *Medical Decision Making*, 2016. **36**(3): p. 335-348.
203. NHS Digital. *Press Release: Patient Opt Out*. 2016 [cited 2016 01/05/2016]; Available from: <http://www.hscic.gov.uk/catalogue/PUB20527>.
204. Clarke, M., *Standardising outcomes for clinical trials and systematic reviews*. *Trials*, 2007. **8**.
205. Appleyard, S.E. and D.C. Gilbert, *Innovative Solutions for Clinical Trial Follow-up: Adding Value from Nationally Held UK Data*. *Clinical Oncology*, 2017.
206. HTMR, M. *Studies Within A Trial*. 2018 6/2/18]; Available from: <http://www.methodologyhubs.mrc.ac.uk/resources/swat>.
207. Ryan, M. and S. Farrar, *Using conjoint analysis to elicit preferences for health care*. *British Medical Journal*, 2000. **320**(7248): p. 1530-1533.
208. Ethier, J.F., et al., *eSource for clinical trials: Implementation and evaluation of a standards-based approach in a real world trial*. *International Journal of Medical Informatics*, 2017. **106**: p. 17-24.



## **Appendix A**

### **Chapter Two: Literature Review Search Strategies, PRISMA**

#### **Checklist and Further Results**

Table A.1: Included Studies; Characteristics and Use of UK Secondary Care Routinely Recorded Data in RCTs

Study Reference	Trial Summary	Outcome Measures	Implementation of Routinely Collected Data	Appraisal
<b>Office for National Statistics (ONS) (General Register Office (GRO), General Statistics Office of Ireland (GSOI))</b>				
Ashton et al: 2002 [119]  The Multicentre Aneurysm Screening Study (MASS) into the effect of abdominal aortic aneurysm (AAA) screening on mortality in men: a randomised controlled trial  Parallel Group RCT	67,800 participants randomised to Ultrasound (USS) abdominal screening or no intervention. Those with normal USS had no further intervention. Abnormal USS had follow up scans and vascular input where indicated.  Data such as investigation results and clinical follow up were reported through standard prospective methods.  Mortality was measured by accessing data from ONS.	Primary Outcome: Aneurysm related mortality.  Secondary Outcomes: All-cause mortality, frequency of ruptured aneurysm, quality of life	Participants were identified by NHS Number and ONS requested to provide a copy of the death certificates.  International Classification of Diseases (ICD) codes were used to identify relevant diagnoses such as 'ruptured aortic aneurysms'.  'Additional information' was sought from hospital / GP records if needed although the information requested and participant numbers where this additional information was sought is not explicit.	ONS Mortality follow-up was available for 67 274 (99%) of the randomised sample. Cost and resources required for access not reported.  Following review of all death certificates, 8% (14 of 177) of those certified as having died from a AAA were considered to have died from other causes and 0.1% (9 of 7407) of those certified as having died from other causes were considered to have died from ruptured AAA. The study examined the impact of these discrepancies on trial results and there was no significant difference - HR of 0.62 compared to 0.58.
Bale et al: 2008 [121]  Long-term mortality follow-up of the ISOLDE participants: causes of death during 13 years after trial completion  Parallel Group RCT	752 patients with moderate/severe Chronic Obstructive Pulmonary Disease (COPD) randomised to fluticasone or placebo for 3 years. Mortality at 3 years examined in the initial trial along with lung function and COPD exacerbations using standard prospective methods.  375 participants (inclusion dictated by ethical approval) were included in this long term follow up study.  Mortality was measured by accessing data from ONS.	Primary Outcome: All-cause mortality.	ONS requested to provide a copy of the death certificates. Participant identifiers used or data extracted not detailed.  Life status was also confirmed by the NHS strategic tracing service.	Data available for 206 (98%) participants included. Cost and resources required for access not reported.  No explanation of the methods for data retrieval or review.  Study reported 'limitations associated with ONS data'. No further details.

<p>Brown et al: 2012 [120]</p> <p>The UK EndoVascular Aneurysm Repair (EVAR) trials: randomised trials of EVAR versus standard therapy</p> <p>Parallel Group RCT</p>	<p>EVAR 1: EVAR vs Laparotomy, 1252 participants with AAA randomised to EVAR or laparotomy.</p> <p>EVAR 2: EVAR vs No Intervention, 404 participants with AAA randomised to EVAR or conservative management.</p> <p>Mean follow up 8 years. Data to inform clinical and cost effectiveness measures were reported through standard prospective methods.</p> <p>Mortality was measured by accessing data from ONS.</p>	<p>Primary Outcome: Mortality (all-cause, operative, aneurysm-related).</p>	<p>ONS requested to provide a copy of the death certificates. Participant identifiers used not detailed.</p> <p>ICD codes were obtained and reviewed by an endpoints committee. Included codes and timing of death were pre-specified in the protocol. No reported additional data sources accessed, including comparison to clinical sources of data to assess agreement.</p>	<p>Mortality data available for 99% of patients. Cost and resources required for access not reported.</p> <p>No explanation of the methods for data retrieval.</p> <p>Study reported the 'couple of months' delay between death and notification of death through ONS. For patients where notification of death was not received in the last few months of the RCT, letters were sent to participants. Where no response was received, phone calls were made to confirm life status. No further limitations detailed.</p>
<p>Henderson et al: 2015 [126]</p> <p>10-Year Mortality Outcome of a Routine Invasive Strategy Versus a Selective Invasive Strategy in Non-ST-Segment Elevation Acute Coronary Syndrome</p> <p>Parallel Group RCT</p>	<p>1810 patients with non-ST-segment elevation acute coronary syndrome were randomised to an early invasive strategy (coronary arteriography and myocardial revascularization) or a selective invasive strategy (coronary arteriography for recurrent ischemia only).</p> <p>Trial follow up for 5 years in previous paper assessed clinical measures and mortality (99.6% of patients) using standard prospective methods.</p> <p>10 year mortality was measured in this study by accessing data from ONS and the General Register Office (GRO) for Scotland.</p>	<p>Primary Outcome: All-cause mortality.</p> <p>Secondary Outcomes: Mortality (cardiovascular or non-cardiovascular).</p>	<p>ONS for England and GRO for Scotland requested to provide a copy of the death certificates. Participant identifiers used not detailed.</p> <p>ICD codes were reviewed to identify relevant diagnoses.</p>	<p>457 deaths identified at 10 years. Cost and resources required for access not reported.</p> <p>No explanation of the methods for data retrieval or review.</p> <p>Study states missing data is a possibility and completeness of data not reported - example provided of patients emigrating.</p>
<p>Molyneux et: 2015 [127]</p> <p>The durability of endovascular coiling versus neurosurgical clipping: 18 year follow-up of the UK cohort of the International Subarachnoid Aneurysm Trial (ISAT)</p> <p>Parallel Group RCT</p>	<p>2143 participants with ruptured cerebral aneurysm were randomised to neurosurgical clipping or endovascular coiling in the ISAT Trial.</p> <p>Clinical outcomes and mortality measured using standard prospective methods and reported in the initial ISAT publication.</p> <p>18 year mortality for 1624 participants was measured in this study by accessing data from ONS. Additional measures included functional status, assessed by questionnaire.</p>	<p>Primary Outcome: All-cause mortality.</p> <p>Secondary Outcomes: Functional status, dependency.</p>	<p>ONS requested to provide a copy of the death certificates to determine mortality status. Participant identifiers used or data extracted not detailed.</p>	<p>Follow up for 99% of cohort at 10 years, 338 patients (24%) died. Cost and resources required for access not reported.</p> <p>No explanation of the methods for data retrieval or review.</p> <p>No limitations relating to ONS data self-reported.</p>

<p>Perera et al: 2012 [128]</p> <p>Long-Term Mortality Data From the Balloon Pump-Assisted Coronary Intervention Study (BCIS-1) A Randomized, Controlled Trial of Elective Balloon Counter-pulsation During High-Risk Percutaneous Coronary Intervention (PCI)</p> <p>Parallel Group RCT</p>	<p>301 patients with severe ventricular impairment and coronary artery disease were randomised to Intra-Aortic Balloon Pump (IABP) during PCI, or PCI alone.</p> <p>Clinical outcomes including death, acute myocardial infarction, cerebrovascular event, or urgent further revascularization at hospital discharge and 6 month mortality were measured using standard prospective methods.</p> <p>This follow up study assessed long term all-cause mortality (median duration 51 months) through accessing ONS/GRO data.</p>	<p>Primary Outcome: All-cause mortality.</p>	<p>ONS requested to provide mortality status, cause of death not requested. Participant identifiers used or data extracted not detailed.</p>	<p>'100% data capture' reported, but no explanation as to how patients who had not died, were confirmed to be alive.</p> <p>No explanation of the methods for data retrieval or review.</p> <p>No limitations relating to ONS data self-reported.</p>
<p>Scholefield: 2012 [118]</p> <p>Nottingham trial of faecal occult blood (FOB) testing for colorectal cancer: a 20-year follow-up</p> <p>Parallel Group RCT</p>	<p>152 850 individuals by household were randomised to biennial FOB screening vs no intervention.</p> <p>Clinical outcomes including incidence of colorectal cancer, incidence and complications of interventions (colonoscopy) and mortality were measured using standard prospective methods.</p> <p>This follow up study assessed long term (median 19.5 years) mortality through access to ONS data.</p>	<p>Primary Outcomes: Mortality (all-cause, colorectal cancer related).</p>	<p>ONS requested to provide a copy of the death certificates to determine mortality status. Participant identifiers used or data extracted not detailed.</p>	<p>99% follow up based on ONS flagging. 875 (0.6%) participants could not be traced by ONS, suggests patients may have emigrated. Cost and resources required for access not reported.</p> <p>No explanation of the methods for data retrieval.</p> <p>Reported good agreement between 'verified' CRC deaths and 'certified' CRC deaths on the ONS data, assessed by comparison to case note review in the initial publication.</p>
<p>Simmons: 2012 [129]</p> <p>Screening for type 2 diabetes and population mortality over 10 years (ADDITION-Cambridge)</p> <p>Cluster RCT</p>	<p>20 185 participants in 33 GP's were randomised, by GP to Type 2 Diabetes Mellitus (T2DM) screening followed by intensive treatment for people diagnosed with diabetes (n=15); screening plus routine care of diabetes according to national guidelines (n=13); and no-screening control group (n=5).</p> <p>This study reports long term (median 9.6 years) mortality through access to ONS/GRO and Central Statistics Office of Ireland (CSOI) data.</p>	<p>Primary Outcome: All-cause mortality.</p> <p>Secondary Outcomes: Death from cardiovascular disease, cancer, DM related death.</p>	<p>'GP medical records' used to screen for patients at risk of T2DM. No further details provided on the method or primary care data source.</p> <p>Mortality data sought from ONS/GRO/CSOI. NHS Number provided to each source and copy of death certificate requested for review.</p> <p>ICD codes reviewed to classify cause of death.</p>	<p>99% follow up based on ONS flagging by NHS Number. The effect of missing data was examined in the study and deemed not to affect the result, but the data regarding this was not shown. Resources required for access not reported.</p> <p>Reported 'high level of agreement for cause of death'. However, this refers to the inter-rater agreement of ICD code classification, rather than accuracy of ONS data.</p>

<b>NHA Digital and Office of National Statistics (ONS)</b>				
<p>Turner et al: 2014 [85]</p> <p>Design and preliminary recruitment results of the Cluster randomised trial of PSA testing for Prostate cancer (CAP)</p> <p>Cluster RCT</p>	<p>785 GP's randomised to prostate specific antigen (PSA) screening vs standard care.</p> <p>This study reports the design and recruitment results.</p> <p>Clinical details including diagnoses and cause of death will be obtained through access to HES / ONS data. Follow up at 10 years due 2017.</p>	<p>Primary Outcome: 10 year 'definite' or 'probable' prostate cancer mortality.</p> <p>Secondary Outcomes: All-cause mortality (10 and 15 years), cost effectiveness.</p>	<p>All participants prospectively flagged with NHS Digital/ONS.</p> <p>Clinical details will be reviewed in addition to death certificate data through access to HES/ONS.</p> <p>Non-identifiable ONS data was also used to determine death rates and inform the sample size calculation.</p>	<p>All participants (intervention and control) were flagged with NHS Digital prospectively.</p> <p>Anticipated inaccuracies with coding / classification of cause of death will be addressed with committee review and consensus of the available documentation - it is not clear if this will involve comparison to clinical / source data.</p>

Table A.2: Included Studies; Characteristics and Use of UK Primary Care Routinely Recorded Data in RCTs

Study Reference	Trial Summary	Outcome Measures	Implementation of Routinely Collected Data	Appraisal
<b><i>The Secure Anonymised Information Linkage Databank (SAIL)</i></b>				
Brooks et al: 2009 [122]  Use of a patient linked data warehouse to facilitate diabetes trial recruitment from primary care  RCT Recruitment Feasibility Assessment	SAIL Databank was used as the data source to perform a recruitment feasibility assessment for two fictitious RCTs. Both RCTs involved diabetes mellitus with pragmatic inclusion criteria such as diagnosis of diabetes, prescription of specified medicines, BMI and smoking status.  Of 250,086 individuals in SAIL, 284 were identified for the first RCT and 711 for the second.	N/A	Patients whose details were anonymously recorded in SAIL Databank were searched using specific inclusion and exclusion criteria. The total numbers of eligible patients in the participating GP's and their location by GP were then determined.	SAIL represents a useful data source to prospectively perform a recruitment feasibility assessment for selected RCTs.  Pragmatic RCTs with simple inclusion criteria are likely most appropriate for such a feasibility assessment.  SAIL Databank cannot routinely re-identify patients and therefore the transition to GP recruitment when the RCT opens is limited.
McGregor et al: 2010 [62]  The Health Informatics Trial Enhancement Project (HITE): Using routinely collected primary care data to identify potential participants for a depression trial  RCT Recruitment Feasibility Assessment	SAIL Databank was used as the data source to perform a recruitment feasibility assessment for an existing RCT assessing folate use in patients with depression. The existing inclusion/exclusion criteria were translated into READ codes and a database query was run using Structured Query Language (SQL) within SAIL for 5 GP's. The inclusion/exclusion criteria are pragmatic and include variable such as prescribed drugs, investigation results, previous diagnoses.  867 potential participants were identified.	N/A	Patients whose details were anonymously recorded in SAIL Databank were searched using specific inclusion and exclusion criteria. The total numbers of eligible patients in the participating GP's and their location by GP were then determined.	SAIL represents a useful data source to prospectively perform a recruitment feasibility assessment for selected RCTs.  Pragmatic RCTs with simple inclusion criteria are likely most appropriate for such a feasibility assessment.  In this study the sensitivity/specificity (>96%) of the SQL query was determined by manual review of the electronic medical record.  SAIL Databank cannot routinely re-identify patients and therefore the transition to GP recruitment when the RCT opens is limited.

<b>The Clinical Practice Research Datalink (CPRD)</b>				
<p>Dregan et al: 2014 [123]</p> <p>Point-of-Care Cluster Randomized Trial in Stroke Secondary Prevention Using Electronic Health Records</p> <p>Cluster RCT</p>	<p>106 participating GP's contributing data to CPRD were allocated to intervention or control study arms. The intervention was installation of IT decision support tools to improve adherence to secondary care stroke prevention measures during the patient consultation. Control was standard practice.</p> <p>RCT duration 12 months. Pragmatic clinical details including BP and blood tests were recorded through CPRD to measure the study outcomes.</p>	<p>Primary Outcome: Systolic blood pressure (BP).</p> <p>Secondary Outcomes: Diastolic BP, total cholesterol, prescription of cardiovascular drugs.</p>	<p>GP's contributing data to CPRD were invited to participate. Practices in the intervention arm received IT software that flagged up during the consultation of a suitable patient. Patient consent was not required.</p> <p>Pragmatic, commonly recorded clinical parameters were selected as outcome measures and data was collected solely through CPRD.</p>	<p>BP data was available for 90% and cholesterol data for 84%.</p> <p>Data for 12 month follow up were recorded at 15 months, to account for the delay in CPRD data becoming available. Baseline data regarding the cause of stroke (ischaemic, haemorrhagic) was limited with 60% classed as 'undefined'.</p>
<p>Guilliford et al: 2014 [124]</p> <p>Electronic Health Records for Intervention Research: A Cluster Randomized Trial to Reduce Antibiotic Prescribing in Primary Care (eCRT Study)</p> <p>Cluster RCT</p>	<p>104 participating GP's contributing data to CPRD were allocated to intervention or control study arms. The intervention was installation of decision support tools to improve adherence to antibiotic prescribing guidelines during the patient consultation for respiratory tract infection (RTI).</p> <p>RCT duration 12 months. Pragmatic clinical details including prescription of antibiotics and record of respiratory diagnoses were recorded through CPRD to measure the study outcomes.</p>	<p>Primary Outcome: Proportion of consultations for RTI with antibiotics prescribed.</p> <p>Secondary Outcomes: Proportion of antibiotics prescribed in other respiratory/ENT infective diagnoses.</p>	<p>GP's contributing data to CPRD were invited to participate. Practices in the intervention arm received IT software that flagged up during the consultation of a suitable patient. Patient consent was not required.</p> <p>Pragmatic, commonly recorded clinical parameters were selected as outcome measures and data was collected solely through CPRD.</p>	<p>Reported feasibility in performing large scale public health RCTs in a routine primary care database.</p> <p>No mention of missing data – the proportion of inaccurate prescription records was not assessed.</p>
<p>Horspool et al: 2013 [125]</p> <p>Preventing and lessening exacerbations of asthma in school-age children associated with a new term (PLEASANT): study protocol for a cluster randomised control trial</p> <p>Cluster RCT Protocol</p>	<p>Protocol for a cluster RCT involving 140 GP's contributing data to CPRD.</p> <p>GP's will be recruited, half to usual care, half to the intervention - a letter informing parents of the importance of adherence to their child's asthma treatment throughout the summer holidays.</p> <p>The pragmatic clinical outcomes of unscheduled medical contact, prescriptions and respiratory diagnoses will be recorded through CPRD to measure the study outcomes.</p>	<p>Primary Outcome: Unscheduled medical contact in September.</p> <p>Secondary Outcomes: Unscheduled medical contacts at other time points associated with prescriptions, respiratory diagnoses.</p>	<p>GP's contributing data to CPRD will be invited to participate. Practices in the intervention arm will receive a standard letter to be sent to eligible patients. Patient consent will not be required.</p> <p>Pragmatic outcomes that should be reliably recorded have been selected and data will be collected solely through CPRD. Committee review of coded data is proposed.</p>	<p>N/A</p>

<b><i>The Clinical Practice Research Datalink (CPRD) and ResearchOne</i></b>				
Herrett et al: 2014 [92]  Text messaging reminders for influenza vaccine in primary care: protocol for a cluster randomised controlled trial  Cluster RCT Protocol	<p>Protocol for a cluster RCT involving GP's contributing data to CPRD or ResearchOne.</p> <p>GP's will be randomised to either standard care or text messaging campaign to increase uptake of the flu vaccine. Unknown planned duration / completion date.</p> <p>The pragmatic clinical outcome of flu vaccine administration will be recorded through CPRD and ResearchOne to measure the study outcome.</p>	Primary Outcome: Flu vaccine administration.	<p>GP's contributing data to CPRD and ResearchOne will be invited to participate. Practices in the intervention arm will receive software to enable text messages to be sent to eligible patients. Patient consent will not be required.</p> <p>A pragmatic outcome that should be reliably recorded has been selected as the outcome measure and data will be collected solely through CPRD and ResearchOne.</p>	N/A

Table A.3: PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
<b>TITLE</b>			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	Title Page
<b>ABSTRACT</b>			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	Abstract
<b>INTRODUCTION</b>			
Rationale	3	Describe the rationale for the review in the context of what is already known.	Introduction
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	Introduction
<b>METHODS</b>			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Methods The review was not eligible for registration in the PROSPERO database.
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	Methods
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	Methods
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Methods Appendix II
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Methods
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	Methods
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Methods

Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	Methods A narrative appraisal was performed. Formal assessment was performed where routinely recorded data was implemented in the study and there was potential for the introduction of bias.
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	Methods
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I <sup>2</sup> ) for each meta-analysis.	Methods

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	Methods A narrative appraisal has been performed. Formal assessment has not been performed in view of the objective of the review.
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/a
<b>RESULTS</b>			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Results
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Results
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Results

Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Results
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/a
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	N/a
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/a
<b>DISCUSSION</b>			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	Discussion
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	Discussion
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	Discussion
<b>FUNDING</b>			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	To be included in journal submission.

Table A.4: Search Strategy: Medline (OVID)

1	Administrative Data Research Network.tw.
2	Clinical Practi\$ Research Datalink.tw.
3	((Driv\$ adj2 Vehicle Licen?ing Agency) or (Driv\$ adj2 Vehicle Licen?ing Authority)).tw.
4	(Department adj2 "Work and Pensions").tw.
5	General Practi\$ Extraction Service.tw.
6	(General Practi\$ Research Database or General Practi\$ Registry Database).tw.
7	Hospital Episode Statistics.tw.
8	"Revenue and Customs".tw.
9	"Health and Social Care Information Centre".tw.
10	NorthWest eHealth.tw.
11	Office for National Statistics.tw.
12	Secure Anonymised Information Linkage Databank.tw.
13	(NHS Wales Informatics Service or Patient Episode Database for Wales).tw.
14	"The Health Improvement Network".tw.
15	QResearch.tw.
16	ResearchOne.tw.
17	Information Services Division.tw.
18	or/1-17
19	(ADRN or CPRD or DVLA or DWP or GPES or GPRD or HES or HMRC or HSCIC or NWEH or ONS or SAIL or PEDW).tw.
20	("dose-width product" or Dose width product).tw.
21	General Practice Education\$ Supervi\$.tw.
22	(Geriatric Psychiatry Research Division or Aspergillus or gastropharyngeal reflux disease\$ or (gene\$ and prion disease\$)).tw.
23	(balanced starch\$ or hydroxyethyl starch\$ or hydroxyethylstarch\$ or hydroxy ethyl starch\$ or Hospital Eye Service\$ or hip extensor stretch).tw.
24	("Add-ons" or Oral nutri\$ supplement\$ or Occipital nerve stimul\$.tw.
25	(Sheffield Assessment for Letters or Sheffield Assessment Instrument for Letters or Stimulation Assistance through Iterative Learning).tw.
26	or/20-25
27	19 not 26
28	((routine or clinic\$ or primary or general practic\$ or general practis\$ or general practitioner\$ or GP?) and data and source?).ti.
29	((routine or clinic\$ or primary or general practic\$ or general practis\$ or general practitioner\$ or GP?) adj2 data).ab.
30	data source?.ab. /freq=2
31	29 and 30
32	27 or 28 or 31
33	exp Great Britain/
34	(national health service* or nhs*).ti,ab,in.
35	(english not ((published or publication* or translat* or written or language* or speak* or literature or citation*) adj5 english)).ti,ab.
36	(gb or "g.b." or britain* or (british* not "british columbia") or uk or "u.k." or united kingdom* or (england* not "new england") or northern ireland* or northern irish* or scotland* or scottish* or ((wales or "south wales") not "new south wales") or welsh*).ti,ab,jw,in.

37	(bath or "bath's" or ((birmingham not alabama*) or ("birmingham's" not alabama*) or bradford or "bradford's" or brighton or "brighton's" or bristol or "bristol's" or carlisle* or "carlisle's" or (cambridge not (massachusetts* or boston* or harvard*)) or ("cambridge's" not (massachusetts* or boston* or harvard*)) or (canterbury not zealand*) or ("canterbury's" not zealand*) or chelmsford or "chelmsford's" or chester or "chester's" or chichester or "chichester's" or coventry or "coventry's" or derby or "derby's" or (durham not (carolina* or nc)) or ("durham's" not (carolina* or nc)) or ely or "ely's" or exeter or "exeter's" or gloucester or "gloucester's" or hereford or "hereford's" or hull or "hull's" or lancaster or "lancaster's" or leeds* or leicester or "leicester's" or (lincoln not nebraska*) or ("lincoln's" not nebraska*) or (liverpool not (new south wales* or nsw)) or ("liverpool's" not (new south wales* or nsw)) or ((london not (ontario* or ont or toronto*)) or ("london's" not (ontario* or ont or toronto*)) or manchester or "manchester's" or (newcastle not (new south wales* or nsw)) or ("newcastle's" not (new south wales* or nsw)) or norwich or "norwich's" or nottingham or "nottingham's" or oxford or "oxford's" or peterborough or "peterborough's" or plymouth or "plymouth's" or portsmouth or "portsmouth's" or preston or "preston's" or ripon or "ripon's" or salford or "salford's" or salisbury or "salisbury's" or sheffield or "sheffield's" or southampton or "southampton's" or st albans or stoke or "stoke's" or sunderland or "sunderland's" or truro or "truro's" or wakefield or "wakefield's" or wells or westminster or "westminster's" or winchester or "winchester's" or wolverhampton or "wolverhampton's" or (worchester not (massachusetts* or boston* or harvard*)) or ("worchester's" not (massachusetts* or boston* or harvard*)) or (york not ("new york*" or ny or ontario* or ont or toronto*)) or ("york's" not ("new york*" or ny or ontario* or ont or toronto*))))).ti,ab,in.
38	(bangor or "bangor's" or cardiff or "cardiff's" or newport or "newport's" or st asaph or "st asaph's" or st davids or swansea or "swansea's").ti,ab,in.
39	(aberdeen or "aberdeen's" or dundee or "dundee's" or edinburgh or "edinburgh's" or glasgow or "glasgow's" or inverness or (perth not australia*) or ("perth's" not australia*) or stirling or "stirling's").ti,ab,in.
40	(armagh or "armagh's" or belfast or "belfast's" or lisburn or "lisburn's" or londonderry or "londonderry's" or derry or "derry's" or newry or "newry's").ti,ab,in.
41	or/33-40
42	(exp africa/ or exp americas/ or exp antarctic regions/ or exp arctic regions/ or exp asia/ or exp australia/ or exp oceania/) not (exp great britain/ or europe/)
43	41 not 42
44	32 and 43
45	18 or 44
46	randomized controlled trial.pt.
47	controlled clinical trial.pt.
48	randomized.ab.
49	placebo.ab.
50	clinical trials as topic.sh.
51	randomly.ab.
52	trial.ti.
53	Or/46-52
54	exp animals/ not humans.sh.
55	53 not 54
56	45 and 55
57	remove duplicates from 56

i. 33-43 NICE UK Search Filter

ii. 46-55 Cochrane RCT Highly Sensitive Search Strategy



## **Appendix B**

### **Chapter Three: Systematic Review Search Strategies, PRISMA**

#### **Checklist and Further Results**

Table B.1: Included Studies: Agreement between Routinely Recorded Clinical Data and Data Collected through Standard Prospective Methods

Study Reference	Study Summary	Data Sources	Assessment of Agreement	Appraisal
<p>Barry et al: 2013 [65]</p> <p>Are Routinely Collected NHS Administrative Records Suitable for Endpoint Identification in Clinical Trials? Evidence from the West of Scotland Coronary Prevention Study</p> <p>RCT</p>	<p>6595 patients recruited into a RCT were randomised to pravastatin or placebo. Cardiovascular clinical outcomes and mortality were measured using standard RCT follow up.</p> <p>Participants were linked by the Information Services Division of the NHS National Services Scotland and RCT outcomes were identified from mortality and secondary care medical records and compared.</p>	<p>Standard Prospective: RCT clinical follow up</p> <p>Routinely Recorded: Scottish Morbidity Record Secondary care medical records</p>	<p>The primary outcomes were broadly comparable. Cardiovascular (CV) death or myocardial infarction (MI) in the placebo compared to pravastatin group was 212 vs 147 (P&lt;0.001, RR 32 (16,45)) in the RCT dataset and 195 vs 121 (P&lt;0.001, RR 39 (24,51)) in the routine dataset.</p> <p>241 deaths in the RCT dataset were matched with 240 in the routine dataset.</p> <p>217/268 (81%) of the first events recorded in the RCT matched first recorded routine events. 217/230 (94.3%) of the first recorded routine events matched first recorded RCT events.</p> <p>Agreement was reduced for events following the first recorded event and secondary outcomes, such as the diagnosis of stroke 44/61 (72.1%).</p>	<p>There is no statistical assessment of agreement between patient completed questionnaires and routinely recorded data despite the data being comparable.</p> <p>There is good agreement between mortality and the primary outcome of major cardiovascular events.</p>
<p>Britton et al: 2012 [178]</p> <p>Validating self-reported strokes in a longitudinal UK cohort study (Whitehall II): Extracting information from hospital medical records versus the Hospital Episode Statistics database</p> <p>Cohort Study</p>	<p>10,308 patients recruited into the Whitehall cohort study completed self-report health questionnaires. Between 2002-2009 self-reported episodes of stroke were identified.</p> <p>Episodes of stroke were identified from the Hospital Episode Statistics Inpatient Dataset (HES IP), primary and secondary care medical records and compared to patient reported events.</p>	<p>Standard Prospective: Patient completed questionnaire - Interval not reported</p> <p>Routinely Recorded: Hospital Episode Statistics Inpatient Dataset (HES IP) Primary and secondary care medical records</p>	<p>106 episodes of stroke were self-reported.</p> <p>8 (7.5%) self-reported strokes were recorded as 'false positives' where primary or secondary care medical records provided evidence the event was not a stroke. 4 self-reported strokes had no evidence in routinely recorded data.</p> <p>66 (62.3%) of self-reported strokes were validated in HES data. 16 (15.1%) were validated by hospital records alone and were not recorded in HES. 11 (10.4%) were recorded in HES alone. 12 (11.3%) were validated by GP only.</p> <p>47 episodes of stroke were recorded in HES but not self-reported.</p>	<p>There is no statistical assessment of agreement between patient completed questionnaires and routinely recorded data despite the data being comparable.</p> <p>There are discrepancies between self-reported episodes of stroke and those routinely recorded. However, discrepancies are minimised if multiple sources of routinely recorded data are accessed, although this approach has resource limitations.</p>

<p>Bryant et al: 2015 [149]</p> <p>Agreement between routine and research measurement of infant height and weight</p> <p>Cohort Study</p>	<p>A subgroup of patients recruited into the Born in Bradford cohort study had infant height and weight recorded in the study at 6, 12, 18, 24 and 36 months.</p> <p>Height and weight, respectively, was measured routinely in the Personal Child Health Record at 6 months (n=158 and 560), 12 months (n=101 and 166) and 24 months (n=307 and 434). Measurements at 18 and 36 months were excluded due to small sample size. Agreement was determined using mean differences and Bland Altman Plots.</p>	<p>Standard Prospective: Cohort study follow up</p> <p>Routinely Recorded: Personal Child Health Record</p>	<p>The mean difference in age was &lt;1 month for all assessments.</p> <p>There was agreement, greater for weight than height:</p> <p>Routinely recorded height was underestimated at 6 months (0.46 (-3.99 to 4.91)) and overestimated at 12 (-0.25 (-4.50 to 4.00)) and 24 (-0.32 (-4.00 to 3.36)) months.</p> <p>Routinely recorded weight was overestimated at 6 (-0.04 (-0.67 to 0.59)), 12 (-0.06 (-1.10 to 0.98)) and 24 (-0.14 (-1.19 to 0.91)) months.</p>	<p>There is acceptable agreement between research and routinely recorded height and weight measurements although wide limits of agreement are noted.</p> <p>For limited numbers of individuals there were marked discrepancies. For height, the differences ranged from -0.4-4.91 cm and for weight -1.19-0.98 kg.</p>
<p>Cleland et al: 2007 [150]</p> <p>An exploratory, pragmatic, cluster randomised trial of practice nurse training in the use of asthma action plans</p> <p>Cluster RCT</p>	<p>629 patients with asthma were randomised to receive standard care or enhanced care from a practice nurse who had received advanced asthma training.</p> <p>Prescriptions of relevant medications were obtained from primary care medical records for all (629) patients to measure asthma control. Additionally, the Asthma Control Questionnaire was completed by 236 patients to provide an indirect measure of asthma control.</p>	<p>Standard Prospective: Asthma Control Questionnaire (ACQ)</p> <p>- Baseline, 6 months</p> <p>Routinely Recorded: Primary care medical records</p>	<p>The ACQ provides a summary statistic of asthma control and is therefore not directly comparable to the prescription of medications.</p> <p>Between intervention and control groups there was no difference between ACQ results (Intervention: 2.66 (1.92-3.67), Control: 2.50 (1.67-3.67), P = 0.27). There was similarly no difference between prescription rates obtained from medical records (Intervention: 19 (12-32), Control: 18 (13-30), P = 0.51).</p>	<p>There is no statistical assessment of agreement between ACQ and primary care medical records and the measures are not directly comparable. However, broadly the results were similar and both identified no significant difference between intervention and control groups.</p> <p>The study also reports issues with missing routinely recorded data, despite the existence of relevant clinical codes.</p>

<p>Doshi et al: 2008 [152]</p> <p>Post-tonsillectomy morbidity statistics: are they underestimated?</p> <p>Cohort Study</p>	<p>92 consecutive patients undergoing tonsillectomies in a single centre were invited to complete a questionnaire.</p> <p>Post-operative complication rates over a 30 day period were retrieved from secondary care medical records and compared to patient reported complication rates assessed through questionnaire completion.</p>	<p>Standard Prospective: Patient completed questionnaire - 1 month</p> <p>Routinely Recorded: Secondary care medical records</p>	<p>70 of 92 patients returned the questionnaire. 11/70 patients reported post-operative bleeding, just 4 patients were identified in secondary care medical records (Rate 15.7% versus 5.7%). The remaining 7 patients contacted their GP and an additional 15 contacted their GP for pain control.</p>	<p>There is no statistical assessment of agreement between patient completed questionnaires and routinely recorded data despite the data being comparable.</p> <p>There is marked discrepancy between patient reported and secondary care routinely recorded data. However, all patients not contacting hospital had contacted their GP and therefore would likely be identified if primary in addition to secondary care medical records were accessed.</p>
<p>Hutchings et al: 2005 [154]</p> <p>Can electronic routine data act as a surrogate for patient-assessed outcome measures?</p> <p>RCT</p>	<p>93 patients diagnosed with inflammatory bowel disease recruited into a RCT completed the UK Inflammatory Bowel Disease Questionnaire and Short Form 36.</p> <p>Data retrieved from primary and secondary care electronic medical records were used to complete the questionnaires and results were compared to the patient-reported responses.</p>	<p>Standard Prospective: Patient completed UK Inflammatory Bowel Disease Questionnaire (UKIBD) and Short Form 36 (SF36) - Interval not reported</p> <p>Routinely Recorded: Primary and secondary care medical records</p>	<p>UKIBD: 29/30 questions had a comparable READ (CTV3) code and 20/30 an ICD 10 code. Symptoms were only recorded for 10 questions. Routinely recorded data resulted in higher UKIBDQ subscales and total scores. Intraclass correlation coefficients were poor (0.04-0.22).</p> <p>SF36: 34/36 questions had a comparable READ (CTV3) code and 30/36 an ICD10 code. Symptoms were only recorded for 10 questions. Routinely recorded data resulted in higher SF36 subscales. Intraclass correlation coefficients were poor (0.00 to 0.27).</p>	<p>The ICD10 codes, more commonly used in secondary care, do not translate as well as READ (CTV3) codes to commonly used existing patient-reported outcomes measures.</p> <p>Broadly, the patient-reported questions can in theory be identified by existing coding systems, but just 10/30 UKIBD and 10/36 SF36 codes were ever used in the medical records.</p> <p>In this study, the under-utilisation of codes, rather than the lack of available codes resulted in the poor correlation.</p>

<p>Iyer et al: 2013 [155]</p> <p>Patient-reporting improves estimates of postoperative complication rates: a prospective cohort study in gynaecological oncology</p> <p>Cohort Study</p>	<p>2152 patients undergoing gynaecological surgery were invited to complete an annual follow up questionnaire to ascertain post-operative complications.</p> <p>1462 patient completed questionnaire responses were compared to routinely recorded data retrieved from secondary care medical records.</p>	<p>Standard Prospective: Patient completed questionnaire - 1 year</p> <p>Routinely Recorded: Secondary care medical records</p>	<p>In total, 452 surgical Grade II-V complications were reported.</p> <p>Grade II: 158/280 (56.4%) patient reported complications were recorded in medical records.</p> <p>Grade III-V: 36/36 (100%) patient reported complications were recorded in medical records.</p> <p>The post-operative complication rate using data retrieved from medical records is 11.8% (172/1462; 95% CI 11–14) and using patient completed questionnaires 15.8% (231/1462; 95% CI 14–17.8).</p>	<p>There is a descriptive assessment of agreement between patient completed questionnaire and medical records.</p> <p>For more serious complications, frequently requiring hospital treatment such as further surgery, there was 100% concordance. For Grade II (mild) complications, the concordance rate was only 56.4%. This may be explained by the mild nature of symptoms and the authors suggest relevant details may be recorded in the routinely recorded primary care medical records.</p>
<p>Kingston et al: 2010 [153]</p> <p>Assessing the Amount of Unscheduled Screening (“Contamination”) in the Control Arm of the UK “Age” Trial</p> <p>RCT</p>	<p>3706 patients enrolled in the control arm (‘standard care’) of a RCT assessing annual mammographic screening were sent questionnaires to assess the occurrence of unrecorded mammography.</p> <p>2115 patients returned completed questionnaires and data were compared to 946 responses to the routinely recorded ONS Omnibus Surveys.</p>	<p>Standard Prospective: Patient completed questionnaire - 1 year</p> <p>Routinely Recorded: ONS Omnibus Surveys</p>	<p>2115 patients enrolled in the control arm of the RCT returned questionnaires. A total of 24.9% (95% confidence interval, 23.0-26.8) reported having a mammogram. A different cohort of 223 patients (23.6%) of similar age reported having a mammogram over the same time period in the ONS Omnibus Surveys.</p>	<p>There is no statistical assessment of agreement between patient completed questionnaires and the routinely recorded ONS Omnibus Surveys despite the data being comparable.</p> <p>The methodology involves patient completed questionnaires either during a RCT or as part of a routine population survey. The results are expectedly comparable although the ONS Omnibus Surveys would not be a ‘typical’ source of routinely recorded data.</p>

<p>Lewsey et al: 2000 [46]</p> <p>Using routine data to complement and enhance the results of randomised controlled trials</p> <p>Cohort Study</p>	<p>Report involving a meta-analysis of 8 RCT's examining coronary artery bypass graft (CABG) vs percutaneous transluminal coronary angioplasty (PTCA). Data were compared to routinely recorded data retrieved from Scottish Morbidity Records and secondary care medical records.</p> <p>3371 patients enrolled in RCTs were compared with routinely recorded data regarding 12,238 patients.</p>	<p>Standard Prospective: RCT clinical follow up</p> <p>Routinely Recorded: Scottish Morbidity Record Secondary care medical records</p>	<p>Baseline characteristics are broadly comparable between RCT and routinely recorded data.</p> <p>The primary outcome of the meta-analysis was myocardial infarction or cardiac death. The RR of PTCA compared to CABG was 1.03 (95% CI 0.84 to 1.27). The RR estimated from routinely recorded data was 1.15 (95% CI 0.90 to 1.48).</p>	<p>There is no statistical assessment of agreement between RCT data and routinely recorded data despite data being comparable.</p> <p>Acknowledging the prospective access to medical records during the RCT, the primary outcome is comparable to that calculated from routinely recorded data.</p>
<p>Mitchell et al: 2016 [166]</p> <p>Is there a difference between hospital verified and self-reported self-harm? Implications for repetition</p> <p>Cohort Study</p>	<p>774 patients presenting to a single emergency department with self harm were enrolled in a cohort study. Previous episodes of self harm were recorded based on patient self-report.</p> <p>Self-reported self-harm was compared with self harm episodes verified in secondary care medical records.</p>	<p>Standard Prospective: Patient completed questionnaire</p> <p>Routinely Recorded: Secondary care medical records</p>	<p>774 patients were enrolled in the study.</p> <p>432 patients had no previous self-harm evident on medical records, but only 134 (31%) had no previous episodes on self-report.</p> <p>340 patients had previous self-harm evident on medical records but only 113 (33.2%) had previous self-harm on self-report.</p> <p>Cohen's Kappa agreement was low (Kappa=0.353, CI 0.287–0.419, S.E. of kappa=0.034 P=not significant).</p>	<p>There is a descriptive and statistical assessment of agreement.</p> <p>There is poor agreement with patients underreporting both the occurrence and absence of previous episodes of self harm.</p>
<p>Pastorino et al: 2015 [180]</p> <p>Validation of self-reported diagnosis of diabetes in the 1946 British birth cohort</p> <p>Cohort Study</p>	<p>230 patients recruited into the Medical Research Council National Survey of Health and Development Study self-reported a diagnosis of diabetes on questionnaire.</p> <p>Self-reported information regarding diagnosis and treatment of diabetes was compared with primary care medical records.</p>	<p>Standard Prospective: Patient completed questionnaire</p> <p>Routinely Recorded: Primary care medical records</p>	<p>230 patients reported a diagnosis of diabetes, 184 were reviewed at the most recent study follow up and 172 provided consent to request primary care medical records.</p> <p>Primary care medical records were available for 157 patients. 149 self-reported diagnoses were confirmed (PPV: 94.9%). Of the 8 non-confirmed, 2 patients had 'pre-diabetes' (impaired fasting glucose).</p> <p>The mean difference in age at diagnosis is 0.6 years (95% CI: 0.2-1.1), with patients self-reporting earlier.</p>	<p>There is a descriptive and statistical assessment of agreement.</p> <p>There is good agreement between self-reported and recorded diagnosis of diabetes. However, patients significantly over estimated the duration of their disease.</p>

<p>Steward et al: 1993 [156]</p> <p>Chemotherapy Administration and Data Collection in an EORTC Collaborative Group Can we Trust the Results?</p> <p>RCT</p>	<p>78 (of a total 111) patients from 14 centres were enrolled in a RCT assessing chemotherapy for sarcoma.</p> <p>Information recorded prospectively in Case Report Forms was compared to local secondary care medical records.</p>	<p>Standard Prospective: RCT clinical follow up</p> <p>Routinely Recorded: Secondary care medical records</p>	<p>8776 data items entered on CRFs were compared to local secondary care medical records.</p> <p>&lt;1% of data on CRFs was missing. Agreement, determined by the percentage of data items entered 'incorrectly' on the CRF was &lt;3% in 9 centres and between 4-7.5% in the remaining 6 centres. Drug dosages and patient performance status were the commonly incorrectly entered data items. However, in 7/14 centres &lt;80% of data recorded on the CRF could be verified in medical records.</p>	<p>There is no statistical assessment of agreement between RCT data and routinely recorded data despite data being comparable.</p> <p>The percentage of data transcribed 'incorrectly' was relatively modest.</p> <p>However, a greater percentage of data recorded in the CRF could not be verified in the medical records. This may be explained as certain data items may not normally be routinely recorded in the medical records, although this is not discussed.</p>
<p>Tannen et al: 2006 [144]</p> <p>Simulation of the Syst-Eur randomized control trial using a primary care electronic medical record was feasible</p> <p>Simulated RCT</p>	<p>4695 patients were enrolled in a RCT assessing anti-hypertensives vs placebo in patients with systolic hypertension.</p> <p>16,771 patients identified from the General Practice Research Database were identified that met the RCT inclusion criteria and were included in a simulated RCT, replicating all stages of the original RCT with the exception of randomisation.</p>	<p>Standard Prospective: RCT clinical follow up</p> <p>Routinely Recorded: General Practice Research Database (GPRD)</p>	<p>2815 'exposed' and 13,956 'unexposed' patients were identified in GPRD (compared to 2398 and 2297 in the RCT). Baseline characteristics were comparable. The average baseline BP readings were similar, however, the baseline BP in the 'exposed' GPRD cohort was raised (Exposed: 176.2, Unexposed: 171.9) compared to the RCT (173.8, 173.9). The difference was significant (P&lt;0.001).</p> <p>The incidence rates for clinical measures were comparable between RCT and GPRD, except for incidence of heart failure and transient ischemic attack, that were higher (P&lt;0.001) and cardiovascular death that was lower (P&lt;0.001) in the 'unexposed' GPRD than in the RCT placebo.</p> <p>For clinical outcomes statistically significant differences between GPRD and RCT data were observed for 2 of the total 12 outcomes including death (GPRD IRR: 1.23, 95% CI: 1.00–1.50, RCT IRR: 0.86, 95% CI: 0.67–1.09, P=0.03) and peripheral vascular disease (GPRD IRR: 1.13, 95% CI: 0.85–1.51, RCT IRR: 0.68, 95% CI: 0.46–1.02, P=0.04).</p>	<p>Baseline characteristics were comparable with the exception of baseline BP – likely explained by the greater probability of patients with higher BP readings being commenced on treatment.</p> <p>There were some statistically significant differences in the incidence rates of clinical measures and 2 of 12 clinical outcomes. However, broadly the results of the GPRD study and RCT were comparable.</p>

<p>Tannen et al: 2007 [147]</p> <p>A simulation using data from a primary care practice database closely replicated the women's health initiative trial</p> <p>Simulated RCT</p>	<p>17,407 patients were enrolled in a RCT assessing the risks and benefits of oestrogen plus progestin hormone replacement therapy vs placebo in postmenopausal women.</p> <p>51,388 patients identified from the General Practice Research Database were identified that met the RCT inclusion criteria and were included in a simulated RCT, replicating all stages of the original RCT with the exception of randomisation.</p>	<p>Standard Prospective: RCT clinical follow up</p> <p>Routinely Recorded: General Practice Research Database (GPRD)</p>	<p>13,658 'exposed' and 37,730 'unexposed' patients were identified in GPRD (compared to 8506 and 8902 in the RCT). Baseline characteristics were broadly comparable. However, patients in the GPRD study were younger, less obese, more were current smokers and had a lower incidence of diabetes, hypercholesterolemia and hypertension. Prior use of HRT was greater in the GPRD exposed group than the RCT treated group, lower in the GPRD Unexposed than RCT placebo. In the RCT baseline characteristics between treated and placebo groups were identical. Between exposed and unexposed groups in GPRD, significant differences were observed in BMI, diabetes, hypertension, smoking status, prior cardiovascular disease and prior HRT use (P&lt;0.01).</p> <p>The incidence rates for clinical measures were broadly comparable with the exception of GPRD adjusted hazard ratios for cancer (increased) and death (decreased) which were not significant in the RCT. In addition, myocardial infarction was significantly increased in the RCT and unchanged in GPRD.</p> <p>For clinical outcomes statistically significant differences between GPRD and RCT data were observed in the ITT HR's (95% CI) for 4 of the total 10 outcomes including myocardial infarction (RCT: 1.32 1.02-1.72, GPRD: 0.79 0.65-0.95, P=0.05), pulmonary embolus (RCT: 2.13 1.39-3.25, GPRD: 1.26 0.90-1.76, P=0.04), breast cancer (RCT: 1.26 1.00-1.59, GPRD: 1.67 1.45-1.93, P=0.05) and death (RCT: 0.98 0.82-1.18, GPRD: 0.64 0.57-0.73, P=0.02).</p>	<p>Baseline characteristics differ in some variables and are potentially important to the agreement between study results. For example the RCT inclusion criteria could not be rigidly applied with regards to prior HRT use, medication and HRT doses were different and the GPRD study included a significantly younger population.</p> <p>Resultantly, statistically significant differences were observed for clinical outcomes including myocardial infarction, cancer and death. An 'adjusted' analysis, using multiple imputation or propensity scores to account for missing data, mediated some of these differences.</p>
--	--	--	---	--

<p>Tannen et al: 2007 [148]</p> <p>Estrogen affects post-menopausal women differently than estrogen plus progestin replacement therapy</p> <p>Simulated RCT</p>	<p>10,739 patients were enrolled in a RCT assessing the risks and benefits of oestrogen monotherapy hormone replacement therapy vs placebo in postmenopausal women.</p> <p>18,462 patients from the General Practice Research Database were identified that met the RCT inclusion criteria and were included in a simulated RCT, replicating all stages of the original RCT with the exception of randomisation.</p>	<p>Standard Prospective: RCT clinical follow up</p> <p>Routinely Recorded: General Practice Research Database (GPRD)</p>	<p>6890 'exposed' and 11,572 'unexposed' patients were identified in GPRD (compared to 5310 and 5429 in the RCT). Baseline characteristics were broadly comparable. However, patients in the GPRD study were younger, weighed less and had fewer cardiovascular risk factors. Current oestrogen use was higher in the exposed and lower in the unexposed GPRD groups compared to the RCT and women in the RCT had earlier hysterectomy.</p> <p>In the RCT baseline characteristics between treated and placebo groups were identical. Between exposed and unexposed groups in GPRD, significant differences were observed in cardiovascular risk factors (reduced in exposed) and prior use of HRT (increased in exposed) (P&lt;0.001).</p> <p>For clinical outcomes statistically significant differences between GPRD and RCT data were observed in the ITT HR's (95% CI) for 5 of the 10 total outcomes including myocardial infarction (RCT: 0.89 (0.7–1.12) , GPRD: 0.42 (0.31–0.55), P=0.002), stroke (RCT: 1.39 (1.1–1.77), GPRD: 0.82 (0.64–1.05), P=0.032), DVT (RCT: 1.47 (1.04–2.08), GPRD: 0.84 (0.71–1.00), P=0.008), breast cancer (RCT: 0.77 (0.59–1.01), GPRD: 1.15 (0.92–1.42), P=0.031) and death (RCT: 1.04 (0.88–1.22), GPRD: 0.62 (0.52–0.74), P&lt;0.001).</p>	<p>Baseline characteristics differ between RCT and GPRD cohorts and GPRD exposed and unexposed cohorts. These differences are likely important to the disagreements between selected study outcomes.</p> <p>For selected clinical outcomes there are statistically significant and clinically important differences between RCT and GPRD studies, for example death is significantly reduced in the GPRD study.</p>
---	--	--	--	---

<p>Tannen et al: 2008 [145]</p> <p>Replicated studies of two randomized trials of angiotensin converting enzyme inhibitors: further empiric validation of the 'prior event rate ratio' to adjust for unmeasured confounding by indication</p> <p>Simulated RCT</p>	<p><b>RCT1:</b> 9297 patients were enrolled in a RCT assessing Ramipril vs placebo in preventing cardiovascular outcomes including heart failure in patients with stable cardiovascular disease.</p> <p>35,521 patients identified from the General Practice Research Database were included in a simulated RCT, replicating all stages of the original RCT with the exception of randomisation.</p> <p><b>RCT2:</b> 9297 patients were enrolled in a RCT assessing Perindopril vs placebo in preventing cardiovascular outcomes including heart failure in patients with stable cardiovascular disease.</p> <p>19,958 patients identified from the General Practice Research Database were included in a simulated RCT, replicating all stages of the original RCT with the exception of randomisation.</p>	<p>Standard Prospective: RCT clinical follow up</p> <p>Routinely Recorded: General Practice Research Database (GPRD)</p>	<p><b>Baseline Characteristics:</b> Both the intervention and placebo groups of the RCTs were comparable. In both GPRD studies, there were statistically significant differences in the majority of variables including cardiovascular risk factors and prior medication use both between the RCT groups and between GPRD exposed and unexposed groups.</p> <p><b>RCT1:</b> For clinical outcomes statistically significant differences between GPRD and RCT data were observed in the ITT RR's (HR's for GPRD data) (95% CI) for all 5 outcomes including myocardial infarction (RCT: 0.79 (0.70–0.89), GPRD: 1.61 (1.47–1.76), P&lt;0.001), stroke (RCT: 0.68 (0.56–0.84), GPRD: 1.30 (1.18–1.43), P&lt;0.001), death (RCT: 0.84 (0.75–0.95), GPRD: 1.23 (1.17–1.30), P&lt;0.001), congestive heart failure (RCT: 0.77 (0.67–0.87), GPRD: 4.26 (3.98–4.56), P&lt;0.001) and coronary revascularisation (RCT: 0.82 (0.74–0.92), GPRD: 1.90 (1.66–2.68), P&lt;0.001).</p> <p><b>RCT 2:</b> For clinical outcomes statistically significant differences between GPRD and RCT data were observed in the ITT RR's (HR's for GPRD data) (95% CI) for 4 of the total 5 outcomes including myocardial infarction (RCT: 0.76 (0.66–0.89), GPRD: 1.55 (1.39–1.73), P&lt;0.001), death (RCT: 0.89 (0.77–1.02), GPRD: 1.34 (1.25–1.45), P&lt;0.001), congestive heart failure (RCT: 0.61 (0.44–0.84), GPRD: 3.92 (3.60–4.26), P&lt;0.001) and coronary revascularisation (RCT: 0.96 (0.85–1.08), GPRD: 1.93 (1.69–2.20), P&lt;0.001).</p>	<p>Baseline characteristics differ between RCT and GPRD cohorts and GPRD exposed and unexposed cohorts. These differences are likely important to the statistically significant differences between RCT and GPRD study outcomes.</p> <p>Further analysis used the Prior Event Rate Ratio (PERR) technique which assumes that the ratio of an outcome event rate in the exposed to unexposed cohorts preceding the study reflects the composite effect of all 'unmeasured' confounders on that specific outcome so long as neither the exposed or unexposed groups received the intervention prior to the commencement of the study. Repeating the analysis using only the 'no prior intervention' patients, produced results more comparable to the RCT.</p>
--	--	--	---	--

<p>Tudur-Smith et al: 2012 [141]</p> <p>The Value of Source Data Verification in a Cancer Clinical Trial</p> <p>RCT</p>	<p>533 patients were enrolled in a RCT to assess control and experimental treatments for advanced cancer.</p> <p>Information recorded prospectively in Case Report Forms was compared to local secondary care medical records and data retrieved from the Office for National Statistics.</p>	<p>Standard Prospective: RCT clinical follow up</p> <p>Routinely Recorded: Secondary care medical records and Office for National Statistics (ONS) data</p>	<p>Results for the primary outcome (overall survival) are comparable between RCT data (HR: 1.18 (95% CI: 0.99 to 1.42), secondary care medical records (1.18 (0.99 to 1.41) and ONS data (1.18 (0.99 to 1.40). There were discrepancies for 13 patients with date of death ranging 1-366 days.</p> <p>When assessing response rates there was some discrepancy between RCT and routine data, with routine data likely picking up additional investigation results. For overall response, the OR (95% CI) in the RCT was 2.45 (1.49 to 4.04) versus 1.67 (1.04 to 2.68) in the routine data. Both were significant (P=0.003 vs P=0.03).</p> <p>There were discrepancies between reported Serious Adverse Events (SAE's). 20 patients reported additional SAE's in the RCT data and 33 in the secondary care medical records.</p>	<p>Data regarding the primary outcome of mortality are comparable.</p> <p>However, data for other clinical outcomes including 'overall response' and SAE's demonstrated some discrepancy. However, the overall results and significance of the RCT outcomes are consistent whether using RCT or routinely recorded data.</p>
<p>Weiner et al: 2008 [157]</p> <p>Replication of the Scandinavian Simvastatin Survival Study using a primary care medical record database prompted exploration of a new method to address unmeasured confounding</p> <p>Cohort Study</p>	<p>A RCT recruited 4444 patients and involved the prescription of simvastatin, assessing cardiac outcomes and mortality. The study was replicated in all aspects except randomisation in 4151 similar patients present in the GPRD.</p> <p>Patient characteristics and study outcomes were compared between RCT and GPRD studies.</p>	<p>Standard Prospective: RCT clinical follow up</p> <p>Routinely Recorded: General Practice Research Database (GPRD)</p>	<p>There were significant baseline differences between the 4151 GRPD patients and the 4444 RCT patients, including a higher rate of prior coronary revascularisation, BMI, diabetes mellitus and lower rate of aspirin therapy.</p> <p>Outcomes were similar for mortality (RR 0.70 95% CI: 0.58-0.85 vs HR 0.66 95% CI: 0.49-0.89 P = 0.94) and myocardial infarction (RR 0.67 95% CI: 0.58-0.77 vs HR 0.78 95% CI: 0.61-1.01 P = 0.28) between RCT and GRPD respectively. However, coronary revascularisation increased in the GPRD cohort (HR 2.17 95% CI: 1.76-2.68) and decreased in the RCT (RR 0.63 95% CI: 0.54-0.74), differing significantly (P &lt; 0.001).</p>	<p>There were significant baseline differences between the RCT and GPRD cohorts.</p> <p>Outcomes calculated using the GPRD cohort are not significantly different to those calculated in the RCT, with the exception of coronary revascularisation, which was aberrantly higher in the statin treated group in the GRPD cohort. However following further analyses, involving only a subgroup of GPRD patients that had not taken cholesterol lowering medications prior to the pre-study period and using Prior Event Rate Ratio analysis to address unmeasured confounding and described in the summary of Tannen et al 2008, the outcomes were similar to the RCT.</p>

<p>Williams et al: 2003 [64]</p> <p>Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment</p> <p>Simulated RCTs</p>	<p><b>RCT1:</b> Open access vs routine follow up for inflammatory bowel disease</p> <p><b>RCT 2:</b> Community vs inpatient diagnosis of obstructive sleep apnoea</p> <p><b>RCT 3:</b> Comparison of two surgical techniques for the treatment of urinary stress incontinence</p>	<p><b>RCT1:</b> Standard Prospective: Patient completed questionnaire: - 0,6,12,18,24 months RCT clinical follow up</p> <p>Routinely Recorded: Primary and secondary care medical records</p> <p><b>RCT 2:</b> Standard Prospective: Patient completed questionnaire RCT clinical follow up</p> <p>Routinely Recorded: Primary and secondary care medical records</p> <p><b>RCT 3:</b> Standard Prospective: Patient completed questionnaire: - 0,3,6,12 months RCT clinical follow up</p> <p>Routinely Recorded: Secondary care medical records</p>	<p><b>RCT1:</b> Surrogate coded terms in routine data were identified to substitute data collected during the RCT. This was available for 9/13 outcomes. There is similarity in calculated outcomes and no significant differences were identified between groups using either dataset. The biggest mean difference between groups was for SF-36 'Vitality' (RCT data -3.72 (-10.72 to 3.27) 'Tired all the time' -0.01 (-0.17 to 0.14). Similarly the outcomes regarding healthcare resource use were comparable.</p> <p><b>RCT 2:</b> Data was available for only 90/102 patients from the individual hospital's medical records. Primary care data was available for only 34/102 due to refusal of consent. Only 19/102 patients had 2 episodes of care routinely recorded and therefore met the RCT inclusion criteria. Only 1/5 outcomes could be calculated – positive diagnosis and results were comparable to RCT data with no significant difference identified. Data assessing healthcare resource use were comparable.</p> <p><b>RCT3 :</b> Data for enrolled patients was missing from all routine sources, most pronounced on a single hospital's records where only 39 of 95 patients could be identified. Of the 25 clinical measures, 14 could be determined from routine data. Broadly, results were comparable including the primary outcome (no significant difference between surgical techniques). For outcome (post-op pain at 5 days) the significant difference or site of pain could not be determined using routine data.</p>	<p><b>RCT 1:</b> Routinely recorded data was broadly comparable to RCT data and similar study results were concluded.</p> <p><b>RCT 2:</b> Routinely recorded data was limited with data available for only a proportion of the RCT sample. Furthermore, although data regarding healthcare resource use were comparable, these data were extracted from the same source. Just 1/5 clinical outcomes could be assessed.</p> <p><b>RCT 3:</b> Routinely recorded data was limited with data available for only a proportion of the RCT sample. With the exception of one outcome involving the assessment of post-operative pain, clinical outcomes were comparable.</p>
--	---	--	---	---

	<p><b>RCT 4:</b> Autologous blood transfusion vs donor blood transfusion in total knee replacement surgery</p>	<p><b>RCT 4:</b> Standard Prospective: Patient completed questionnaire: - 0, 7D, 4W, 3M RCT clinical follow up</p> <p>Routinely Recorded: Secondary care medical records</p>	<p><b>RCT 4:</b> 223/231 patients could be identified from the hospital medical records. 2 out of 11 clinical outcomes could not be calculated. For 9/11 the results were comparable but one significant difference was not found (post-operative infection rates) and one was found (increased post-operative circulatory disorders) in comparison to the RCT results. For the assessment of QOL (EuroQoL) only 1/5 variables could be calculated from routine data, the result was comparable. Data assessing healthcare resource use were comparable.</p>	<p><b>RCT 4:</b> Routinely recorded data were relatively complete and broadly data were comparable. However, unlike the previous RCTs, significantly different results were obtained for 2/11 outcome measures.</p> <p>There were extremely limited surrogate routinely recorded measures in the assessment of QOL, compared to standard self-report questionnaires.</p>
--	--	--	--	--

Table B.2: Included Studies: Agreement between Routinely Recorded Health Economic Data and Data Collected through Standard Prospective Methods

Study Reference	Study Summary	Data Sources	Assessment of Agreement	Appraisal
Byford et al: 2007 [158]  Comparison of alternative methods of collection of service use data for the economic evaluation of health care interventions  RCT	397 patients recruited into a RCT assessing treatments for deliberate self harm completed a questionnaire (Client Services Receipt Inventory (CSRI)) assessing healthcare resource use.  Primary care medical record data was retrieved for 272/397 patients for GP and allied health service resource use and compared to data from the completed CSRI.	Standard Prospective: Patient completed questionnaire (Client Services Receipt Inventory) - 6, 12 months  Routinely Recorded: Primary care medical records	A mean 8.78 GP contacts were reported on the CSRI compared to 10.66 from medical records (95% limits of agreement: 13.71 to 16.94, Mean Difference: 1.88, 95% CI: 0.96 to 2.81), representing an underreporting.  Broadly, the record of services not administered in primary care was under recorded in primary care medical records. This includes a mean difference of -2.27 (95% CI: 3.39 to -1.15) for total outpatient appointments and -2.47 (95% CI: 4.70 to 0.23) for total inpatient days.	There is under-reporting of the number of GP contacts and 'over-reporting' of contacts with other health services although it is likely that these services are poorly recorded on primary care records.  Wide limits of agreement are noted indicating the presence of outliers where there is likely marked discrepancy.
Chishti et al: 2013 [174]  How reliable are stroke patients' reports of their numbers of general practice consultations over 12 months?  RCT	115 patients enrolled in a RCT assessing home blood pressure monitoring completed a questionnaire assessing healthcare resource use.  Primary care medical record data was retrieved for 87 patients and compared to data from the patient completed questionnaires.	Standard Prospective: Patient completed questionnaire - 12 months  Routinely Recorded: Primary care medical records	For the patient completed questionnaires, a mean 5.7 (SD: 5.4, Range: 0 to 24, n=83) GP appointments were reported compared to 7.2 (SD: 5.9, Range: 0 to 28, n=73) in the medical record. The mean difference is 1.6 (95% CI 0.5-2.7); P = 0.004, representing an under reporting of 22%.  For consultations with a nurse, a patient reported mean of 1.4 (SD: 2.5, Range: 0 to 12, n=73) compared to 1.9 (SD: 2.3, Range: 0 to 14, n=73) determined from medical records. The mean difference was not significant 0.5 (95% CI: -0.2 to 1.2; P = 0.12).	There is significant patient under reporting of healthcare utilisation compared to medical record data.  Recall bias may be responsible with the questionnaire completed at 12 months.

<p>Dixon et al: 2009 [151]</p> <p>Is it cost effective to introduce paramedic practitioners for older people to the ambulance service? Results of a cluster randomised controlled trial</p> <p>Cluster RCT</p>	<p>2854 elderly patients were randomised to receive intervention from a paramedic practitioner compared to standard care.</p> <p>Healthcare resource use was determined through routinely recorded secondary care medical records and ambulance records. Additionally, to examine quality of life and healthcare resource use, a patient completed questionnaire including the EQ-5D was administered.</p>	<p>Standard Prospective: Patient completed questionnaire including EQ-5D - 28 days</p> <p>Routinely Recorded: Secondary care and ambulance medical records</p>	<p>The questionnaire regarding healthcare resource use provides is directly comparable to routinely recorded data.</p> <p>However, routinely recorded data was available for 2854 patients and questionnaire data available for a subset of 938. For the intervention and control groups respectively, the total costs of healthcare resource use using routinely recorded data were £3966 and £4166. For the subset completing the questionnaire costs were £2102 and £2641.</p>	<p>There is no statistical assessment of agreement between patient completed questionnaires and routinely recorded data despite the data being comparable.</p> <p>There is marked discrepancy between costs derived from questionnaires and routinely recorded data and the study provides explanation in the likely existence of bias with patients returning questionnaires less unwell at baseline.</p>
<p>Ford et al: 2007 [175]</p> <p>The children's services interview: validity and reliability</p> <p>Cross Sectional Study</p>	<p>The parents of 87 children attending a complex mental health service completed the Children's Services Interview, a parent-completed questionnaire assessing mental healthcare resource use. 25 parents completed a second questionnaire.</p> <p>Parent reported healthcare resource use was compared with local secondary care medical records.</p>	<p>Standard Prospective: Parent completed Children's Services Interview: - 2 years - present</p> <p>Routinely Recorded: Secondary care medical records</p>	<p>The mean parent reported number of total appointments was 5.6 (Range 1-30) compared to 7.0 (Range 1-50) retrieved from the medical records (P=0.1). For appointments at the primary mental health trust, there was good agreement (ICC = 0.77, 95% CI: 0.67-0.85).</p> <p>There was good agreement for the type of intervention received (e.g. Cognitive Behavioural Therapy, Kappa: 0.81 (SE: 0.07)) but poor agreement for type of professional seen (e.g. Psychiatrist, Kappa: 0.07 (SE: 0.12) and healthcare resource use other than the primary mental health trust (e.g. Primary Care and Contact with Teachers, respectively, Kappa 0.30 (SE: 0.08) and 0.03 (SE: 0.03)).</p>	<p>Parents' under-reported healthcare resource use compared to data retrieved from medical records although the difference was non-significant.</p> <p>Data is presented regarding agreement for other variables. Parents reported type of intervention received in agreement to medical records but reported greater resource use for care received outside of the primary mental health trust, most evident for primary care and contact involving the educational sector.</p>

<p>Hussain et al: 2012 [179]</p> <p>HERALD (Health Economics using Routine Anonymised Linked Data)</p> <p>Cohort Study</p>	<p>500 patients diagnosed with ankylosing spondylitis and enrolled in a cohort study were invited to complete questionnaires reporting attendance at healthcare services.</p> <p>Patient reported healthcare use was compared to routinely recorded datasets accessed through the Secure Anonymised Information Linkage (SAIL) Databank.</p>	<p>Standard Prospective: Patient completed questionnaire: - 3 months</p> <p>Routinely Recorded: Secure Anonymised Information Linkage (SAIL) Databank (Inpatient, Outpatient, Primary Care, Emergency Care)</p>	<p>Primary care data was available for 183 patients. In all cases, self-reported 'visits' were underreported compared to the 'events' recorded in the primary care records. 'Events' also includes non-visits. E.g. Patients with high disease severity 1.78 (95% CI: 1.32-2.21) visits compared to 4.25 (3.22-5.28) events.</p> <p>Outpatient data was available for 236 patients. In all cases, self-reported visits were over-reported. Most evident in patients with high disease severity, reporting 2.55 (1.54-3.55) visits compared to 1.51 (1.06-1.97) over the previous 3 month period.</p> <p>Inpatient data was available for 296 patients. Patients tended to under-report admissions. Most evident in young patients, reporting 0.05 (0.01-0.09) admissions compared to 0.16 (0.08-0.25) recorded.</p>	<p>There is no statistical assessment of agreement between patient completed questionnaires and routinely recorded data.</p> <p>GP data were not comparable for assessing 'visits'.</p> <p>Outpatient attendances were over-reported and inpatient attendances under-reported by patients compared to medical records over the 3 month recall period.</p>
<p>Kennedy et al: 2002 [160]</p> <p>Resource use data by patient report or hospital records: Do they agree?</p> <p>RCT</p>	<p>315 patients enrolled in an orthopaedic RCT, receiving care from an orthopaedic surgeon or orthopaedic medical specialist, were invited to complete a self-report questionnaire assessing healthcare resource use at 3 months and 1 year.</p> <p>243 patients completed the questionnaire. Patient reported healthcare resource use was compared with local secondary care medical records.</p>	<p>Standard Prospective: Patient completed questionnaire: - 3 months, 1 year</p> <p>Routinely Recorded: Secondary care medical records</p>	<p>The questionnaire regarding healthcare resource use provides a direct comparator to routinely recorded data.</p> <p>Overall the intraclass correlation coefficient was 0.54 (95% CI 0.31 to 0.70), indicating moderate agreement.</p> <p>The mean patient reported hospital attendance was 5.6 compared to 4.3 recorded in medical records (P=0.006).</p> <p>21% of patients who had been referred but did not report attendance, had attended based on medical records.</p>	<p>The intraclass correlation coefficient indicates moderate agreement. However, there are a significantly increased number of patient reported attendances compared to hospital records. The authors suggest this is likely explained by attendance at other centres where medical records have not been accessed as part of this study.</p> <p>Furthermore, a subgroup of patients did not report attendance when this was evident from the medical records.</p>

<p>Mistry et al: 2005 [176]</p> <p>Comparison of general practitioner records and patient self-report questionnaires for estimation of costs</p> <p>RCT</p>	<p>324 patients enrolled in a RCT assessing the cost-effectiveness of antidepressant treatments had primary care medical record data available for a 12 month period.</p> <p>85 patients completed 6 questionnaires regarding healthcare resource use and 82 patients completed 4 or 5 questionnaires. Healthcare resource use was compared between patient completed questionnaires and primary care medical records.</p>	<p>Standard Prospective: Patient completed questionnaire: - 1, 2, 3, 6, 9 and 12 months</p> <p>Routinely Recorded: Primary care medical records</p>	<p>The total mean patient reported number of healthcare contacts was 17.20 (SD: 30.24) compared to 12.64 (SD: 10.11) determined from medical records for patients completing all 6 questionnaires. This difference was not significant (paired t test: <math>t=-1.755</math>, <math>P=0.083</math>).</p> <p>Patients reported a mean 9.94 (SD: 6.11) contacts with the GP which moderately agreed with data from medical records (9.79 SD: 5.06, Kappa: 0.370). However, for contact with services not delivered by the GP, patient reported contact was greater (e.g. patient reported social services 3.27 SD: 21.45 compared to data from medical records 0.09SD: 0.77, Kappa: 0.021).</p> <p>The total costs of healthcare resource use using patient reported data was 680.04 (SD: 1,634.63) compared to 545.75 (SD: 1,260.20) calculated from medical records. The difference was not significant (<math>P=0.549</math>).</p>	<p>There was agreement between patient reported GP contact and data retrieved from medical records.</p> <p>Although the total numbers of healthcare contacts were not statistically significantly different, patients reported greater contact with healthcare resources not delivered by the GP. This may indicate primary care medical records poorly record data on the access of allied healthcare services.</p>
<p>Morrell et al: 2000 [177]</p> <p>Costs and benefits of community postnatal support workers: a randomised controlled trial</p> <p>RCT</p>	<p>623 women enrolled in a RCT assessing the clinical and cost-effectiveness of providing community postnatal support completed self-report questionnaires at 6 weeks and 6 months postnatally.</p> <p>Self-reported healthcare resource use in 266 women was compared to data retrieved from primary care medical records.</p>	<p>Standard Prospective: Patient completed questionnaire: - 6 weeks, 6 months</p> <p>Routinely Recorded: Primary care medical records</p>	<p>At 6 weeks: There was a total mean 3.8 (SD, 2.6) self-reported GP contacts, compared with 3.4 (SD, 2.3) recorded in the medical records. Self-reported contacts were over-reported by a mean 0.5 of a contact (95% CI, 0.2, 0.7). There was no difference between 'over-reporting' between intervention and control groups (mean difference, 0.2; 95% CI, -0.4, 0.7; <math>t = 0.67</math> on 1 df; <math>p = 0.51</math>).</p> <p>At 6 months: There was under-reporting, the difference between medical records and self-reports was -0.1 contacts (95% CI, -0.7, 0.5).</p>	<p>In the short term (6 weeks), self-reported healthcare contacts were significantly over-reported by a mean of 0.5 of 1 contact.</p> <p>However, at 6 months, there was a non-significant modest under-reporting.</p>

<p>Petrou et al: 2002 [159]</p> <p>The Accuracy of Self-Reported Healthcare Resource Utilisation in Health Economic Studies</p> <p>RCT</p>	<p>82 women enrolled in a RCT assessing a preventative intervention for women at risk of postpartum depression completed self-report questionnaires at 4 and 12 months postpartum, retrospectively detailing healthcare contacts.</p> <p>Patient reported healthcare resource use was compared to data retrieved from primary and secondary care medical records.</p>	<p>Standard Prospective: Patient completed questionnaire: - 4, 12 months</p> <p>Routinely Recorded: Primary and secondary care medical records</p>	<p>For primary care visits, at months 1-4 29.4% of patients recorded visits accurately with 56.9% underreporting. At months 5-12, the proportions were 28.0% and 58.0%. Community midwifery was accurately reported in 21.1% and underreported in 49.3%. The differences between self-reported primary care attendances and medical record review were significant in all cases (Wilcoxon Sign-Rank Test <math>P &lt; 0.001</math>).</p> <p>The accuracy of other healthcare contacts including secondary care attendances ranged 90.8%-100%, differences were non-significant.</p> <p>Multivariate linear regression identified a reduced absolute number of GP visits, the presence of postnatal depression and accommodation problems were significantly associated with reduced under-reporting. The presence of financial difficulties and reduced years spent with current partner were significantly associated with over-reporting.</p>	<p>Significant differences were identified between self-reported primary care visits and primary care medical records, with the majority of patients under-reporting.</p> <p>Notably, the differences remain very similar between 4 and 12 months, indicating a potentially static effect of recall bias.</p> <p>There were no significant differences between self-reported and medical record data regarding other healthcare contacts including secondary care.</p>
<p>Richards et al: 2003 [167]</p> <p>Patient-reported use of health service resources compared with information from health providers</p> <p>RCT</p>	<p>185 elderly patients enrolled in a RCT assessing a 'hospital-at-home' service with routine care completed questionnaires at 4 and 12 weeks post admission assessing healthcare resource use.</p> <p>Patient reported healthcare resource use was compared to data retrieved from primary, secondary and community medical records.</p>	<p>Standard Prospective: Patient completed questionnaire: - 4, 12 weeks</p> <p>Routinely Recorded: Primary, secondary and community medical records</p>	<p>Agreement was greatest for 'Hospital Readmission' (90.6%, Kappa: 0.68 (95% CI: 0.54–0.82)) and GP surgery visits (88.0%, Kappa: 0.60 (0.44–0.76)) and weakest for 'Physiotherapy' (72.1%, Kappa: 0.23 (0.13–0.33)).</p> <p>Patients under-reported primary and secondary care attendances. Medical records indicated a greater number of hospital admissions in the previous 12 weeks (McNemar <math>\chi^2 = 4.8</math>, d.f.=1, <math>P = 0.03</math>), and a greater number than that reported by patients (Stuart–Maxwell <math>\chi^2 = 6.8</math>, d.f. =2, <math>P = 0.03</math>). GPs reported a greater number of visits than patients (Stuart–Maxwell <math>\chi^2 = 6.5</math>, d.f. =2, <math>P = 0.04</math>).</p>	<p>Agreement was 'Good' or 'Moderate' for the majority of healthcare settings including hospital admissions and GP visits.</p> <p>However, patients under reported contact with all healthcare services with the exception of 'Requesting GP Visit'.</p>

<p>Thorn et al: 2016 [142]</p> <p>Validation of the Hospital Episode Statistics Outpatient Dataset in England</p> <p>RCT</p>	<p>370 men enrolled in a RCT assessing PSA testing vs standard care for prostate cancer were flagged and had their data prospectively extracted from medical records.</p> <p>Healthcare resource use was compared to data retrieved from the Hospital Episode Statistics Outpatient dataset.</p>	<p>Standard Prospective: RCT clinical follow up, primary care medical records</p> <p>Routinely Recorded: Hospital Episode Statistics Outpatient Dataset (HES OP)</p>	<p>4922 outpatient appointments with urology services were identified from primary care medical records. 12,154 appointments were identified in the HES OP dataset (all specialities). 0.4% of appointments did not record speciality. 7452 appointments occurred with a relevant speciality and 4088/4922 appointments recorded in medical records were identified in HES OP (83.1 %; 95 % CI 82.0–84.1). Diagnosis codes were present in only 0.9% and operation codes in 6.7%. 215/370 men (58.1 %) had at least one appointment that was unmatched in HES OP. 2195/2755 (79.7 %; 95 % CI 78.2–81.2) matches were observed pre-2008, while 1893/2167 (87.4 %; 95 % CI 86.0–88.9) matches were observed post-2008 (P&lt;0.001).</p>	<p>There is moderate agreement of clinic appointments that were recorded in the RCT. However, there are significant numbers of ‘additional’ appointments identified and also ‘missing’ appointments.</p> <p>Routinely recorded data has improved since 2008 and for assessing healthcare resource use currently represents a useful data source. However, clinical data such as diagnostic details is extremely poorly recorded.</p>
<p>Thorn et al: 2016 [66]</p> <p>Validating the use of Hospital Episode Statistics data and comparison of costing methodologies for economic evaluation: an end-of-life case study from the Cluster randomised trial of PSA testing for Prostate cancer (CAP)</p> <p>RCT</p>	<p>282 men enrolled in a RCT assessing PSA testing vs standard care for prostate cancer were included in this methodological study.</p> <p>Healthcare resource use costs derived from data collected during the RCT from medical record review were compared to costs derived from data retrieved from the Hospital Episode Statistics Inpatient dataset.</p>	<p>Standard Prospective: RCT clinical follow up, primary care medical records</p> <p>Routinely Recorded: Hospital Episode Statistics Inpatient Dataset (HES)</p>	<p>The inpatient cost profiles derived from RCT and HES were comparable, the 95% CIs overlapped at each month, and values were not significantly different at the 1% significance level.</p> <p>The mean resource use per man over the final year of life was £11 122 (95% CI £9083 to £13 161) using RCT data and national reference costs, and £10 223 (95% CI £8880 to £11 565) using HES data with reference costs. Costs associated with HES data were slightly lower (about 8%) than those associated with RCT data, but the difference was not significant (p=0.3).</p> <p>The definition of a ‘single event’ was not consistent between the two sets of resource use, with some episodes recorded as single events in RCT data appearing as multiple events in HES and vice versa. 11 men (3.8%) for whom events were recorded in HES had all these events missing from RCT data and 7 men (2.4%) with no events according to HES had events identified in RCT data.</p>	<p>There was good agreement between the inpatient cost profiles derived from HES compared to RCT data and between the cost per patient.</p>

Table B.3: Included Abstracts: Agreement between Routinely Recorded Data and Data Collected through Standard Prospective Methods

Study Reference	Study Summary	Data Sources	Assessment of Agreement	Appraisal
<p>Breeman et al: 2011 [172]</p> <p>Patient reported clinical outcomes: the challenges and implications for randomised controlled trials</p> <p>RCT</p>	<p>Four RCTs assessing knee surgery included the collection of data to measure outcomes through patient completion of questionnaires.</p> <p>Self-reported data was compared to data from 'routine sources' including medical records.</p>	<p>Standard Prospective: Patient completed questionnaire - Interval not reported</p> <p>Routinely Recorded: Medical records and 'routine data sources'</p>	<p>The questionnaires potentially provide a direct comparator to routinely recorded data.</p> <p>'15% of patient reported knee-related hospital re-admission could not be verified through routine data sources or medical records.'</p>	<p>There is no statistical assessment of agreement presented between patient completed questionnaires and routinely recorded data despite the data being comparable.</p> <p>There is marked discrepancy between patient-reported outcomes and outcomes retrieved from medical records. The authors propose the underlying reasons are a result of patient misunderstanding and inaccuracies of routine data.</p>
<p>Embleton et al: 2015 [183]</p> <p>Impact of retrospective data verification on the results of the academic led ICON6 trial</p> <p>RCT</p>	<p>282 patients were enrolled in a RCT assessing novel chemotherapy + standard treatment vs placebo + standard treatment. Standard prospective methods including Case Report Forms were completed to measure the primary outcome of progression free survival.</p> <p>RCT data was compared to secondary care medical records for all patients during source data verification.</p>	<p>Standard Prospective: RCT clinical follow up</p> <p>Routinely Recorded: Secondary care medical records</p>	<p>There were 253 'progressions'. Source data verification identified 2 additional progressions, 1 in each trial arm.</p> <p>The result of the log-rank test changed marginally, remaining <math>P &lt; 0.001</math>. The HR of 0.57 was unchanged, but the CI changed from 0.45-0.74 to 0.44-0.73. Median time to event remained at 8.7 months.</p>	<p>There is good agreement for the primary outcome based on the descriptive comparison presented.</p>

<p>Herrington et al: 2015 [182]</p> <p>Can vascular mortality be reliably ascertained from the underlying cause of death recorded on a medical death certificate? Evidence from the 2800 adjudicated Heart Protection Study deaths</p> <p>RCT</p>	<p>2835 of total 20,536 participants of the Heart Protection Study died during study follow up.</p> <p>Data retrieved from UK mortality registers was compared with data collected using standard prospective methods ('clinically adjudicated' data).</p>	<p>Standard Prospective: RCT clinical follow up</p> <p>Routinely Recorded: 'UK Mortality Registers'</p>	<p>2778/2835 deaths were recorded in UK mortality registers.</p> <p>1152 certified compared to 1260 adjudicated coronary heart disease (CHD) deaths. 108 adjudicated CHD deaths were wrongly certified to another cause and 81 non-CHD deaths were wrongly certified as CHD (Kappa: 0.86, 95% CI: 0.84-0.88).</p> <p>161 certified compared to 214 adjudicated stroke deaths. 53 were certified as other causes and 26 non-stroke deaths were wrongly certified as stroke (Kappa: 0.78, 95% CI: 0.74-0.83).</p> <p>7 certified did not compare to 60 adjudicated ischaemic stroke deaths (K: 0.19, 95% CI: 0.07-0.31).</p> <p>The allocation to intervention vs control in HPS reduced vascular deaths (RR: 0.83, 95% CI: 0.75-0.91). Using only certified deaths the outcome is comparable (RR: 0.81, 95% CI: 0.74-0.90).</p>	<p>There is good agreement for all clinical measures with the exception of 'ischaemic stroke'.</p> <p>The primary trial outcomes are almost identical whether using 'adjudicated' or certified deaths.</p> <p>The authors conclude that using routinely recorded data from the UK Mortality Registers is likely to be sufficiently reliable to assess vascular deaths.</p>
<p>Shaw et al: 1998 [168]</p> <p>Can we trust the quality of routine hospital outpatient information in the UK? Validating outpatient data from the patient administration system</p> <p>Cohort Study</p>	<p>Attendance details for 140 patients at 4 NHS hospitals were examined.</p> <p>The grade of doctor seen and management decision were recorded by clinician report immediately following the appointment and compared to the information recorded in the Patient Administration System (PAS).</p>	<p>Standard Prospective: Clinician report</p> <p>Routinely Recorded: Secondary care medical records</p>	<p>There was agreement between the clinician report and the PAS data in 118/140 (84.3%) cases for grade of doctor seen and in 105/139 (76.7%) cases for the management decision. There was complete agreement for both items in 88/139 (62.6%) cases.</p> <p>Kappa values indicated 'good' agreement between the two data sources. However, sensitivity statistics suggested that the likely accuracy of each data item varied.</p>	<p>There is a descriptive and statistical assessment of agreement.</p> <p>Data extraction is limited as the full report was not retrievable through available resources during this review.</p>
<p>Smith et al: 2015 [181]</p> <p>Assessing the accuracy of routinely collected data and their use in pressure ulcer trials. Cross Sectional Study</p>	<p>A cross sectional study compared routinely recorded data on the incidence of pressure ulcers, recorded through the English NHS Safety Thermometer, to research nurse review.</p>	<p>Standard Prospective: Research nurse review</p> <p>Routinely Recorded: English NHS Safety Thermometer</p>	<p>Reported low accuracy of routinely recorded data.</p> <p>Weighted sensitivity estimate: 48.2% (95% CI: 35.4%-56.7%).</p>	<p>There is marked discrepancy between routinely recorded data and research nurse review. The study concludes routinely recorded data is not satisfactory for measuring the outcomes of pressure ulcer RCTs.</p>

<p>Wright-Hughes et al: 2013 [184]</p> <p>Can the use of routine data enhance collection of the primary outcome in the SHIFT trial?</p> <p>RCT</p>	<p>The SHIFT RCT involved randomising adolescents following an episode of self harm to family therapy vs standard care. The primary outcome of 'repetition, leading to hospital attendance' involved research nurse prospective review of local hospital record data.</p> <p>Data was also retrieved from secondary care medical records and agreement compared.</p>	<p>Standard Prospective: RCT clinical follow up, secondary care medical records</p> <p>Routinely Recorded: Hospital Episode Statistics Inpatient Dataset (HES)</p>	<p>'Our comparison found advantages of data collection via NHS Digital to include the acquisition of more comprehensive and timely trial outcome data, potentially at a reduced cost, whilst disadvantages included ambiguity in the classification of self-harm relatedness for a proportion of episodes.'</p>	<p>There is no presented data regarding the assessment of agreement performed.</p> <p>However, as a result of this analysis, the researchers altered the RCT methodology to access HES data as the primary dataset for measuring RCT outcomes.</p>
--	--	--	---	--

Table B.4: PRISMA Checklist

Section/topic	#	Checklist item	Reported on page #
<b>TITLE</b>			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	Title Page
<b>ABSTRACT</b>			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	Abstract
<b>INTRODUCTION</b>			
Rationale	3	Describe the rationale for the review in the context of what is already known.	Introduction
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	Introduction
<b>METHODS</b>			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	Methods The review was not eligible for registration in the PROSPERO database.
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	Methods
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	Methods
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	Methods Appendix II
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	Methods
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	Methods
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	Methods

Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	Methods A narrative appraisal was performed. Formal assessment was performed where relevant to the assessment of agreement and when possible considering the methods used in the included study.
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	Methods
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I <sup>2</sup> ) for each meta-analysis.	Methods

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	Methods A narrative appraisal has been performed. Formal assessment has not been performed due to the heterogeneity of included studies.
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/a
<b>RESULTS</b>			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	Results
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Results
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	Results Presented in tables where relevant.

Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Results
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	N/a
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	N/a
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/a
<b>DISCUSSION</b>			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	Discussion
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	Discussion
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	Discussion
<b>FUNDING</b>			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	To be included in journal submission.

Table B.5: Search Strategy: MEDLINE (OVID) "Prospective Studies" 1.10.16

1	Administrative Data Research Network.tw.
2	Clinical Practice Research Datalink.tw.
3	((Driving Licence Agency) or (Driving Licence Authority)).tw.
4	(Department of Work and Pensions).tw.
5	General Practice Extraction Service.tw.
6	(General Practice Research Database or General Practice Registry Database).tw.
7	Hospital Episode Statistics.tw.
8	"Revenue and Customs".tw.
9	"Health and Social Care Information Centre".tw.
10	NorthWest eHealth.tw.
11	Office for National Statistics.tw.
12	Secure Anonymised Information Linkage Databank.tw.
13	(NHS Wales Informatics Service or Patient Episode Database for Wales).tw.
14	"The Health Improvement Network".tw.
15	QResearch.tw.
16	ResearchOne.tw.
17	Information Services Division.tw.
18	or/1-17
19	(ADRN or CPRD or DVLA or DWP or GPES or GPRD or HES or HMRC or HSCIC or NWEH or ONS or SAIL or PEDW).tw.
20	("dose-width product" or Dose width product).tw.
21	General Practice Education Supervisors.tw.
22	(Geriatric Psychiatry Research Division or Aspergillus or gastropharyngeal reflux disease or (gene and prion disease)).tw.
23	(balanced starch or hydroxyethyl starch or hydroxyethylstarch or hydroxy ethyl starch or Hospital Eye Service or hip extensor stretch).tw.
24	("Add-ons" or Oral nutrient supplement or Occipital nerve stimulator).tw.
25	(Sheffield Assessment for Letters or Sheffield Assessment Instrument for Letters or Stimulation Assistance through Iterative Learning).tw.
26	or/20-25
27	19 not 26
28	((routine or clinic or primary or general practice or general practitioner or GP?) and data and source?).ti.
29	((routine or clinic or primary or general practice or general practitioner or GP?) adj2 data).ab.
30	data source?.ab. /freq=2
31	29 and 30
32	27 or 28 or 31
33	((Electronic or Routine or Administrative) adj (health or medical or record or data)).tw.

34	((Regist\$ or Hospital) adj (data or recor\$)).tw.
35	33 or 34
36	32 or 35
37	exp Great Britain/
38	(national health service* or nhs*).ti,ab,in.
39	(english not ((published or publication* or translat* or written or language* or speak* or literature or citation*) adj5 english)).ti,ab.
40	(gb or "g.b." or britain* or (british* not "british columbia") or uk or "u.k." or united kingdom* or (england* not "new england") or northern ireland* or northern irish* or scotland* or scottish* or ((wales or "south wales") not "new south wales") or welsh*).ti,ab,jw,in.
41	(bath or "bath's" or ((birmingham not alabama*) or ("birmingham's" not alabama*) or bradford or "bradford's" or brighton or "brighton's" or bristol or "bristol's" or carlisle* or "carlisle's" or (cambridge not (massachusetts* or boston* or harvard*)) or ("cambridge's" not (massachusetts* or boston* or harvard*)) or (canterbury not zealand*) or ("canterbury's" not zealand*) or chelmsford or "chelmsford's" or chester or "chester's" or chichester or "chichester's" or coventry or "coventry's" or derby or "derby's" or (durham not (carolina* or nc)) or ("durham's" not (carolina* or nc)) or ely or "ely's" or exeter or "exeter's" or gloucester or "gloucester's" or hereford or "hereford's" or hull or "hull's" or lancaster or "lancaster's" or leeds* or leicester or "leicester's" or (lincoln not nebraska*) or ("lincoln's" not nebraska*) or (liverpool not (new south wales* or nsw)) or ("liverpool's" not (new south wales* or nsw)) or ((london not (ontario* or ont or toronto*)) or ("london's" not (ontario* or ont or toronto*)) or manchester or "manchester's" or (newcastle not (new south wales* or nsw)) or ("newcastle's" not (new south wales* or nsw)) or norwich or "norwich's" or nottingham or "nottingham's" or oxford or "oxford's" or peterborough or "peterborough's" or plymouth or "plymouth's" or portsmouth or "portsmouth's" or preston or "preston's" or ripon or "ripon's" or salford or "salford's" or salisbury or "salisbury's" or sheffield or "sheffield's" or southampton or "southampton's" or st albans or stoke or "stoke's" or sunderland or "sunderland's" or truro or "truro's" or wakefield or "wakefield's" or wells or westminster or "westminster's" or winchester or "winchester's" or wolverhampton or "wolverhampton's" or (worcester not (massachusetts* or boston* or harvard*)) or ("worcester's" not (massachusetts* or boston* or harvard*)) or (york not ("new york*" or ny or ontario* or ont or toronto*)) or ("york's" not ("new york*" or ny or ontario* or ont or toronto*))))).ti,ab,in.
42	(bangor or "bangor's" or cardiff or "cardiff's" or newport or "newport's" or st asaph or "st asaph's" or st davids or swansea or "swansea's").ti,ab,in.
43	(aberdeen or "aberdeen's" or dundee or "dundee's" or edinburgh or "edinburgh's" or glasgow or "glasgow's" or inverness or (perth not australia*) or ("perth's" not australia*) or stirling or "stirling's").ti,ab,in.
44	(armagh or "armagh's" or belfast or "belfast's" or lisburn or "lisburn's" or londonderry or "londonderry's" or derry or "derry's" or newry or "newry's").ti,ab,in.
45	or/37-44
46	(exp africa/ or exp americas/ or exp antarctic regions/ or exp arctic regions/ or exp asia/ or exp australia/ or exp oceania/) not (exp great britain/ or europe/)
47	45 not 46
48	36 and 47
49	18 or 48
50	randomized controlled trial.pt.
51	controlled clinical trial.pt.
52	randomized.ab.
53	placebo.ab.

54	clinical trials as topic.sh.
55	randomly.ab.
56	trial.ti.
57	Or/50-56
58	cohort studies/ or longitudinal studies/ or follow-up studies/ or cohort.ti,ab. or longitudinal.ti,ab.
59	exp prospective studies/
60	exp case-control studies/ or exp retrospective studies/ or exp cross-sectional studies/ or (case control or case-control or case series or cross?section\$ or "cross section\$" or retrospective).ti. or case reports.pt.
61	Retrospective adj cohort.ti,ab.
62	58 and 59
63	62 not 60
64	63 not 61
65	57 or 64
66	exp animals/ not humans.sh.
67	65 not 66
68	49 and 67
69	Remove duplicates from 68

Table B.6: Search Strategy: SCOPUS "Prospective Studies" 19.10.16

1	TITLE-ABS-KEY ("Administrative Data Research Network" OR "Clinical Practi* Research Datalink" OR "Driv* W/1 Vehicle Licen?ing Agency" OR "Driv* W/1 Vehicle Licen?ing Authority" OR "Department W/1 Work and Pensions" OR "General Practi* Extraction Service" OR "General Practi* Research Database" OR "General Practi* Registry Database" OR "Hospital Episode Statistics" OR "Revenue and Customs" OR "Health and Social Care Information Centre" OR "NorthWest eHealth" OR "Office for National Statistics" OR "Secure Anonymised Information Linkage Databank" OR "NHS Wales Informatics Service" OR "Patient Episode Database for Wales" OR "The Health Improvement Network" OR QResearch OR ResearchOne OR "Information Services Division")
2	TITLE-ABS-KEY (ADRN OR CPRD OR DVLA OR DWP OR GPES OR GPRD OR HES OR HMRC OR HSCIC OR NWEH OR ONS OR SAIL OR PEDW)
3	TITLE-ABS-KEY ("dose-width product" OR "Dose width product")
4	TITLE-ABS-KEY ("General Practice Education Supervis*")
5	TITLE-ABS-KEY ("Geriatric Psychiatry Research Division" OR Aspergillus OR "gastropharyngeal reflux disease*" OR ("gene* AND prion disease*"))
6	TITLE-ABS-KEY ("balanced starch*" OR "hydroxyethyl starch*" OR "hydroxyethyl starch*" OR "hydroxy ethyl starch*" OR "Hospital Eye Service*" OR "hip extensor stretch")
7	TITLE-ABS-KEY ("Add-ons" OR "Oral nutri* supplement*" OR "Occipital nerve stimul*")
8	TITLE-ABS-KEY ("Sheffield Assessment for Letters" OR "Sheffield Assessment Instrument for Letters" OR "Stimulation Assistance through Iterative Learning")
9	#3 OR #4 OR #5 OR #6 OR #7 OR #8
10	#2 AND NOT #9
11	TITLE-ABS-KEY (routine OR clinic* OR primary OR "general practice*" OR "general practis*" OR "general practitioner*" OR GP) AND ("data source")
12	TITLE-ABS-KEY ((Electronic OR Routine OR Administrative) W/0 health OR medical OR recor* OR data)
13	TITLE-ABS-KEY ((Regist* OR Hospital) W/0 data OR recor*)
14	#10 OR #11 OR #12 OR #13
15	TITLE-ABS-KEY (Great Britain)
16	TITLE-ABS-KEY ("national health service*" OR nhs* OR (english AND NOT (published OR publication* OR translat* OR written OR language* OR speak* OR literature OR citation*) W/5 english))
17	TITLE-ABS-KEY (gb OR "g.b." OR britain* OR (british* AND NOT "british columbia") OR uk OR "u.k." OR "united kingdom*" OR (england* AND NOT "new england") OR "northern ireland*" OR "northern irish*" OR scotland* OR scottish* OR ((wales OR "south wales") AND NOT "new south wales") OR welsh*)
18	TITLE-ABS-KEY (bath OR "bath's" OR ((birmingham AND NOT alabama*) OR ("birmingham's" AND NOT alabama*) OR bradford OR "bradford's" OR brighton OR "brighton's" OR bristol OR "bristol's" OR carlisle* OR "carlisle's" OR (cambridge AND NOT (massachusetts* OR boston* OR harvard*)) OR ("cambridge's" AND NOT (massachusetts* OR boston* OR harvard*)) OR (canterbury AND NOT zealand*) OR ("canterbury's" AND NOT zealand*) OR chelmsford OR "chelmsford's" OR chester OR "chester's" OR chichester OR "chichester's" OR coventry OR "coventry's" OR derby OR "derby's" OR (durham AND NOT (carolina* OR nc)) OR ("durham's" AND NOT (carolina* OR nc)) OR ely OR "ely's" OR exeter OR "exeter's" OR gloucester OR "gloucester's" OR hereford OR "hereford's" OR hull OR "hull's" OR lancaster OR "lancaster's" OR leeds* OR leicester OR "leicester's" OR (lincoln AND NOT nebraska*) OR ("lincoln's" AND NOT nebraska*) OR (liverpool AND NOT (new south wales* OR nsw)) OR ("liverpool's" AND NOT (new south wales* OR nsw)) OR ((london AND NOT (ontario* OR ont OR toronto*)) OR ("london's" AND NOT (ontario* OR ont OR toronto*)) OR manchester OR "manchester's" OR (newcastle AND NOT (new south wales* OR nsw)) OR ("newcastle's" AND NOT (new south wales* OR nsw)) OR norwich OR "norwich's" OR nottingham OR "nottingham's" OR oxford OR "oxford's" OR peterborough OR "peterborough's" OR plymouth OR "plymouth's" OR portsmouth OR "portsmouth's" OR preston

	OR "preston's" OR ripon OR "ripon's" OR salford OR "salford's" OR salisbury OR "salisbury's" OR sheffield OR "sheffield's" OR southampton OR "southampton's" OR st albans OR stoke OR "stoke's" OR sunderland OR "sunderland's" OR truro OR "truro's" OR wakefield OR "wakefield's" OR wells OR westminster OR "westminster's" OR winchester OR "winchester's" OR wolverhampton OR "wolverhampton's" OR (worchester AND NOT (massachusetts* OR boston* OR harvard*)) OR ("worchester's" AND NOT (massachusetts* OR boston* OR harvard*)) OR (york AND NOT ("new york*" OR ny OR ontario* OR ont OR toronto*)) OR ("york's" AND NOT ("new york*" OR ny OR ontario* OR ont OR toronto*))))
19	TITLE-ABS-KEY (bangor OR "bangor's" OR cardiff OR "cardiff's" OR newport OR "newport's" OR st asaph OR "st asaph's" OR st davids OR swansea OR "swansea's" OR aberdeen OR "aberdeen's" OR dundee OR "dundee's" OR edinburgh OR "edinburgh's" OR glasgow OR "glasgow's" OR inverness OR (perth AND NOT australia*) OR ("perth's" AND NOT australia*) OR stirling OR "stirling's" OR armagh OR "armagh's" OR belfast OR "belfast's" OR lisburn OR "lisburn's" OR londonderry OR "londonderry's" OR derry OR "derry's" OR newry OR "newry's")
20	#15 OR #16 OR #17 OR #18 OR #19
21	ALL ((Africa OR Americas OR Antarctic OR arctic OR Asia OR Australia OR Oceania) AND NOT Great Britain OR Europe)
22	#20 AND NOT #21
23	#14 AND #22
24	#1 OR #23
25	TITLE-ABS-KEY ("randomized controlled trial" OR "controlled clinical trial" OR "clinical tria*")
26	TITLE (trial)
27	ABS (randomized OR placebo OR randomly)
28	#25 OR #26 OR #27
29	TITLE-ABS-KEY (cohort OR longitudinal OR "follow-up")
30	TITLE-ABS-KEY (prospective)
31	TITLE-ABS-KEY (case-control OR retrospective OR cross-sectional OR "case control" OR case series OR cross?section* OR "cross section*" OR retrospective OR case report*)
32	TITLE-ABS-KEY (Retrospective W/0 cohort)
33	#29 AND #30
34	#33 AND NOT #31
35	#34 AND NOT #32
36	#28 OR #35
37	ALL (anima* AND NOT huma*)
38	#36 AND NOT #37
39	#24 AND #38

Table B.7: Search Strategy: Cochrane Methodology Register “No Study Filter” 1.10.16

1	“Administrative Data Research Network”:ti,ab
2	“Clinical Practi* Research Datalink”:ti,ab
3	((Driv* and Vehicle Licen?ing Agency) or (Driv* and Vehicle Licen?ing Authority)):ti,ab
4	(Department of "Work and Pensions"):ti,ab
5	“General Practi* Extraction Service”:ti,ab
6	(“General Practi* Research Database” or “General Practi* Registry Database”):ti,ab
7	“Hospital Episode Statistics”:ti,ab
8	"Revenue and Customs":ti,ab
9	"Health and Social Care Information Centre":ti,ab
10	“NorthWest eHealth”:ti,ab
11	“Office for National Statistics”:ti,ab
12	“Secure Anonymised Information Linkage Databank”:ti,ab
13	(“NHS Wales Informatics Service” or “Patient Episode Database for Wales”):ti,ab
14	"The Health Improvement Network":ti,ab
15	QResearch:ti,ab
16	ResearchOne:ti,ab
17	“Information Services Division”:ti,ab
18	#1 or #2 or #3 or #4 or #5 or #6 or #7 or #8 or #9 or #10 or #11 or #12 or #13 or #14 or #15 or #16 or #17
19	(ADRN or CPRD or DVLA or DWP or GPES or GPRD or HES or HMRC or HSCIC or NWEH or ONS or SAIL or PEDW):ti,ab
20	("dose-width product" or Dose width product):ti,ab
21	“General Practice Education* Supervis*”:ti,ab
22	(Geriatric Psychiatry Research Division or Aspergillus or gastropharyngeal reflux disease* or (gene* and prion disease*)):ti,ab
23	(balanced starch* or hydroxyethyl starch* or hydroxyethylstarch* or hydroxy ethyl starch* or Hospital Eye Service* or hip extensor stretch):ti,ab
24	("Add-ons" or Oral nutri* supplement* or Occipital nerve stimul*):ti,ab
25	(Sheffield Assessment for Letters or Sheffield Assessment Instrument for Letters or Stimulation Assistance through Iterative Learning):ti,ab
26	#20 or #21 or #22 or #23 or #24 or #25
27	#19 not #26
28	((routine or clinic* or primary or general practice* or general practis* or general practitioner* or GP?) and data and source?):ti,ab
29	#27 or #28
30	((Electronic or Routine or Administrative) NEXT (health or medical or recor* or data)):ti,ab
31	((Regist* or Hospital) NEXT (data or recor*)):ti,ab
32	#30 or #31
33	#29 or #32

34	"Great Britain"
35	(national health service* or nhs*):ti,ab
36	(english not ((published or publication* or translat* or written or language* or speak* or literature or citation*) near english)):ti,ab
37	(gb or "g.b." or britain* or (british* not "british columbia") or uk or "u.k." or united kingdom* or (england* not "new england") or northern ireland* or northern irish* or scotland* or scottish* or ((wales or "south wales") not "new south wales") or welsh*):ti,ab
38	(bath or "bath's" or ((birmingham not alabama*) or ("birmingham's" not alabama*) or bradford or "bradford's" or brighton or "brighton's" or bristol or "bristol's" or carlisle* or "carlisle's" or (cambridge not (massachusetts* or boston* or harvard*)) or ("cambridge's" not (massachusetts* or boston* or harvard*)) or (canterbury not zealand*) or ("canterbury's" not zealand*) or chelmsford or "chelmsford's" or chester or "chester's" or chichester or "chichester's" or coventry or "coventry's" or derby or "derby's" or (durham not (carolina* or nc)) or ("durham's" not (carolina* or nc)) or ely or "ely's" or exeter or "exeter's" or gloucester or "gloucester's" or hereford or "hereford's" or hull or "hull's" or lancaster or "lancaster's" or leeds* or leicester or "leicester's" or (lincoln not nebraska*) or ("lincoln's" not nebraska*) or (liverpool not (new south wales* or nsw)) or ("liverpool's" not (new south wales* or nsw)) or ((london not (ontario* or ont or toronto*)) or ("london's" not (ontario* or ont or toronto*)) or manchester or "manchester's" or (newcastle not (new south wales* or nsw)) or ("newcastle's" not (new south wales* or nsw)) or norwich or "norwich's" or nottingham or "nottingham's" or oxford or "oxford's" or peterborough or "peterborough's" or plymouth or "plymouth's" or portsmouth or "portsmouth's" or preston or "preston's" or ripon or "ripon's" or salford or "salford's" or salisbury or "salisbury's" or sheffield or "sheffield's" or southampton or "southampton's" or st albans or stoke or "stoke's" or sunderland or "sunderland's" or truro or "truro's" or wakefield or "wakefield's" or wells or westminster or "westminster's" or winchester or "winchester's" or wolverhampton or "wolverhampton's" or (worcester not (massachusetts* or boston* or harvard*)) or ("worcester's" not (massachusetts* or boston* or harvard*)) or (york not ("new york*" or ny or ontario* or ont or toronto*)) or ("york's" not ("new york*" or ny or ontario* or ont or toronto*))))):ti,ab
39	(bangor or "bangor's" or cardiff or "cardiff's" or newport or "newport's" or st asaph or "st asaph's" or st davids or swansea or "swansea's"):ti,ab
40	(aberdeen or "aberdeen's" or dundee or "dundee's" or edinburgh or "edinburgh's" or glasgow or "glasgow's" or inverness or (perth not australia*) or ("perth's" not australia*) or stirling or "stirling's"):ti,ab
41	(armagh or "armagh's" or belfast or "belfast's" or lisburn or "lisburn's" or londonderry or "londonderry's" or derry or "derry's" or newry or "newry's"):ti,ab
42	#34 or #35 or #36 or #37 or #38 or #39 or #40 or #41
43	(africa or americas or antarctic regions or arctic regions or asia or australia or oceania) not (great britain or europe):ti,ab
44	#42 not #43
45	#33 and #44
46	#18 or #45

Table B.8: Search Strategy: DIRUM MEDLINE 2012-2016 21.10.16

1	Patient Readmission/
2	Office Visits/
3	House Calls/
4	"Referral and Consultation"/
5	Remote Consultation/
6	Patient Admission/
7	Hospitalization/
8	exp Ambulatory Care Facilities/
9	exp mass screening/
10	exp Diagnostic Imaging/
11	visits.tw.
12	appointment\$.tw.
13	hospitali?ation\$.tw.
14	exp Health Services/
15	Health Resources/
16	"episode of care"/
17	exp General Practice/
18	exp Drug Therapy/
19	Outpatients/
20	exp therapeutics/
21	or/1-20
22	ut.fs.
23	utili?ation.tw.
24	(valid\$ adj5 self report\$.tw.
25	or/22-24
26	21 and 25
27	exp Drug Utilization/
28	((medicine\$ or medication\$ or hospital\$) adj1 "use").tw.
29	("health care use" or "healthcare use").tw.
30	("medical care use" or "use of medical care").tw.
31	("use of health care" or "use of healthcare").tw.
32	("health service\$ use" or "use of health service\$").tw.
33	("clinic use" or "use of clinic#").tw.
34	("hospital\$ use" or "use of hospital\$" or "emergency use" or "use of emergency").tw.
35	"resource use".tw.
36	"use of resource\$".tw.

37	((health care or healthcare) adj1 visit\$.tw.
38	(utilization adj3 (resource\$ or healthcare or health care or medical care or hospital\$ or emergency or service\$)).tw.
39	(usage adj3 (resource\$ or healthcare or health care or medical care or hospital\$ or emergency or service\$)).tw.
40	or/27-39
41	26 or 40
42	exp Questionnaires/
43	exp Interviews as Topic/
44	Health Care Surveys/
45	Data Collection/
46	Self Disclosure/
47	Self-Assessment/
48	Self Report/
49	Mental Recall/
50	Self Care/
51	(self adj1 (report\$ or disclos\$ or record\$)).tw.
52	(patient\$ adj1 (report\$ or disclos\$ or recorded or recall)).tw.
53	(patient completed or completed by patient\$).tw.
54	self assess\$.tw.
55	(patient assess\$ or assess\$ by patient\$).tw.
56	(questionnaire\$ or survey\$ or interview\$ or diary or diaries).tw.
57	or/42-56
58	57 and 41
59	"reproducibility of results"/
60	accuracy.tw.
61	reliability.tw.
62	reliable.tw.
63	valid\$.tw.
64	exp utilization review/
65	recall bias.tw.
66	internal consisten\$.tw.
67	precision.tw.
68	test-retest.tw.
69	missing data.tw.

70	"Bias (Epidemiology)"/
71	gold standard.tw.
72	exp medical records/
73	records as topic/
74	registries/
75	or/59-74
76	58 and 75
77	limit 76 to english language
78	77 and 2012:2016.(sa_year).
79	exp Great Britain/
80	(national health service* or nhs*).ti,ab,in.
81	(english not ((published or publication* or translat* or written or language* or speak* or literature or citation*) adj5 english)).ti,ab.
82	(gb or "g.b." or britain* or (british* not "british columbia") or uk or "u.k." or united kingdom* or (england* not "new england") or northern ireland* or northern irish* or scotland* or scottish* or ((wales or "south wales") not "new south wales") or welsh*).ti,ab,jw,in.
83	(bath or "bath's" or ((birmingham not alabama*) or ("birmingham's" not alabama*) or bradford or "bradford's" or brighton or "brighton's" or bristol or "bristol's" or carlisle* or "carlisle's" or (cambridge not (massachusetts* or boston* or harvard*)) or ("cambridge's" not (massachusetts* or boston* or harvard*)) or (canterbury not zealand*) or ("canterbury's" not zealand*) or chelmsford or "chelmsford's" or chester or "chester's" or chichester or "chichester's" or coventry or "coventry's" or derby or "derby's" or (durham not (carolina* or nc)) or ("durham's" not (carolina* or nc)) or ely or "ely's" or exeter or "exeter's" or gloucester or "gloucester's" or hereford or "hereford's" or hull or "hull's" or lancaster or "lancaster's" or leeds* or leicester or "leicester's" or (lincoln not nebraska*) or ("lincoln's" not nebraska*) or (liverpool not (new south wales* or nsw)) or ("liverpool's" not (new south wales* or nsw)) or ((london not (ontario* or ont or toronto*)) or ("london's" not (ontario* or ont or toronto*)) or manchester or "manchester's" or (newcastle not (new south wales* or nsw)) or ("newcastle's" not (new south wales* or nsw)) or norwich or "norwich's" or nottingham or "nottingham's" or oxford or "oxford's" or peterborough or "peterborough's" or plymouth or "plymouth's" or portsmouth or "portsmouth's" or preston or "preston's" or ripon or "ripon's" or salford or "salford's" or salisbury or "salisbury's" or sheffield or "sheffield's" or southampton or "southampton's" or st albans or stoke or "stoke's" or sunderland or "sunderland's" or truro or "truro's" or wakefield or "wakefield's" or wells or westminster or "westminster's" or winchester or "winchester's" or wolverhampton or "wolverhampton's" or (worcester not (massachusetts* or boston* or harvard*)) or ("worcester's" not (massachusetts* or boston* or harvard*)) or (york not ("new york*" or ny or ontario* or ont or toronto*)) or ("york's" not ("new york*" or ny or ontario* or ont or toronto*))))).ti,ab,in.
84	(bangor or "bangor's" or cardiff or "cardiff's" or newport or "newport's" or st asaph or "st asaph's" or st davids or swansea or "swansea's").ti,ab,in.
85	(aberdeen or "aberdeen's" or dundee or "dundee's" or edinburgh or "edinburgh's" or glasgow or "glasgow's" or inverness or (perth not australia*) or ("perth's" not australia*) or stirling or "stirling's").ti,ab,in.
86	(armagh or "armagh's" or belfast or "belfast's" or lisburn or "lisburn's" or londonderry or "londonderry's" or derry or "derry's" or newry or "newry's").ti,ab,in.
87	or/79-86
88	(exp africa/ or exp americas/ or exp antarctic regions/ or exp arctic regions/ or exp asia/ or exp australia/ or exp

	oceania/) not (exp great britain/ or europe/)
89	87 not 88
90	78 and 89

Table B.9: Search Strategy: DIRUM SCOPUS 2012-2016 21.10.16

1	TITLE-ABS-KEY (visits OR appointment* OR hospitali?ation* OR "Health Services/" OR "Drug Therapy/" OR Outpatients OR hospital readmission OR "ambulatory care" OR "professional practice" OR "patient referral" OR teleconsultation OR "hospital admission" OR hospitalization OR "outpatient department" OR "mass screening" OR "diagnostic imaging" OR "health service/" OR "health care delivery" OR "general practice" OR therapy/)
2	TITLE-ABS-KEY (utili?ation OR utili?ation OR ("valid* W/5 self report*"))
3	#1 AND #2
4	TITLE-ABS-KEY ("Drug Utilization/" OR ((medicine* OR medication* OR hospital*) W/1 use) OR ("health care use" OR "healthcare use") OR ("medical care use" OR "use of medical care") OR ("use of health care" OR "use of healthcare") OR ("health service* use" OR "use of health service*") OR ("clinic use" OR "use of clinic*") OR ("hospital* use" OR "use of hospital*") OR ("emergency use" OR "use of emergency") OR (("health care" OR healthcare) W/1 visit*) OR "resource use" OR "use of resource*" OR (utili?ation W/3 (resource* OR healthcare OR "health care" OR "medical care" OR hospital* OR emergency OR service*)) OR (usage W/3 (resource* OR healthcare OR "health care" OR "medical care" OR hospital* OR emergency OR service*)) OR "health care utilization" OR "hospital utilization" OR utilization OR (valid* W/5 "self report*"))
5	#3 OR #4
6	TITLE-ABS-KEY ("Questionnaires/" OR "Health Care Surveys/" OR "Self Disclosure/" OR "Self Report/" OR "Mental Recall/" OR "Self Care/" OR (self W/1 (report* OR disclos* OR record*)) OR (patient* W/1 (report* OR disclos* OR recorded)) OR ("patient completed" OR "completed by patient*") OR "self assess*" OR ("patient assess*" OR "assess* by patient*") OR (questionnaire* OR survey* OR diary OR diaries OR interview*) OR "self evaluation/" OR "recall/" OR "self repor*" OR "patient report*")
7	TITLE-ABS-KEY (accuracy OR reliability OR reliable OR valid* OR "utilization review/" OR "recall bias" OR "reliability/" OR "reproducibility/" OR "accuracy/" OR "validity/" OR "recall bias/" OR "information processing/" OR "medical record/" OR "register/" OR "medical record review/" OR (accura* OR reliab* OR reproduce* OR valid*) OR precision OR "test retest" OR "missing data" OR "gold standard")
8	#5 AND #6 AND #7
9	TITLE-ABS-KEY ("conference abstract")
10	#8 AND NOT #9
11	TITLE-ABS-KEY (Great Britain)
12	TITLE-ABS-KEY ("national health service*" OR nhs* OR (english AND NOT (published OR publication* OR translat* OR written OR language* OR speak* OR literature OR citation*)) W/5 english))
13	TITLE-ABS-KEY (gb OR "g.b." OR britain* OR (british* AND NOT "british columbia") OR uk OR "u.k." OR "united kingdom*" OR (england* AND NOT "new england") OR "northern ireland*" OR "northern irish*" OR scotland* OR scottish* OR ((wales OR "south wales") AND NOT "new south wales") OR welsh*)
14	TITLE-ABS-KEY (bath OR "bath's" OR ((birmingham AND NOT alabama*) OR ("birmingham's" AND NOT alabama*) OR bradford OR "bradford's" OR brighton OR "brighton's" OR bristol OR "bristol's" OR carlisle* OR "carlisle's" OR (cambridge AND NOT (massachusetts* OR boston* OR harvard*)) OR ("cambridge's" AND NOT (massachusetts* OR boston* OR harvard*)) OR (canterbury AND NOT zealand*) OR ("canterbury's" AND NOT zealand*) OR chelmsford OR "chelmsford's" OR chester OR "chester's" OR chichester OR "chichester's" OR coventry OR "coventry's" OR derby OR "derby's" OR (durham AND NOT (carolina* OR nc)) OR ("durham's" AND NOT (carolina* OR nc)) OR ely OR "ely's" OR exeter OR "exeter's" OR gloucester OR "gloucester's" OR hereford OR "hereford's" OR hull OR "hull's" OR lancaster OR "lancaster's" OR leeds* OR leicester OR "leicester's" OR (lincoln AND NOT nebraska*) OR ("lincoln's" AND NOT nebraska*) OR (liverpool AND NOT (new south wales* OR nsw)) OR ("liverpool's" AND NOT (new south wales* OR nsw)) OR ((london AND NOT (ontario* OR ont OR toronto*)) OR ("london's" AND NOT (ontario* OR ont OR toronto*)) OR manchester OR "manchester's" OR (newcastle AND NOT (new south wales* OR nsw)) OR ("newcastle's" AND NOT (new south wales* OR nsw)) OR norwich OR "norwich's" OR nottingham OR "nottingham's" OR oxford OR "oxford's" OR peterborough OR "peterborough's" OR plymouth OR "plymouth's" OR portsmouth OR "portsmouth's" OR preston OR "preston's" OR ripon OR "ripon's" OR salford OR "salford's" OR salisbury OR "salisbury's" OR sheffield OR "sheffield's" OR southampton OR "southampton's" OR st albans OR stoke OR "stoke's" OR sunderland OR "sunderland's" OR truro OR "truro's" OR wakefield OR "wakefield's" OR wells OR westminster OR "westminster's" OR

	winchester OR "winchester's" OR wolverhampton OR "wolverhampton's" OR (worcester AND NOT (massachusetts* OR boston* OR harvard*)) OR ("worcester's" AND NOT (massachusetts* OR boston* OR harvard*)) OR (york AND NOT ("new york*" OR ny OR ontario* OR ont OR toronto*)) OR ("york's" AND NOT ("new york*" OR ny OR ontario* OR ont OR toronto*))))
15	TITLE-ABS-KEY (bangor OR "bangor's" OR cardiff OR "cardiff's" OR newport OR "newport's" OR st asaph OR "st asaph's" OR st davids OR swansea OR "swansea's" OR aberdeen OR "aberdeen's" OR dundee OR "dundee's" OR edinburgh OR "edinburgh's" OR glasgow OR "glasgow's" OR inverness OR (perth AND NOT australia*) OR ("perth's" AND NOT australia*) OR stirling OR "stirling's" OR armagh OR "armagh's" OR belfast OR "belfast's" OR lisburn OR "lisburn's" OR londonderry OR "londonderry's" OR derry OR "derry's" OR newry OR "newry's")
16	#11 OR #12 OR #13 OR #14 OR #15
17	ALL ((Africa OR Americas OR Antarctic OR arctic OR Asia OR Australia OR Oceania) AND NOT Great Britain OR Europe)
18	#16 AND NOT #17
19	#10 AND #18
20	INDEX(embase)
21	#19 AND #20
22	Results limited manually to 2012-2016

## **Appendix C**

### **Chapter Five: Participant Information Leaflet and Consent Form**

## An Assessment of Data from Routine Sources Applied to a Randomised Controlled Trial

### Information Leaflet

*We would like to invite you to take part in a research study. Before you decide we would like you to understand why the research is being done and what it would involve.*

*This information leaflet will introduce the study. Please contact us to discuss any questions you may have or clarify anything that is not clear.*

### What is the purpose of the study?

All new medical treatments need to be studied to see how they compare to standard treatments for intended benefits and risks of side effects. ***The Study of Standard and New Antiepileptic Drugs (SANAD II)*** is an example of a medical study in which people with newly diagnosed epilepsy are given different treatments and followed over time to see which treatment is best.

In studies such as SANAD II the way in which we measure benefits and side effects of treatment can be time-consuming for participants and expensive. For example, participants may be asked to come back to hospital for a clinic appointment, be telephoned by the study team or asked to fill in questionnaires.

We plan to find out whether information routinely collected by the NHS and other organisations in the UK can be used to assess the benefits and side effects of treatments for epilepsy. When any patient attends their GP or hospital, information regarding this visit is recorded on NHS computers forming your ***electronic medical record***. These records are stored securely but shared between a number of organisations to inform your care and the wider functions of the NHS. In this research study we will compare the information collected during the SANAD II study with information recorded in your GP and hospital electronic medical records. This will help us assess if routinely recorded information from electronic medical records is useful when measuring the effects of the different treatments in the SANAD II study.

## Why have I been invited to take part in this study?

We would like to invite you to participate in this study as you are currently taking part in SANAD II.

## What will happen during the study?

If you are willing to take part in this study, we ask that you complete and return the attached consent form. Following this, you would not be asked to do anything else.

The study team will then retrieve information from your GP and hospital electronic medical records. The time period covered will be the date you entered SANAD II until the most recently available date. We will not be obtaining information for any time you are not taking part in SANAD II or retrieving information from your electronic medical records on an ongoing basis in the future. We will extract information on only one occasion.

We will assess if information from your electronic medical records can accurately measure seizures, possible side effects of treatment, hospital attendances and admissions and other factors relevant to patients with epilepsy. We will compare the information from electronic medical records to the information obtained in SANAD II to see if it is useful. In order to provide a thorough assessment of information in electronic medical records we will retrieve all of the information that may be relevant to SANAD II. This information is collected routinely in both GP and hospital electronic medical records for all patients in the UK and includes:

- **Clinical Information** such as details about any hospital attendances and admissions, results of tests, other medical conditions you may have or treatments that are prescribed
- **Information about Patients'** such as your age and lifestyle factors
- **Administrative Information** such as waiting times, details of any hospital referrals and methods of admission
- **Geographical Information** such as details about where you are treated and the area where you live

### ***For Patients in England:***

The study team will retrieve information from your electronic medical records from:

- **The Health and Social Care Information Centre** ([www.hscic.gov.uk](http://www.hscic.gov.uk))  
Who collect information in electronic medical records about your hospital care (***Hospital Episode Statistics***) recorded by hospital care providers via ***The Secondary Uses Service***, on behalf of the NHS
- For patients' living in the North West of England, your **General Practice Surgery**  
Who record information in electronic medical records about your GP care on behalf of the NHS

The study team will retrieve information from your hospital electronic medical records from the *Hospital Episode Statistics* recorded by *The Health and Social Care Information Centre* for the time period you have been taking part in SANAD II.

For patients living in the North West of England, the study team are working with ***North West EHealth*** ([www.nweh.org.uk](http://www.nweh.org.uk)) an organisation sponsored by Salford Royal NHS Hospital, Salford Clinical Commissioning Group and the University of Manchester. *North West EHealth*, on behalf of the study team will approach your GP to retrieve information from your GP electronic medical records and retrieve information from your hospital electronic medical records from the *Secondary Uses Service* for the time period you have been taking part in SANAD II.

To retrieve your electronic medical records, information to identify you, including your name, date of birth, NHS Number and SANAD II Study Number will be securely transferred to ***The Health and Social Care Information Centre***. For patients living in the North West of England, your details will also be securely transferred to the ***Secondary Uses Service*** and ***North West EHealth***.

### ***For Patients in Wales:***

The study team will retrieve information from your electronic medical records from:

- ***The Secure Anonymised Information Linkage (SAIL) Databank***  
([www.saildatabank.com](http://www.saildatabank.com)) an organisation based in Swansea University and funded by the Welsh government to make GP and hospital electronic medical records accessible for research

SAIL holds information from electronic medical records provided by the *The NHS Wales Informatics Service* ([www.wales.nhs.uk/nwis](http://www.wales.nhs.uk/nwis)), who record information about your GP and hospital care on behalf of the NHS. All information held by SAIL is de-identified, meaning SAIL will not know the identity of individuals.

To allow the study team to obtain information from electronic medical records in identifiable form for this study, your name, date of birth, NHS Number and SANAD II Study Number will be securely transferred to **The NHS Wales Informatics Service** who then transfer your *SAIL Code Number* and the *SANAD II Study Number* to SAIL. This allows SAIL to securely transfer your electronic medical records to the study team, identified by SANAD II Study Number, without knowing the identity of individuals.

### ***For All Patients:***

Information from your GP and hospital electronic medical records will be securely transferred to the University of Liverpool using an encrypted electronic transfer system.

Your electronic medical records will be stored on secure University of Liverpool computer servers which meet NHS data security standards. The SANAD II data team will receive the electronic medical records and then link all the information collected about you from your GP and hospital electronic medical records and information collected about you for the SANAD II study. Information from your electronic medical records will then be compared to the information collected about you for the SANAD II study. The SANAD II data team will remove your identifying details (name, date of birth, address, NHS Number) from your medical records before the members of the study team who will be making this comparison have access to the information.

Data protection and security is of paramount importance and all information collected about you during this study will be used in accordance with the Data Protection Act 1998. Personally identifiable information collected for this study will be deleted by 31/12/18, within 12 months of study completion. All of the organisations involved in this study have established systems in place for securely retrieving and transferring information from electronic medical records for research.

If you would like to discuss any aspect of this study with the research team, we would be pleased if you would contact us on the details listed below.

## **What are the possible disadvantages of taking part in this study?**

There are no significant, direct disadvantages or risks to participants during this study.

This study involves the collection of information from your electronic medical records that can identify you (personally identifiable information). As with all research studies, data security and confidentiality is of paramount importance and will be protected at all times, meeting NHS data security standards.

## **What are the possible benefits of taking part in this study?**

There are no expected significant, direct benefits to the participants of this study.

However, the information collected from electronic medical records may be found to contribute to the information collected as part of the SANAD II study and inform the analyses at the end of the SANAD II study.

## **What if there is a problem during the study?**

If you have any questions or concerns about any aspect of this study, we would encourage you to contact the study team. Should you wish to submit a formal complaint, we recommend following the NHS Complaints Procedure. Details can be obtained from the *NHS Patient Advice and Liaison Services*:

*PALS Office, The Walton Centre for Neurology and Neurosurgery NHS Trust  
Liverpool  
L9 7LJ  
Telephone: 0151 529 6100*

The University of Liverpool is the sponsor for this study and the professional indemnity insurance will apply as appropriate. If you are harmed by taking part in this research study, there are no special compensation arrangements in place. If harm occurs to you and is due to someone's negligence, you may have grounds for legal action for compensation against the treating NHS Trust or Hospital.

## **Will my taking part in this study be kept confidential?**

Yes, data security and confidentiality will be protected throughout the study. Only authorised members of the study team and authorised persons directly supporting the study from *The Health and Social Care Information Centre and Secondary Uses Service, NorthWest EHealth, NHS Wales Informatics Service* and *The Secure Anonymised Information Linkage Databank* will have access to information from your electronic medical records. Data protection and security is of paramount importance and all information collected about you during this study will be used in accordance with the Data Protection Act 1998.

## **What will happen to the results of this study?**

We hope to present the results of this study at academic conferences and publish in the medical and scientific literature in order to inform other researchers of our findings and help them perform medical studies more efficiently. Your confidentiality will be ensured at all times and individual participants will not be identified in any presentation or publication.

## **What will happen if I don't want to carry on with this study?**

Taking part in this study is voluntary. You may decide not to take part, or after an original decision to take part you may decide to withdraw from the study. You do not have to give a reason and it will not affect your participation in SANAD II or the standard of medical care you receive now or in the future.

If you initially decide to take part and then change your mind in the future no more information from your electronic medical records will be collected. All information collected up until this time will be included in the study unless you specifically request that it is not included. If you decide in the future you would not like to continue with the study we would ask that you contact the study team using the contact details provided below.

## **Who is performing this study?**

This study is funded by the *Medical Research Council Hubs for Trials Methodology Research* ([www.methodologyhubs.mrc.ac.uk/research](http://www.methodologyhubs.mrc.ac.uk/research)), a governmental body aiming to improve the way in which medical studies are performed to benefit patients, the public and researchers. The research is being led by the University of Liverpool and The Walton Centre NHS Foundation Trust, by the same research team that are leading the SANAD II study.

## **Who has reviewed this study?**

The aims, methods and ethics of this study have been reviewed and approved by the *Health Research Authority and The North of Scotland Research Ethics Committee*.

## **Contact Details**

For further information about the study, please contact the principal researcher:

Dr Graham Powell  
The Department of Molecular and Clinical Pharmacology  
University of Liverpool  
Telephone: 0151 529 5464

*Thank you for taking the time to read this information leaflet.*

*If you would be willing to take part in this study, please complete the consent form and return to the research team in the enclosed pre-paid, addressed envelope.*

## An Assessment of Data from Routine Sources Applied to a Randomised Controlled Trial: Consent Form

*If you are willing to take part in this study, please consider each statement below and sign with your initials. Please return one copy of the consent form in the enclosed pre-paid, addressed envelope and keep one copy for your records.*

I confirm I have read and understood the information leaflet (version 2.7, dated 28/01/16), had the opportunity to ask questions and if asked, had them answered satisfactorily.	Please Initial: <input style="width: 100%; height: 40px;" type="text"/>
I understand that my taking part is voluntary and that I am free to withdraw from the study at any time without giving a reason and without my care or legal rights being affected.	Please Initial: <input style="width: 100%; height: 40px;" type="text"/>
I understand and give permission for personally identifiable information relevant to this study to be collected from my GP and hospital electronic medical records held by my General Practice and held and maintained by <i>The Health and Social Care Information Centre, Secondary Uses Service</i> and <i>The NHS Wales Informatics Service</i> for the time I have been taking part in SANAD II. I give permission for the secure transfer of my name, date of birth and NHS Number to allow information from my electronic medical records to be collected.	Please Initial: <input style="width: 100%; height: 40px;" type="text"/>
I understand and give permission for authorised persons from <i>The Secure Anonymised Information Linkage Databank, North West E-Health</i> and the study team at the University of Liverpool to access information collected from my GP and hospital electronic medical records where it is relevant to taking part in this study.	Please Initial: <input style="width: 100%; height: 40px;" type="text"/>
I understand and give permission for members of the SANAD II study team in the University of Liverpool to link information from my GP and hospital electronic medical records and information collected during SANAD II. I give permission for the secure electronic transfer of information in my electronic medical records to the University of Liverpool.	Please Initial: <input style="width: 100%; height: 40px;" type="text"/>
I agree to take part in this study.	Please Initial: <input style="width: 100%; height: 40px;" type="text"/>

**Name:** \_\_\_\_\_

**Date of Birth:** \_\_\_\_\_

**Signature:**  
\_\_\_\_\_

**Date:** \_\_\_\_\_

## **Appendix D**

### **Chapter Eight: Assessment of Feasibility and Efficiency – Further Results and Publication**

Table D.1: Summary of Key Application Milestones

Routine Data Source	Summary of Key Application Milestones	Cost Structure
NHS Digital	<p><b>August 2015:</b> First request to review Participant Information Sheet (PIS) and Consent Form. Directed by enquiries desk to a member of the Data Access Request Service (DARS) and subsequently reportedly forwarded to the Information Governance Team.</p> <p><b>4<sup>th</sup> November 2015:</b> No feedback received. Second request to review PIS and Consent Form. Forwarded by enquiries desk to Data Access and Information Sharing Team (DAIS).</p> <p><b>23<sup>rd</sup> November 2015:</b> No feedback received. Third request. PIS and Consent Form discussed with a member of the DARS Team in person at a NHS Digital Engagement Event. Informed that a full application would be required in order for NHS Digital to provide feedback on the PIS and Consent Form. This was completed and submitted on 26<sup>th</sup> November.</p> <p><b>7<sup>th</sup> December 2015:</b> Response regarding PIS and Consent Form following a further email. Informative and useful teleconference with a member of the DARS Team, informing development of the PIS / Consent Form.</p> <p><b>22<sup>nd</sup> December 2015:</b> Feedback received via email from the DAIS Team, in response to the second request on 4<sup>th</sup> November. Subsequent teleconference arranged to discuss. Useful feedback received, largely in agreement with that received from the DARS Team on the 7<sup>th</sup> December.</p> <p><b>25<sup>th</sup> January 2016:</b> Final review requested from DARS Team prior to ethics and governance application, prompt feedback received.</p> <p><b>8<sup>th</sup> February 2016:</b> NHS Digital requested completion of a new application using the existing application process.</p> <p><b>29<sup>th</sup> February 2016:</b> Submission of full application.</p> <p><b>1<sup>st</sup> March 2016:</b> Informed via automated email that there will be a 2 week suspension of applications whilst the system is updated.</p> <p><b>March 2016:</b> Enquires desk contacted on multiple occasions, repeatedly reassured that application has been accepted.</p> <p><b>18<sup>th</sup> April 2016:</b> First formal acknowledgment of submission of application. NHS Digital advised that they are no longer accepting applications using the previous system and the DARS Online Portal application must be used.</p> <p><b>22<sup>nd</sup> April 2016:</b> Third full application submitted, via DARS Online portal.</p> <p><b>28<sup>th</sup> April 2016:</b> Application acknowledged.</p> <p><b>16<sup>th</sup> May 2016:</b> Review at 'pre-DAAG' meeting. Recommended for full review at DAAG meeting.</p> <p><b>24<sup>th</sup> May 2016:</b> DAAG meeting. Recommended for approval once caveats addressed.</p> <p><b>26<sup>th</sup> May 2016:</b> Caveats addressed, application updated and re-submitted to NHS Digital.</p> <p><b>7<sup>th</sup> June 2016:</b> DAAG Approved. Data Sharing Agreement provided.</p> <p><b>13<sup>th</sup> July 2016:</b> Hospital Episode Statistics (HES) data available for download.</p>	<p>Standard cost recovery structure applied:</p> <p><i>£1000 New application</i></p> <p><i>£900 Release fee</i></p> <p><i>£500 3 year agreement</i></p> <p><i>£300 Per dataset per year</i></p>

<p><b>The Secure Anonymised Information Linkage Databank (SAIL)</b></p>	<p><b>22<sup>nd</sup> April 2015:</b> First contact regarding application process and association with ADNR. Prompt response confirming a single ADNR application would also include access to SAIL datasets if required.  <b>June 2015:</b> Request for information regarding application process. Teleconference arranged. Request by SAIL for protocol to inform scoping procedure.  <b>7<sup>th</sup> July 2015:</b> SAIL protocol submitted, specific points clarified by SAIL.  <b>August 2015:</b> Request to review PIS and Consent Form. Forwarded to Information Governance officer for review.  <b>September 2015:</b> Response from Information Governance Officer with feedback on PIS and Consent Form. Scoping document received and verified.  <b>October 2015:</b> Further advice requested and received regarding PIS / Consent Form.  <b>January 2016:</b> Final review of PIS / Consent Form requested following alterations required for the other organisations, received.  <b>February 2016:</b> Submission of full application.  <b>March 2016:</b> Feedback received regarding application following internal review, alterations suggested prior to submission for formal review. Re-emailed with alterations.  <b>April 2016:</b> Application re-submitted for formal IGRP review.  <b>21<sup>st</sup> July 2016:</b> IGRP Approved.  <b>August 2016:</b> SAIL data available for download.</p>	<p>Standard cost recovery structure applied:</p> <p><i>£500 Base cost  £291 Data transfer to SAIL  £1455 Individual level data processing  £500 Data transfer</i></p>
<p><b>The Clinical Practice Research Network (CPRD)</b></p>	<p><b>November 2014:</b> First request regarding feasibility of the study, response received broadly confirming feasibility.  <b>August 2015:</b> Further contact regarding feasibility and quote. Estimated quote received. Informed by CPRD that the Confidentiality Advisory Group and Ethical approvals need to be updated to permit identifiable, linked data release and the timelines to resolve these are unclear. Furthermore, informed that compliance with NHS Digital's governance framework needs to be approved. No further contact as the issues with linked data release, cost and population coverage make CPRD unfeasible for this study.</p>	<p>Standard cost recovery structure applied:</p> <p><i>£7500 CPRD GOLD for &lt;1000 patients  £4250 Linked HES inpatient  £850 Linked HES Outpatient  £3-5000 Extraction, specification, assurance</i></p>
<p><b>QResearch   The Health Improvement Network (THIN) Database   ResearchOne</b></p>	<p><b>September 2015:</b> All organisations contacted via email or telephone call. Confirmed that data is de-identified only, with no facility to re-identify patients as would be needed for this study. Data sources are therefore not feasible for inclusion in this study.</p>	<p>N/A</p>
<p><b>NorthWest eHealth</b></p>	<p><b>October 2015:</b> First contact regarding study. Telephone call to discuss the study outline and confirm feasibility in principle.  <b>November 2015:</b> Correspondence via email to request review of the updated protocol and PIS / Consent Form. Further correspondence to determine the methodology, refine the protocol and determine provisional costings. Advice also provided on PIS / Consent Form. Face to face meeting at the end of November clarified the methodological details. All feedback and correspondence prompt and detailed.  <b>December 2016:</b> Correspondence via email and telephone discussion with Apollo regarding the development of the data query to allow the extraction of data. Response received confirming the existing data query can be used for GP practices in Salford already holding a data sharing agreement with NWEH.  <b>January 2016:</b> Final review of PIS / Consent Form requested and received.  <b>May 2016:</b> &lt;5 participants consented to inclusion in the study are registered in eligible GP practices, therefore accessing data through NWEH is not cost effective for this study.</p>	<p>Bespoke NWEH costing:  <i>£11027 Data handling  £1575 Data check  £1326 Project manager</i></p> <p>Apollo Medical costing:  <i>£7200 Data query development</i></p> <p>CK Aspire costing:  <i>£6800 GP Recruitment</i></p>

<p><b>The Driver and Vehicle Licensing Agency (DVLA)</b></p>	<p><b>October 2014:</b> Multiple points of contact (emails and telephone calls) regarding the feasibility of accessing DVLA data for this study. No response received.</p> <p><b>February 2015:</b> Following discussion with a member of a DVLA expert committee, a DVLA medical advisor was contacted. The study was discussed with the DVLA Data Sharing Team and the response indicated that the DVLA would not have the capacity to assist with the study and the data security requirements are 'over and above the NHS or University'.</p>	<p>N/A</p>
<p><b>The Department for Work and Pensions (DWP)</b></p> <p><b>HM Revenue and Customs (HMRC)</b></p>	<p><b>November 2014:</b> First contact via email regarding feasibility of accessing DWP and HMRC data for this study. The email was forwarded to the DWP External Data Sharing and Advice Centre.</p> <p><b>December 2014:</b> External Data Sharing Advice Centre responded. Data access directly with the DWP or HMRC would not be possible and my query should be redirected to the ADRN.</p>	<p>N/A</p>
<p><b>The Administrative Data Research Network (ADRN)</b></p>	<p><b>December 2014:</b> First contact via email regarding feasibility for this study. No response received.</p> <p><b>Feb 2015:</b> Following a telephone call and further email acknowledgement of the query was received. General guidance provided via email and telephone conference arranged.</p> <p><b>March 2015:</b> Teleconference to discuss the study. ADRN confirmed that the study is eligible for their service and they can request access to DWP / HMRC linked to clinical datasets such as HES provided by NHS Digital. They would begin contacting the relevant data sources.</p> <p><b>April 2015:</b> Further teleconference, no significant progress.</p> <p><b>May 2015:</b> Further teleconference, HMRC have declined participation, DWP remains pending. I was informed that if the DWP do not permit access to their data I cannot apply through the ADRN solely for clinical datasets and independent applications must be submitted to the relevant organisations.</p> <p><b>July 2015:</b> Informed via email that the DWP have not been forthcoming but negotiations are on-going and they are unlikely to have a confirmed response until September. They provided no further feedback.</p>	<p>N/A</p>

RESEARCH

Open Access



# Using routinely recorded data in the UK to assess outcomes in a randomised controlled trial: The Trials of Access

G. A. Powell<sup>1\*</sup>, L. J. Bonnett<sup>2</sup>, C. Tudur-Smith<sup>2</sup>, D. A. Hughes<sup>3</sup>, P. R. Williamson<sup>2</sup> and A. G. Marson<sup>1</sup>

## Abstract

**Background:** In the UK, routinely recorded data may benefit prospective studies including randomised controlled trials (RCTs). In an on-going study, we aim to assess the feasibility of access and agreement of routinely recorded clinical and non-clinical data compared to data collected during a RCT using standard prospective methods. This paper will summarise available UK routinely recorded data sources and discuss our experience with the feasibility of accessing routinely recorded data for participants of a RCT before finally proposing recommendations for improving the access and implementation of routinely recorded data in RCTs.

**Methods:** Setting: the case study RCT is the Standard and New Antiepileptic Drugs II (SANAD II) trial, a pragmatic, UK, multicentre, phase IV RCT assessing the clinical and cost-effectiveness of antiepileptic drug treatments for newly diagnosed epilepsy.

**Participants:** 98 participants have provided written consent to permit the request of routinely recorded data.

**Study procedures:** routinely recorded clinical and non-clinical data were identified and data requested through formal applications from available data holders for the duration that participants have been recruited into SANAD II. The feasibility of accessing routinely recorded data during a RCT is assessed and recommendations for improving access proposed.

**Results:** Secondary-care clinical and socioeconomic data is recorded on a national basis and can be accessed, although there are limitations in the application process. Primary-care data are recorded by a number of organisations on a de-identified basis but access for specific individuals has not been feasible. Access to data recorded by non-clinical sources, including The Department for Work and Pensions and The Driving and Vehicle Licensing Agency, was not successful.

**Conclusions:** Recommendations discussed include further research to assess the attributes of routinely recorded data, an assessment of public perceptions and the development of strategies to collaboratively improve access to routinely recorded data for research.

**Trial registration:** International Standard Randomised Controlled Trials, ISRCTN30294119. Registered on 3 July 2012.

**EudraCT No:** 2012-001884-64. Registered on 9 May 2012.

**Keywords:** Routine data, Administrative data, Feasibility, Data collection

\* Correspondence: [gpowell@liverpool.ac.uk](mailto:gpowell@liverpool.ac.uk)

<sup>1</sup>Department of Molecular and Clinical Pharmacology, Clinical Sciences Centre, Lower Lane, Fazakerley, Liverpool L9 7LJ, UK  
Full list of author information is available at the end of the article

© The Author(s). 2017 Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

## Background

There is a plethora of individual-level, routinely recorded data in the UK. These data are recorded to fulfil specific, defined purposes and are regulated for security, confidentiality and disclosure by The Data Protection Act 1998 [1] and The Freedom of Information Act 2000 [2]. Access to routinely recorded data for 'secondary purposes', such as clinical research, is permitted providing that there is demonstrable secondary benefit.

The potential for routinely recorded data to inform clinical research and Health Technology Assessment (HTA) has long been recognised [3]. Presently, there are a number of sources of routinely recorded primary and secondary-care clinical data with regional or national coverage. However, limitations with accuracy of coding, confidentiality, ownership and data access have been previously identified as significant barriers to using routinely recorded data in research [4].

There are numerous examples of retrospective, observational, record-linkage population studies where routine sources have proved a valid and efficient method for providing data for clinical research [5]. In the context of prospective research, such as randomised controlled trials (RCTs), routinely recorded data have been used to inform judgements about the feasibility of sample size and recruitment targets [6] and measuring participant outcomes [3, 7]. Pragmatic cluster RCTs have been coordinated through routine data sources including patient recruitment, randomisation, and administration of intervention and trial assessments, such as through the Clinical Practice Research Datalink (CPRD) [8]. The majority of RCTs incur costs as clinicians assess participants, record outcomes and complete Case Report Forms – hence, using routinely recorded data may provide an efficient alternative method for data collection in addition to reducing the burden on participants. Furthermore, data from non-clinical routine sources may inform outcomes beyond the standard RCT assessments of clinical efficacy and effectiveness. For example, cost data (such as use of health care resources) and socio-economic data (such as employment and means-tested benefits data) may inform health economic analyses and the assessment of the broader societal impact of health care interventions.

The potential benefits of using routinely recorded data in clinical research have resulted in a political drive to increase implementation, detailed in *The Plan for Growth* [9] and *The NHS Constitution* [10], where research is presented as a core activity making the link explicit between the provision of NHS services and research. Consequently, initiatives, such as the Administrative Data Research Network [11], have been established to provide a method of access to individual-level data, linking clinical and non-clinical sources of routinely recorded data.

The objective of this paper is to review relevant sources of routinely recorded data for England, Scotland and Wales and to discuss our experience with the feasibility of accessing individual-level data for a subgroup of participants enrolled into a RCT before finally proposing recommendations for improving the access and implementation of routinely collected data in RCTs. This is an on-going study and in a future publication we aim to assess the agreement of routinely recorded data compared to paired data collected in a RCT using standard prospective methods.

## Methods

The case study RCT is the Standard and New Antiepileptic Drugs (SANAD) II trial. SANAD II is a pragmatic, UK, multicentre, phase IV RCT funded by the National Institute for Health Research (NIHR) Health Technology Assessment (HTA) programme, assessing the clinical and cost-effectiveness of a number of antiepileptic drugs as first-line treatments for newly diagnosed epilepsy. Data for clinical outcomes, including seizure freedom and adverse events, are recorded on Case Report Forms by the treating clinical team during outpatient appointments. Data to inform cost-effectiveness analyses, including health care resource use and quality of life, are recorded through participant completion of questionnaires. SANAD II is currently recruiting and is expected to report in 2019.

Following research ethics and governance approvals, 470 participants enrolled in SANAD II were invited to provide written consent to permit the request of routinely recorded data for the duration of their participation in SANAD II. Ninety-eight (20.9%) participants provided consent and were included in the study. Relevant sources of routinely recorded data were identified and detailed scoping discussions ensued. Subsequently, where accessible, routinely recorded data for participants recruited into SANAD II were requested through formal applications. The routinely recorded data sources included in this study are as follows:

- Clinical routine data sources: secondary care:
  - The Health and Social Care Information Centre (HSCIC)
  - The NHS Wales Informatics Service (NWIS)
  - The NHS National Services Scotland; Information Services Division (ISD)
- Clinical routine data sources: primary care:
  - The Clinical Practice Research Datalink (CPRD)
  - ResearchOne
  - QResearch
  - The Health Improvement Network (THIN) database
  - North West eHealth (NWEH)

- Non-clinical routine data sources:
  - The Office for National Statistics (ONS)
  - HM Revenue and Customs (HMRC)
  - The Department for Work and Pensions (DWP)
  - The Driver and Vehicle Licensing Authority (DVLA)
- 'Linked' routine data sources:
  - The Secure Anonymised Information Linkage (SAIL) databank
  - The Administrative Data Research Network (ADRN)

In a future publication, the agreement between routinely recorded data and data collected using standard prospective methods will be assessed for baseline variables such as gender, age and date of first seizure, and for outcome measures relevant to SANAD II such as time to 12-month remission from seizures. To assess agreement between paired continuous data, Bland-Altman methods will be employed. Acceptable clinical limits of agreement for each variable or SANAD II outcome will be specified a priori and compared to the 95% confidence limits of agreement. To assess agreement between paired, nominal categorical datasets, cross tabulations will be constructed followed by calculation of Cohen's Kappa.

## Results

### Clinical routine data sources: secondary care

Electronic medical records of patients' use of secondary-care services in the UK are routinely managed on a national basis. A number of public service organisations provide national information, data and IT systems for commissioners, analysts and clinicians in health and social care. Data are recorded to inform patient care, provide the data for remuneration for hospital trusts and are subsequently used to monitor and improve clinical services through clinical research. Table 1 summarises the data sources where access to individual-level data is possible.

### Clinical routine data sources: primary care

Electronic medical records of patients' use of primary-care services in the UK are recorded routinely by the general practitioner to inform patient care and remuneration, but are not currently available for clinical research on a national basis. A number of organisations represent collaborations between governmental bodies or academic institutions and providers of primary-care IT systems. Access on a regional basis is possible through a number of data sources summarised in Table 2.

### Non-clinical routine data sources

Non-clinical, individual-level data are routinely recorded by a number of UK governmental departments for a variety of indications. Selected organisations record data that

Table 1 Example sources of routinely recorded secondary-care data

*The Health and Social Care Information Centre (HSCIC) [21]*

*Data access for clinical research:*

*The Data Access Request Service provides a method of access to a number of routinely collected datasets for England. Hospital Episode Statistics (HES) provides clinical, health and socioeconomic data for all secondary-care attendances in England. Datasets include Accident and Emergency, Admitted Patient, Outpatient, Adult Critical Care, Maternity and selected Patient Reported Outcome Measures.*

*Previous experience in clinical research:*

HES data have been accessed for retrospective linkage studies [22] and to provide data for prospective studies; for example, estimation of health care resource use or measuring outcomes such as long-term mortality [23]

*The NHS Wales Informatics Service (NWIS) [24]*

*Data access for clinical research:*

Data access can be facilitated through *The Public Health Wales Observatory. The Patient Episode Database for Wales (PEDW) provides clinical, health and socioeconomic data for all secondary-care attendances in Wales and is broadly comparable to the Admitted Patient HES dataset, with data regarding elective and emergency admissions and maternity care recorded. Additional datasets of relevance to this study include the Emergency Department and Outpatient Datasets. Previous*

*experience in clinical research:*

PEDW data have been accessed for retrospective analyses; for example, analysis of the incidence of obstetric complication rates [25]

*The NHS National Services Scotland; Information Services Division (ISD) [26]*

*Data access for clinical research:*

*The electronic Data Research and Innovation Service (eDRIS) provides a method of access to ISD datasets including Outpatient, General Acute/ Inpatient, Emergency Department, Unscheduled Care, GP Out of Hours and The Prescribing Information System. Clinical, health and socioeconomic data are recorded and datasets are largely comparable to HSCIC HES. Previous*

*experience in clinical research:*

ISD data have been accessed for retrospective linkage studies; for example, analysis of the incidence of gastrointestinal bleeding and complications including mortality [27]

would be informative to prospective clinical research in epilepsy and other diseases, summarised in Table 3.

### 'Linked' routine data sources

In order to provide a 'complete' dataset of the information required to meet research objectives, data from a number of organisations may need to be accessed. This is typically accomplished by linking data sources using identifiers such as patients' name, date of birth, National Insurance number or NHS number. In response to the growing recognition of the potential of routinely recorded data, initiatives have been established to assist with the provision of linked, de-identified, aggregate data between data sources:

- *The Secure Anonymised Information Linkage (SAIL) Databank* is an initiative developed by Swansea University and funded by the Welsh Government. SAIL provides a method of access to individual-level, routinely recorded, de-identified electronic data for patients across Wales to support research [12]. Access to clinical datasets provided by NWIS is

Table 2 Example sources of routinely recorded primary-care data

*The Clinical Practice Research Datalink (CPRD) [28]**Data access for clinical research:*

CPRD is a governmental research service jointly funded by the NHS National Institute for Health Research and the Medicines and Healthcare products Regulatory Agency. Following approval by the *Independent Scientific Advisory Committee*, CPRD provides access to de-identified primary-care clinical, health and socioeconomic data for a geographically representative 13 million patients in England for health care research.

*Previous experience in clinical research:*

CPRD data have been used in retrospective studies for estimating health care resource use, prescription medicines and clinical outcomes [22]. Gulliford conducted two cluster-randomised trials using CPRD: one aimed to reduce inappropriate antibiotic prescribing for acute respiratory infection; the other aimed to increase physician adherence with secondary prevention interventions after first stroke [8]

*ResearchOne [29]**Data access for clinical research:*

ResearchOne is a collaboration between The University of Leeds and The Phoenix Partnership (TTP), developers of the SystmOne clinical database and IT system. De-identified clinical, health and socioeconomic data are available from primary, secondary and out-of-hours care settings for approximately 26 million patients in the UK.

*Previous experience in clinical research:*

ResearchOne data have been used in public health surveillance studies, retrospective studies [29] and, currently, in combination with CPRD data to measure the outcomes of a cluster RCT [30]

*QResearch [31]**Data access for clinical research:*

QResearch is a collaboration between The University of Nottingham and the developers of the EMIS IT systems. De-identified clinical, health and socioeconomic data are available for approximately 18 million patients in the UK.

*Previous experience in clinical research:*

QResearch data have been used to measure clinical outcomes in case-control and cohort studies [32]

*The Health Improvement Network (THIN) Database [33]**Data access for clinical research:*

THIN is a collaboration between IMS Health and In Practice Systems, developers of the IT software Vision. De-identified clinical, health and socioeconomic data are available for approximately 11.1 million patients in the UK.

*Previous experience in clinical research:*

THIN data have been accessed to measure clinical outcomes in cohort and case-control studies [34]

*North West eHealth (NWEH) [35]**Data access for clinical research:*

NWEH is a collaboration between The University of Manchester, Salford Royal Foundation Trust and Salford Clinical Commissioning Group. NWEH has developed the methodology and governance framework to implement the *Salford Integrated Record*, an integrated primary- and secondary-care electronic medical record, into research as part of the Salford Lung Study [14]. The infrastructure permits access to secondary-care electronic medical records accessed through the HSCIC *Secondary Uses Service*. With participant and GP practice enrolment and consent, the Apollo [36] and Graphnet [37] data-extraction tools are employed to extract participant primary-care electronic medical records that can then be linked to data regarding secondary care. North West eHealth is unique in that data are not de-identified and, therefore, participant consent is required. Furthermore, GP practice enrolment and consent is required to permit the installation of third-party software on their systems and subsequent extraction of data.

*Previous experience in clinical research:*

NWEH offers a number of primary-care research tools including a randomised controlled trial (RCT) recruitment feasibility assessment, but does not currently routinely provide a bespoke primary-care data-extraction service for research. However, the methodology for this process has been demonstrated [14]

Table 3 Example sources of routinely recorded non-clinical data

*The Office for National Statistics (ONS) [38]**Data access for clinical research:*

The ONS records individual-level mortality data and aggregate economic and societal statistics that may inform clinical and health economic analyses. Mortality data can be requested through application to the HSCIC DARS. Aggregate data can be accessed via services provided by ONS such as NOMIS [39] and Data for Neighbourhoods and Regeneration [40]. The smallest reported level is the Lower Layer Super Output Area (LSOA) consisting of a population of 1000–3000.

*Previous experience in clinical research:*

ONS mortality data have been accessed to measure mortality in retrospective and prospective studies [23]

*HM Revenue and Customs (HMRC) [41]**Data access for clinical research:*

HMRC is the UK's national tax authority and responsible for taxation including National Insurance and student loan repayments and the administration of tax credits, child benefit and statutory sick and maternity pay. Individual-level data on employment and tax contributions are recorded and likely to inform health and socioeconomic analyses. The *HMRC Datalab* provides a means to access de-identified, aggregate HMRC data for research. An application, once 'approved researcher' status has been gained, must benefit the listed functions of the HMRC.

*Previous experience in clinical research:*

There was no evidence of individual-level, HMRC data being accessed for clinical research in a scoping search performed in MEDLINE via OVID

*The Department for Work and Pensions (DWP) [42]**Data access for clinical research:*

The DWP is responsible for welfare including the provision of state pensions, benefits and child maintenance. Individual-level data regarding employment and welfare are likely to inform health and socioeconomic analyses and de-identified, aggregate data are available for social research.

*Previous Experience in Clinical Research:*

There was no evidence of individual-level, DWP data being accessed for clinical research in a scoping search performed in MEDLINE via OVID

*The Driver and Vehicle Licensing Authority (DVLA) [43]**Data access for clinical research:*

The DVLA is responsible for the licensing of drivers and vehicles in the UK and issuing, reviewing and maintaining guidance regarding driving licence status in the context of medical diagnoses. The legal requirement for driving licence holders to inform the DVLA of the occurrence of seizures and, subsequently, to regain normal driving privileges after a specified period of seizure freedom raises the possibility of DVLA providing an accurate data source to inform the clinical outcome measures in epilepsy research.

*Previous experience in clinical research:*

The DVLA publish limited de-identified, aggregate datasets for research, usually involving driving restrictions. There was no evidence of individual-level, DVLA data being accessed for clinical research in a scoping search performed in MEDLINE via OVID

complemented with numerous non-clinical administrative datasets including births, deaths and demographic data. Following the scoping process a formal application is submitted to the *Information Governance Review Panel* before access to data is granted. SAIL data have been accessed to measure clinical outcomes in retrospective research [13]

- *The Administrative Data Research Network (ADRN)* is a UK-wide partnership between universities, government departments, national statistics authorities, funders and researchers, funded by the *Economic and Social Research Council*. ADRN provides a method of access to a number of

non-clinical administrative routine datasets including employment, socioeconomic, crime and education data [11] in addition to clinical datasets detailed previously such as those recorded by HSCIC. Following development of a project proposal a formal application is reviewed by the *Approvals Panel* before access to data is granted

#### Challenges and feasibility of access

We have requested access to routinely recorded data for individuals enrolled in the SANAD II RCT, resident in England and Wales, who have provided written consent. There were insufficient participants meeting the eligibility criteria resident in Scotland. Data sources were identified and scoping discussions informed the initial assessment of feasibility. Data sources were deemed feasible if individual-level data could be provided for specified individuals providing consent. Resources required including cost and researcher time were also factors important in the assessment of feasibility. Including the preparation, research ethics and governance approval and submission of the applications for data access, significant researcher time and a period of 18 months were required. The feasibility, timeline and key milestones involved for each data source are summarised in Table 4.

#### *Clinical routine data sources*

Routinely recorded secondary-care data can be requested on an individual-level, identifiable basis for patients in England and Wales through HSCIC and NWIS, accessed through SAIL and in our experience this process is feasible as part of a RCT, yet there are notable limitations. In England, HSCIC has set a target time to data access of sixty working days following submission for a complex application, involving bespoke data linkage from multiple datasets. From the date of submission of the Data Access Request Service online application, we have been granted access to the data within this timeframe. However, this positive experience following submission of the application is countered by limitations in the pre-application process. Acknowledging the significant update to online application and approval procedures that occurred during this period, there remains a considerable period of time required in the development of the application. The nature of the request for identifiable data necessitated participant consent as the valid legal basis. HSCIC require ethical and governance approval to be in place prior to DARS review and to prevent future amendments and delays, it was rational to ensure the consent materials had been reviewed by the HSCIC's Information Governance Team, prior to submitting the documents for ethical and governance approval. HSCIC provide written guidance regarding the consent materials and advise that

documents should be reviewed. However, in our experience there is no formalised process for providing this review. Following significant correspondence the consent materials were reviewed by the Data Access and Information Sharing Team. However, this feedback was provided following a formal submission and review by the Data Access Request Service. Formalising the process for the review of consent materials would likely improve the time and resource efficiency for both HSCIC and the researcher.

For participants in Wales, we have requested secondary-care data and, for a proportion of participants, primary-care data through SAIL databank. SAIL provided a streamlined pre-application service, including engaging in multiple discussions and completion of a scoping document outlining the study methods and costs involved. Consent materials were also promptly reviewed by a member of the Information Governance Team.

Common to both sources of secondary-care, routinely recorded data; there are stringent information governance requirements that must be in place prior to application. These include information security measures and assessments, specific inclusion regarding the 'processing of health care data for the subjects of research' in the institutional Data Protection Act registration and, in the case of HSCIC, an institutional Data Sharing Framework Contract. Adequate guidance is provided by the data sources and, if not addressed by the researcher, may cause delay. Furthermore, there is a time lag of approximately 3–6 months before data become available within each data source. This delay potentially limits the utility of such sources in prospective clinical research, such as drug trials, where prompt reporting is clinically important and a regulatory requirement.

Routinely recorded primary-care data for specific participants in England are less accessible. The majority of providers of primary-care data, such as ResearchOne and QResearch, provide data on a de-identified basis with no facility to re-identify individuals. Therefore, where specific participants need to be identified, as for RCTs such as SANAD II, these sources are not applicable. Following our correspondence, CPRD confirmed it may be possible to retrieve identifiable individual-level data linked to HSCIC data in the future, but the required approvals were not in place and the timescale to resolution was unclear. Furthermore, such primary-care sources provide data for only a proportion of the population and can be expensive. North West eHealth employs an alternative methodology whereby primary-care data are extracted directly from the GP through a third party. This process requires participant and GP consent and installation of the required software but is an effective data-extraction method [14]. NWEH offers a number of primary-care research

Table 4 Summary of key application milestones

Routine data source	Summary of key application milestones	Cost structure
The Health and Social Care Information Centre (HSCIC)	<p><i>August 2015</i>: first request to review Participant Information Sheet (PIS) and Consent Form. Sent by enquiries desk to the Data Access Request Service (DARS)</p> <p><i>4 November 2015</i>: second request to review PIS and Consent Form. Sent by enquiries desk to Data Access and Information Sharing Team (DAIS)</p> <p><i>23 November 2015</i>: no feedback yet received. PIS and Consent Form discussed with a member of the DARS team in person at a HSCIC engagement event. Informed that a full, formal application would be required in order for HSCIC to provide feedback on the PIS and Consent Form. This was completed and submitted on 26 November</p> <p><i>7 December 2015</i>: response regarding PIS and Consent Form. Informative teleconference with a member of the DARS team</p> <p><i>22 December 2015</i>: response from the DAIS team in response to the second request on 4 November 2015. Teleconference provided feedback, in agreement with that received from the DARS team on 7 December</p> <p><i>29 February 2016</i>: as directed by HSCIC, submission of a new formal application using the existing application process</p> <p><i>18 April 2016</i>: formal acknowledgment of submission. Requested to submit the application via the DARS Online Portal</p> <p><i>22 April 2016</i>: formal application submitted via DARS Online Portal</p> <p><i>24 May 2016</i>: Data Access Advisory Group (DAAG) review. Caveats to be addressed before approval</p> <p><i>26 May 2016</i>: caveats addressed, application updated and re-submitted</p> <p><i>13 July 2016</i>: DAAG approved. Hospital Episode Statistics (HES) data available for download</p>	<p>Standard cost recovery structure applied:</p> <p>£1000 new application</p> <p>£900 release fee</p> <p>£500 3-year agreement</p> <p>£300 per dataset per year</p>
The Secure Anonymised Information Linkage Databank (SAIL)	<p><i>22 April 2015</i>: first contact regarding application process and association with the Administrative Data Research Network (ADRN)</p> <p><i>June 2015</i>: informative teleconference regarding the SAIL application process and scoping procedure</p> <p><i>7 July 2015</i>: protocol regarding methods specific to SAIL submitted</p> <p><i>August 2015</i>: request to review PIS and Consent Form. Sent to information governance officer for review</p> <p><i>September 2015</i>: feedback on PIS and Consent Form from information governance officer. Scoping document issued by SAIL</p> <p><i>January 2016</i>: final review of PIS/Consent Form requested following revisions required for the other data sources</p> <p><i>February 2016</i>: submission of full, formal application</p> <p><i>March 2016</i>: feedback received following internal review with amendments suggested</p> <p><i>April 2016</i>: application re-submitted for formal Information Governance Review Panel (IGRP) review, outcome pending</p>	<p>Standard cost recovery structure applied:</p> <p>£500 base cost</p> <p>£291 data transfer to SAIL</p> <p>£1455 individual-level data processing</p> <p>£500 data transfer</p>
The Clinical Practice Research Network (CPRD)	<p><i>November 2014</i>: first contact regarding feasibility of the study, response received broadly confirming feasibility</p> <p><i>August 2015</i>: following protocol development, further contact regarding feasibility. Informed by CPRD that the Confidentiality Advisory Group and ethical approvals with HSCIC need to be updated to permit identifiable, linked data release and the timelines to resolve these are unclear. Furthermore, informed that compliance with HSCIC's governance framework needs to be approved. No further contact as the issues with linked data release, cost and population coverage make CPRD not feasible for inclusion in this study</p>	<p>Standard cost recovery structure applied:</p> <p>£7500 CPRD GOLD for &lt;1000 patients</p> <p>£4250 linked HES inpatient</p> <p>£850 linked HES outpatient</p> <p>£3000–5000 extraction, specification, assurance</p>

Table 4 Summary of key application milestones (*Continued*)

QRResearch   ResearchOne The Health Improvement Network (THIN) Database	<p><i>September 2015</i>: all organisations contacted. Confirmed that data are de-identified only, with no facility to re-identify patients as would be needed for this study. Data sources are, therefore, not feasible for inclusion in this study</p>	N/A
North West eHealth	<p><i>October 2015</i>: first contact, the service is not routinely offered but feasibility of the process broadly confirmed <i>November 2015</i>: correspondence via email to request review of the protocol, PIS and Consent Form, confirm the methodology and determine provisional costings. Further discussion during a face-to-face meeting at NWEH <i>December 2016</i>: discussion with the third party, Apollo Medical Software Solutions, regarding the development of the data query to permit the extraction of data. Response received confirming the structure of the existing data query can be used for GP practices in Salford already holding a data-sharing agreement with NWEH, but a bespoke query would be required for this study <i>January 2016</i>: final review of PIS/Consent Form requested and received <i>May 2016</i>: &lt; participants consented to inclusion in the study are registered in eligible GP practices; therefore, accessing data through NWEH is not feasible for this study</p>	<p>Bespoke NWEH costing: £11027 data handling £1575 data check £1326 project manager Apollo Medical costing: £7200 data query development CK Aspire costing: £6800 GP recruitment</p>
The Driver and Vehicle Licensing Agency (DVLA)	<p><i>October 2014</i>: multiple attempts at contact to discuss the feasibility of the study, including telephone calls and email correspondence. No response received <i>February 2015</i>: following discussion with a member of a DVLA expert committee, the DVLA medical advisor was contacted. The study was discussed with the DVLA data-sharing team and the response indicated that the DVLA would not have the capacity to assist with the study and the data-security requirements are 'over and above the NHS or university'</p>	N/A
The Department for Work and Pensions (DWP) HM Revenue and Customs (HMRC)	<p><i>November 2014</i>: first contact regarding feasibility of accessing DWP and HMRC data for this study. Request transferred to the DWP External Data Sharing and Advice Centre <i>December 2014</i>: External Data Sharing Advice Centre responded. Data access directly with the DWP or HMRC would not be possible and my request should be redirected to ADRN</p>	N/A
The Administrative Data Research Network (ADRN)	<p><i>December 2014</i>: first contact regarding feasibility for this study. No response received <i>Feb 2015</i>: further contact regarding feasibility of the study. General information provided via email <i>March 2015</i>: informative teleconference to discuss the study. ADRN confirmed that the study is eligible for their service and they can request access to the DWP/HMRC linked to clinical datasets, such as HES, provided by HSCIC. They agreed to contact the relevant data sources to determine the feasibility <i>April 2015</i>: further teleconference, no significant progress <i>May 2015</i>: further teleconference, HMRC have declined participation, the DWP remains pending. I am informed that if the DWP does not permit access to its data I cannot apply through ADRN solely for clinical datasets and independent applications must be submitted to the relevant organisations such as HSCIC <i>July 2015</i>: informed that the DWP have not been forthcoming but negotiations are on-going and they are unlikely to have a confirmed response until September. No further feedback received</p>	N/A

tools for the wider research community but does not currently routinely provide a bespoke primary-care data-extraction service for research.

#### *Non-clinical routine data sources*

Aggregate economic and societal statistics, provided by Lower Layer Super Output Area (LSOA), can be accessed through the ONS and are in the public domain. Such data may have additional benefits to the analyses of health and socioeconomic outcomes in RCTs. Individual-level, economic data from sources such as the DWP and HMRC would likely be informative to prospective clinical research as such data are often poorly or incompletely recorded using standard methods [15]. However, relevant to this study, there is no previous evidence of access to DWP or HMRC individual-level or aggregate data for clinical research.

During scoping discussions with DWP and HMRC, we were directed to ADRN but this network has not been successful in negotiating data access.

Finally, the outcomes of selected clinical studies may be measured using DVLA data. However, the DVLA declined the request for access, citing insufficient internal resources to process the request and more stringent data protection requirements than those employed in the NHS or academic institutions, without providing explicit details regarding these requirements.

#### Discussion

Routinely recorded data are valid for use in retrospective clinical research [3, 4] and have the potential to be used in prospective research including measuring the outcomes of RCTs [7] and providing additional benefits such as a method to address missing RCT data. Limitations, specifically with respect to accuracy and access have been recognised for some time. Academic, political [9] and health service [10] interest in UK sources of routinely recorded data has resulted in expansion and improvements, notably in the access to linked datasets. However, our experience with accessing individual-level data for specific participants providing written consent, to inform the outcomes of a RCT, highlights persisting limitations.

Clinical routine data sources are numerous and there is comprehensive national coverage of secondary-care data. In our experience, accessing individual-level data is feasible. However, inefficiencies in the application processes persist, particularly during the informal 'pre-application' phase. The notable limitation encountered was obtaining feedback on the Patient Information Sheet and Consent Form prior to ethical and governance review. Formalising an explicit review process for consent materials would improve the efficiency for both the data holders and the research team.

#### Table 5 Recommendations to improve access to routinely recorded data for research

##### *General*

Routinely recorded data are being used to measure randomised controlled trial (RCT) outcomes with the agreement, additional benefits and cost-efficiency of such data compared to data recorded through standard RCT methods being unknown

*Further research should be performed to assess the agreement, additional benefits and cost-efficiency of accessing routinely recorded data to measure RCT outcomes compared to data collected through standard RCT methods*

The costs required for data access from routine data sources vary widely, although all reportedly operate on a cost recovery, not-for-profit basis

*Costs should be standardised and rationalised between routine data sources*

The time lag before data are available in routine data sources represents a significant limitation to the access of routinely recorded data for prospective research, including RCTs

*The infrastructure and procedures should be developed to reduce the time lag seen in routinely recorded data sources*

The requirement for linkage between sources of routinely recorded data has been observed and improvements are on-going; for example, with the establishment of the Administrative Data Research Network (ADRN) A standardised set of identifying variables could be recorded by all (clinical and non-clinical) data sources to improve the accuracy of data linkage, similar to a Core Outcome Set for clinical trials [44]

The public mistrust in the sharing and linking of routinely recorded data will hamper future efforts to develop routinely recorded databases, despite the likely benefits to individual patients and the population

*Further research and public engagement should be undertaken to define the issues of most importance to the public and develop strategies to address these*

##### *Clinical routine data sources*

There are numerous requirements prior to application, and criteria to fulfil on submission, of an application, yet the guidance and support during development of an application remains limited

*Formalise and improve access to guidance and review of study materials during the 'pre-application stage'*

There is national coverage of routinely recorded secondary-care data, yet primary-care coverage remains patchy, based on geographical area or GP IT system

*Develop the primary-care data sources to provide national coverage, either through collaboration of existing sources and data linkage or development of national data sources, such as the General Practice Extraction Service*

##### *Non-clinical routine data sources*

Access to non-clinical data sources to inform clinical research was not possible during this study, despite the significant potential to inform Health Technology Assessment and the increasing importance of such assessments in a health care system where resources are increasingly limited

*To assist with Health Technology Assessment, and particularly the analysis of health economic outcomes, urgent research is required to consider facilitating access to individual-level, identifiable data from non-clinical sources. This would include:*

- 1. Research regarding the public perception and acceptability of using their personal economic data for clinical research*
- 2. Internal review within non-clinical sources, such as the DWP and HMRC, to assess the feasibility and limitations of permitting access to data for clinical research*
- 3. Formalisation of the approval processes through the independent party, the ADRN for access to non-clinical administrative data – currently, following internal approval the ADRN then negotiates access to administrative data on a project-by-project basis*

Access to routinely recorded, individual-level, primary-care data has not been feasible. Each primary-care data source has limited geographical coverage, often based on GP IT systems, which usually process de-identified data and may incur significant expense. The inception of the HSCIC *General Practice Extraction Service*, which records primary-care data nationally for England, represents the most optimistic national source; however, access is currently restricted to Department of Health initiatives such as research involving screening procedures [16].

The access to non-clinical data sources for clinical research has not been possible. ADRN has been established to act on behalf of the researcher in negotiating access to de-identified, linked, routinely recorded data from a number of organisations and the study proposal was promptly directed to ADRN. However, the decision whether to release data remains with the data holder. Ideologically, the next step would be the storage of de-identified linked data from participating organisations in a single repository, similar to those established for RCT data [17]. This would create a single point of access and remove the burden for each organisation to consider each study individually. This would, however, require significant information governance and security barriers to be cleared and, in light of recent developments within the research climate, individual consent. Including patients as stakeholders in the development of such data sources is essential [18].

Although there are examples of pragmatic RCTs being coordinated through routine data sources [8], there are likely to be limitations when accessing routinely recorded data to measure the outcomes of RCTs. Quality assurance is unclear and the level of agreement of routinely recorded data with data recorded through standard RCT methods remains uncertain, particularly when measuring clinical outcomes. The time delay before routinely recorded data become available may have implications for RCTs where prompt reporting is both clinically important and a regulatory requirement. Furthermore the pre-application and application process may introduce further delays. This will have implications for RCTs relying on routinely recorded data. The cost-efficiency of accessing routinely recorded data, compared to standard methods, is unclear. Further research is required to assess the agreement, additional benefits and cost-efficiency of routinely recorded data compared to data collected through standard RCT methods; it may be in the additional benefits, such as addressing missing RCT data, where routinely recorded data is most useful.

## Conclusions

The failure of access to routinely recorded data for a purpose, such as this study with clear secondary benefit

to clinical research methodology, seems inappropriate when the 'public purse' funds the research, the researcher and the public body holding the data. Perhaps a significant cause or contributor to the current limitations is the Care.Data initiative in 2014. The proposal to extract primary-care records from all patients was opposed publicly by a number of groups and, for example, resulted in an internal inquiry within HSCIC. Data applications were suspended during this period and our current experience may be explained by the concurrent revision of the HSCIC application and approval procedures. However, in the medium term, of more concern is the harm in public perception that resulted. Currently, more than 1.2 million individuals in the UK have submitted a 'Type 2 objection', meaning that their data will not be shared for purposes other than direct care [19]. Although the application procedures may improve, and in time we may be able to access data more efficiently, the loss of 2.2% of the population's data will have implications for the routinely recorded data that will then be made available for research. Involving patients as important stakeholders and re-gaining their trust will be an essential factor in realising the individual and population health care benefits of routinely recorded data [20].

## Recommendations

We propose recommendations to improve access and implementation of routinely recorded data during a RCT, summarised in Table 5.

### Abbreviations

ADRN: Administrative Data Research Network; CPRD: Clinical Practice Research Datalink; DVLA: Driver and Vehicle Licensing Agency; DWP: Department for Work and Pensions; HMRC: HM Revenue and Customs; HSCIC: The Health and Social Care Information Centre; ISD: Information Services Division; NHS: National Health Service; NWEH: North West eHealth; NWIS: NHS Wales Information Centre; ONS: The Office for National Statistics; RCT: Randomised controlled trial; SAIL: The Secure Anonymised Information Linkage Databank; SANAD II: The Standard and New Antiepileptic Drugs II trial; THIN: The Health Improvement Network

### Acknowledgements

Not Applicable

### Funding

This report is independent research arising from a Clinical Training Fellowship (GA Powell) awarded by the Medical Research Council Hubs for Trials Methodology Research (Reference: P16-2014-GPC). LJ Bonnett is funded by a Post-Doctoral Fellowship (PDF-2015-08-044) from the National Institute for Health Research and AG Marson is part funded by the National Institute for Health Research Collaboration for Leadership in Applied Health Research and Care North West Coast. The views expressed are those of the authors and not necessarily those of the NHS, the Medical Research Council, the National Institute for Health Research or the routine data sources detailed in the report.

### Availability of data and materials

There is no dataset available for this study and, therefore, data sharing is not applicable.

**Authors' contributions**

GAP performed the background research, prepared the protocol, liaised with the routine data sources, prepared and submitted the routine data applications and drafted and redrafted the manuscript. LJB, CTS, DAH, PRW and AGM provided input into the development of the protocol and drafted and redrafted the manuscript. AGM is the guarantor for the report. The University of Liverpool was the sponsor for the study. All authors read and approved the final manuscript.

**Ethics approval and consent to participate**

To request identifiable data from routine data sources, written participant consent was required. The North of Scotland Research Ethics Service (16/NS/0007) and Health Research Authority (IRAS 189002) approved the study.

**Consent for publication**

Not applicable

**Competing interests**

All authors declare that (1) GAP, LJB, CTS, DAH, PRW and AGM do not have support from any company for the submitted work, (2) GP, LJB, CTS, DAH, PRW and AGM have no financial relationship with any company that might have an interest in the submitted work in the previous 3 years and (3) GP, LJB, CTS, DAH, PRW and AGM have no other relationships or activities that could appear to have influenced the submitted work.

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Author details**

<sup>1</sup>Department of Molecular and Clinical Pharmacology, Clinical Sciences Centre, Lower Lane, Fazakerley, Liverpool L97LJ, UK. <sup>2</sup>Department of Biostatistics, University of Liverpool, Waterhouse Building, Block F, 1-5 Brownlow Street, Liverpool L69 3GL, UK. <sup>3</sup>Centre for Health Economics and Medicines Evaluation, Institute of Medical and Social Care Research, College of Health and Behavioural Sciences, Bangor University, Ardudwy, Normal Site, Gwynedd, North Wales LL57 2PZ, UK.

Received: 13 January 2017 Accepted: 3 August 2017

Published online: 23 August 2017

**References**

- UK Government. The Data Protection Act. 2016. Available from: <https://www.gov.uk/data-protection/the-data-protection-act>. Accessed Mar 2016.
- UK Government. The Freedom of Information Act. 2016. Available from: <http://www.legislation.gov.uk/ukpga/2000/36/contents>. Accessed 4 Mar 2016.
- Lewsey JD, et al. Using routine data to complement and enhance the results of randomised controlled trials. *Health Technol Assess*. 2000;4(22):1–45. i-iv.
- Raftery J, Roderick P, Stevens A. Potential use of routine databases in health technology assessment. *Health Technol Assess (Winch Eng)*. 2005;9(20):1–92. iii-iv.
- Clarson LE, et al. Increased risk of vascular disease associated with gout: a retrospective, matched cohort study in the UK Clinical Practice Research Datalink. *Ann Rheum Dis*. 2015;74(4):642–7.
- McGregor J, et al. The Health Informatics Trial Enhancement Project (HITE): using routinely collected primary care data to identify potential participants for a depression trial. *Trials*. 2010;11:39.
- Williams JG, et al. Can randomised trials rely on existing electronic data? A feasibility study to explore the value of routine data in health technology assessment. *Health Technol Assess (Winch Eng)*. 2003;7(26):1–117. iii, v-x.
- Gulliford MC, et al. Cluster randomized trials utilizing primary care electronic health records: methodological issues in design, conduct, and analysis (eCRT Study). *Trials*. 2014;15(1):220.
- UK Government. The Plan for Growth. 2011. Available from: <http://www.gov.uk/government/publications/plan-for-growth>. Accessed 3 Mar 2016.
- UK Government. The NHS Constitution for England. 2013. Available from: <http://www.gov.uk/government/publications/the-nhs-constitution-for-england>. Accessed 3 Mar 2016.
- ADRN. The Administrative Data Research Network. 2016. Available from: <http://adrn.ac.uk>. Accessed 16 Jun 2016.
- SAIL. The Secure Anonymised Information Linkage Databank. 2016. Available from: <http://www.saildatabank.com>. Accessed 1 Jun 2016.
- Sayers A, et al. Evidence for a persistent, major excess in all cause admissions to hospital in children with type-1 diabetes: results from a large Welsh national matched community cohort study. *BMJ Open*. 2015;5(4):e005644.
- New JP, et al. Obtaining real-world evidence: The Salford Lung Study. *Thorax*. 2014;69(12):1152–4.
- (US) National Research Council. The prevention and treatment of missing data in clinical trials. National Academies Press, Washington DC (US); 2010.
- HSCIC. General Practice Extraction Service. 2016. Available from: <http://www.hscic.gov.uk/gpes>. Accessed 5 Apr 2016.
- Clinical-Study-Data-Request. Clinical Study Data Request. 2016. Available from: <https://www.clinicalstudydatarequest.com>. Accessed 5 Apr 2016.
- Nelson EC, Dixon-Woods M, Batalden PB, Homa K, Van Citters AD, Morgan TS, Eftimovska E, Fisher ES, Ovretveit J, Harrison W, Lind C, Lindblad S. Patient focused registries can improve health, care and science. *BMJ*. 2016;354:i3319.
- HSCIC. Press release: patient opt out. 2016. Available from: <http://www.hscic.gov.uk/catalogue/PUB20527>. Accessed 1 May 2016.
- Van Staa TP, Goldacre B, Buchan I, Smeeth L. Big health data: the need to earn public trust after past management. *BMJ*. 2016;354:95–7.
- HSCIC. The Health and Social Care Information Centre. 2016. Available from: <http://www.hscic.gov.uk>. Accessed 16 Jul 2016.
- Bouras G, et al. Linked hospital and primary care database analysis of the incidence and impact of psychiatric morbidity following gastrointestinal cancer surgery in England. *Ann Surg*. 2016;264(1):93–9.
- Turner EL, et al. Design and preliminary recruitment results of the Cluster randomised trial of PSA testing for Prostate cancer (CAP). *Br J Cancer*. 2014;110(12):2829–36.
- NWIS. NHS Wales Informatics Service. 2016. Available from: <http://www.wales.nhs.uk/nwis/page/52490>. Accessed 4 May 2016.
- Ismail SI, Puyk B. The rise of obstetric anal sphincter injuries (OASIS): 11-year trend analysis using Patient Episode Database for Wales (PEDW) data. *J Obstet Gynaecol*. 2014;34(6):495–8.
- ISD Scotland. The Information Services Division. 2016. Available from: <http://www.isdscotland.org/Products-and-Services/index.asp>. Accessed 5 May 2016.
- Ahmed A, et al. Upper gastrointestinal bleeding in Scotland 2000–2010: improved outcomes but a significant weekend effect. *World J Gastroenterol*. 2015;21(38):10890–7.
- CPRD. The Clinical Practice Research Datalink. 2016. Available from: <http://www.cprd.com/intro.asp>. Accessed 4 May 2016.
- TTP. ResearchOne. 2016. Available from: <http://www.tpp-uk.com/products/systemone>. Accessed 4 May 2016.
- Herrett E, et al. Text messaging reminders for influenza vaccine in primary care: protocol for a cluster randomised controlled trial (TXT4FLUJAB). *BMJ Open*. 2014;4(5):e004633.
- QResearch. QResearch. 2016. Available from: <http://www.qresearch.org/SitePages/Home.aspx>. Accessed 6 May 2016.
- Hill T, et al. Antidepressant use and risk of epilepsy and seizures in people aged 20 to 64 years: cohort study using a primary care database. *BMC Psychiatry*. 2015;15:315.
- THIN. The Health Improvement Network. 2016. Available from: <http://www.thin-uk.net>. Accessed 5 Apr 2016.
- González-Pérez A, et al. Incidence and predictors of hemorrhagic stroke in users of low-dose acetylsalicylic acid. *J Stroke Cerebrovasc Dis*. 2015;24(10):2321–8.
- NWEH. North West eHealth. 2016. Available from: <http://www.nweh.org.uk>. Accessed 5 May 2016.
- ApolloMedical. Apollo Data Extraction. 2016. Available from: <http://www.apollo-medical.com/>. Accessed 6 Apr 2016.
- GraphnetHealth. Graphnet. 2016. Available from: <http://www.graphnethealth.com/what-we-do/overview/what-we-do>. Accessed 6 Apr 2016.
- ONS. The Office for National Statistics. 2016. Available from: <https://www.ons.gov.uk>. Accessed 4 Apr 2016.
- ONS. Official Labour Market Statistics. 2016. Available from: <http://www.nomisweb.co.uk>. Accessed 4 Apr 2016.
- OCSI. Data 4 Neighbourhoods and Regeneration. 2016. Available from: <http://www.data4nr.net/introduction>. Accessed 7 Apr 2016.

41. UK Government. HM Revenue and Customs. 2016. Available from: <https://www.gov.uk/government/organisations/hm-revenue-customs>. Accessed 4 Apr 2016.
42. UK Government. The Department for Work and Pensions. 2016. Available from: <https://www.gov.uk/government/organisations/department-for-work-pensions>. Accessed 1 May 2016.
43. UK Government. Driver and Vehicle Licensing Agency. 2016. Available from: <https://www.gov.uk/government/organisations/driver-and-vehicle-licensing-agency>. Accessed 2 May 2016.
44. Clarke M. Standardising outcomes for clinical trials and systematic reviews. *Trials*. 2007;8:39.