Assessments of Harms in Clinical Trials

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

Alex John Hodkinson

August 2015

Abstract

Introduction and Aims

Healthcare interventions are usually associated with a risk of harmful events that must be balanced against the potential favorable outcomes. However reliable evidence on harms for interventions is often inadequate, and hampered by the many challenges that stem from the reporting, analysis and interpretation of harms data in clinical trials. This thesis addresses some of these issues.

Methods

Reporting of harms data is assessed in a systematic review of reviews and a case study investigating the additional value of harms data reported in clinical study reports (CSRs). A framework for searching and identifying relevant sources of harms data is outlined, and then explored further in a survey assessing current practices in clinical trial units (CTUs). Signal detection methods are introduced, and evaluated using simulated data to assess their performance when detecting safety signals in CTU databases.

Results

The systematic review highlights that the reporting of harms in RCTs is inconsistent, and often inadequate. In the case study, CSRs presented data on harms, including SAEs which are not reported or mentioned in publications, they also provide more detail about the design, conduct and analysis of the trial which facilitate the assessment of risk of bias in evidence synthesis. A wide

range of sources for harms data have been identified, each with distinct strengths and limitations discussed. Selection of appropriate sources depends on the research question, and whether a hypothesis generating or hypothesis testing approach should be taken. Relevant sources have been identified for each approach, with examples of their exploitation in CTUs evaluated in the survey. The simulation study has shown that some of the current available signal detection methods are not able to control the false discovery rate well, and are only able to detect few safety signals for small or sparse data.

Conclusions

The work carried out within this thesis provides some recommendations to address the reporting, conduct, and analysis of harms in clinical trials. Wider adoption of recommendations made by the CONSORT-harms guideline will enhance the quality of reporting and improve subsequent evidence synthesis. Recent initiatives to promote open access to clinical trials data including CSRs is a major step towards supporting better data transparency. It is important to identify and consider different sources that are most likely to yield robust data on harms of interest, rather than relying on studies that cannot reliably detect harm. The survey identified published literature and systematic reviews as the most common source being used in the trial safety monitoring within CTUs. Signal detection methods are potentially unsuitable for use in CTUs. Further tools and guidelines for enhanced signal detection are needed in clinical trials.

Acknowledgements

I would like to start by sincerely thanking my supervisors, Dr Catrin Tudur Smith and Professor Carrol Gamble, for their continued support and guidance.

Several people have helped immensely with the work contained within this thesis. I thank Dr Jamie Kirkham for working as my co-reviewer, in chapter 2. I also thank Su Golder, Fiona Beyer and Alison Beamond for providing support with the search criteria used in chapter 2. For help and support during the research described in chapter 3, I thank Roche Genetech for preparing the trial documents to enable the study to be carried out and for responding to any email queries. For the work in chapter 7, I thank Dr Ismail Ahmed for support using a statistical package during this research.

For my funding, I thank the Medical Research Council (MRC) (grant number G1000397 - 1/1) and the North West Hub for Trials Methodology Research (grant number G0800792).

Finally, I must thank my parents, and brother for their constant love and support.

Table of Contents

Abbreviations used in this thesis

AE	Adverse event
SAE	Serious adverse event
SUSAR	Suspected unexpected serious adverse reaction
ADR	Adverse drug reaction
RCT	Randomised controlled trial
CTU	Clinical trial unit
PV	Pharmacovigilance
EMA	European Medicines agency
WHO	World health organization
MHRA	Medicines healthcare regulatory agency
FDA	Food and Drug Administration
SmPCs	Summary product characteristics
IB	Investigator's brochure
CSR	Clinical study report
CRF	Case report form
SOP	Standard operating procedure
SRS	Spontaneous reporting system
SDA	Signal detection algorithm
DPA	Disproportionality analysis
DSUR	Development safety update report
PSUR	Periodic safety update report
IDSMC	Individual data safety monitoring committee
DMC	Data monitoring committee
PRR	Proportional reporting ratio
IC	Information component
BCPNN	Bayesian confidence propagation neural network
GPS	Gamma Poisson shrinker
FDR	False discovery rate
UMC	Uppsala Monitoring Centre
NHS	National Health Service

Definitions

There exist various definitions used in literature to describe harms in clinical trials. In this thesis I will adhere to the conventional and widely accepted definitions proposed by the Uppsala Monitoring Centre (UMC) and the World Health Organization (WHO), though various other related terms commonly used through this thesis are defined as well.

A ***"drug"*** or ***"medicine"*** is a pharmaceutical product, used in or on the human body for the prevention diagnosis or treatment of disease, or for the modification of physiological function.
A **"health care intervention"** or **"intervention"** is any type of treatment, preventive care, or test that a person could take or undergo to improve health or to help with a particular problem. Health care interventions include drugs (either prescription drugs or drugs that can be bought without a prescription), foods, supplements (such as vitamins), vaccinations, screening tests (to rule out a certain disease), exercises (to improve fitness), hospital treatment, and certain kinds of care (such as physical therapy).
An ***"adverse (drug) reaction"*** is a response to a medicine which is noxious and unintended, and which occurs at doses normally used in humans for the prophylaxis, diagnosis or therapy of disease, or for the modification of physiological function. (Normal dose clause distinguishes adverse reactions from poisoning and this clause was later refined by Meyboom, 2000 [1], to caution on patients experiencing an adverse reaction at normal dose but may indeed be a case of high/toxic dose because of impaired renal/hepatic excretion or other reasons). It is common for the term ***"adverse effect"*** to be used as synonyms for adverse reaction. Adverse effect is seen from the point of view of the drug whereas an adverse reaction from the point of view of the patient. Another commonly used definition for an ADR was put forward by Edwards and Aronson [2], who define an ADR as - an appreciably harmful or unpleasant reaction,

resulting from an intervention related to the use of a medicinal product, which predicts hazard from future administration and warrants prevention or specific treatment, or alteration of the dosage regimen or withdrawal of the product. The Medicines and Healthcare products Regulatory Agency (MHRA) has a broader definition of an ADR - as an unwanted or harmful reaction experienced following the administration of a drug or combination of drugs, which is suspected to be related to the drug. Unlike the WHO definition, the MHRA definition does not exclude overdose or drug misuse.
"Harm(s)" is often the totality of possible adverse consequences of an intervention or therapy; they are the direct opposite of benefits.
"Safety" refers to the substantive evidence of an absence of harm. The term is often misused when there is simply absence of evidence of harm.
A ***"side effect"*** is any unintended effect of a pharmaceutical product occurring at doses normally used in man, which is related to the pharmacological properties of the drug.
An **"*adverse event*"** or **"*experience*"** is defined as any untoward medical occurrence that may present during treatment with a medicine but which does not necessarily have a causal relationship with the treatment.
A ***"signal"*** or ***"safety signal"*** is reported information on a possible causal relationship between an adverse event and a drug, of which the relationship is unknown or incompletely documented previously [2].
"Serious (not synonymous with 'severe' which is used to describe the intensity of a specific outcome) ***AEs/reactions"*** can be defined as those that: - are life threatening or fatal - cause or prolong hospital admission - cause persistent incapacity or disability - concern misuse or dependence. A ***"suspected unexpected serious adverse reaction"*** (SUSAR) is an adverse reaction that is both unexpected and also meets the definition of a SAE/R.
"Complication" is a term widely used to describe adverse events following surgical and other invasive interventions. 'Adverse event' and 'adverse effect'

can be considered synonyms.

Publications arising from the work

Work from chapter 2 has been published and work from 3 is currently under review for publication. Full references for the relevant articles are shown and copies included in the appendix.

Chapter 2

Reporting of harms data in RCTs: a systematic review of empirical assessments against the CONSORT Harms extension

[1]Alex Hodkinson, [1]Jamie J Kirkham, [1]Catrin Tudur-Smith, [1]Carrol Gamble

[1]MRC North West Hub for Trials Methodology Research, Department of Biostatistics, University of Liverpool, UK.

BMJ Open *2013;3:e003436 doi:10.1136/bmjopen-2013-003436*

Chapter 3

Reporting of harms outcomes: A comparison of journal publications with unpublished clinical study reports of orlistat trials

Alex Hodkinson[1], Carrol Gamble[1], Catrin Tudur Smith[1]

[1]MRC North West Hub for Trials Methodology Research, Department of Biostatistics, University of Liverpool, UK.

Under review for publication

Chapter 1: Introduction

In clinical trials harmful effects are generally associated with drug interventions, and so for the majority of this thesis we will focus on drug trials. We start by introducing the drug lifecycle.

1.1 The Drug Lifecycle

Medications are the most frequently employed therapeutic intervention for disease and have led to substantial improvements in morbidity, mortality, and quality of life of patients around the world [3]. However, medications for all their virtues, can also cause harm, and there is growing recognition that our knowledge of a drug's potential for harms is incomplete at the time of licensing. This is well-illustrated by the staggering numbers recently reported by Strom in 2006 [4], that suggest that 51% of drug undergo labeling changes due to major safety issues discovered after marketing.

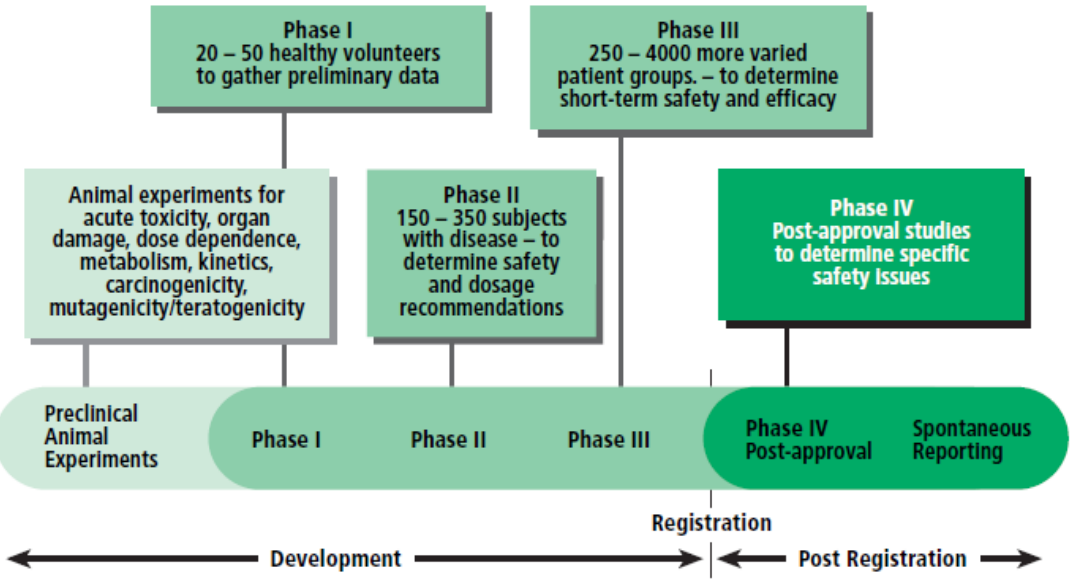
To understand why drugs that initially pass the federal bar for safety and efficacy, and receive the green-light for widespread use, are later discovered to cause harm – we must look at the drug lifecycle [5].

1.1.1 Phases of a Clinical Trial

For safety reasons, before a drug is tested on any humans, preclinical studies are carried out on animals in order to learn more about any toxic effects the drug may have. Once researchers are satisfied with the safety/toxicity of a drug in animals, human clinical trials can start.

Clinical trials are usually split into the four phases, with Phases I to III the development of the drug and Phase IV the post-approval stage, as shown in Figure 1 and explained below:

Figure 1: Phases of clinical development
(Adapted from the World Health Organization (WHO) [6]).



PHASE I A small number, 20-50 (usually young and healthy) volunteers are given the drug to see whether they can tolerate it.

PHASE II The drug is then tested on approximately 150-350 patients with the disease to determine its safety and identify the likely dose(s) that are effective ('phase IIa'). A larger ('phase IIb') trial often follows to identify the efficacy of the drug and to determine how well the drug works at the prescribed dose(s).

PHASE

Trials assess the safety and efficacy of the drug in approximately 250-4000 patients. They may include a comparison group of patients who take a similar drug that is already available. This phase is sometimes called the *"pre-marketing phase"* because it actually measures consumer response to the drug.

PHASE

Trials are carried out after the drug is in general use to find out more about the side-effects and safety of the drug, what the long term harms and benefits are and how well the drug works when it is used more widely. Phase IV is also known as *"post-marketing surveillance"* and includes the safety surveillance of the drug after licensing.

To prevent or reduce harm to patients and thus improve public health, mechanisms for evaluating and monitoring the safety of drugs in clinical use are vital. In practice this means having in place a well-organized pharmacovigilance (PV) programme that takes place continuously throughout the life cycle of a new drug.

1.1.2 Pharmacovigilance and Risk Management

The World Health Organization (WHO) defines PV as the science and activities relating to the detection, evaluation, understanding, and prevention of adverse reactions to medicines or any other drug-related problems [6]. The major aims of PV are:

- Early detection of thus far unknown adverse reactions and interactions

- Detection of increases in frequency of (known) adverse reactions
- Identification of risk factors and possible mechanisms underlying adverse reactions
- Estimation of quantitative aspects of benefit/risk analysis and dissemination of information needed to improve drug prescribing and regulation.

In clinical trials serious adverse events (SAEs) or suspected unexpected serious adverse reactions (SUSARs) are of particular interest in the pre-licensing PV assessment because these are often drug induced. However, systematic safety monitoring in PV systems is also needed to identify previously recognized and unrecognized harms, and to evaluate the safety of medicinal products during clinical trials and in the post-marketing period.

Risk management is the discipline within PV that is responsible for signal detection and the monitoring of the risk-benefit profile of drugs. Risk management has now added focus on safety and risk assessment after a drug has received regulatory approval, when it is placed on the market and prescribed to large populations. Other key activities within the area of Risk management are that of the compilation of Risk Management Plans (RMPs) and aggregate reports such as the periodic Safety update reports (PSURs) and the development safety update report (DSUR), which we discuss later in this chapter.

1.1.3 Pharmacovigilance in the Regulation of Drugs

After Phase 3 clinical trials, regulators have to decide whether to license the drug. Before licensing, drug companies must submit a RMP to the regulator at the time of application for marketing authorization. The RMP includes information on: the drug's safety profile, how risks will be prevented or minimized in patients, plans for further studies to gain more knowledge about the safety of the drug and the risk factors for developing side effects.

Drugs with side effects can be licensed but the beneficial effects must outweigh the risks of harms [7]. The decision takes into account the following:

- The type of illness being treated
- The improvement offered by the drug
- The intensity of side effects
- The likelihood of serious side effects
- The possibility of predicting who is most likely to experience serious side effects.

When treating life-threatening illnesses, more severe side effects are acceptable if the drug could cure or significantly prolong life. For example, chemotherapy can kill cancer cells and lead to recovery, so the risk of severe side effects is accepted. A drug may also still be licensed if a very small number of people respond badly during a trial. To advise prescribers about the possible side-effects of the drug, reported events and their incidence are described in the drug label or the patient information leaflet (PIL) [2]. PILs are a patient friendly-version of the "summaries of product characteristics (SmPCs)". The SmPCs

provide more detailed information to healthcare professionals on how often the side effect may happen, how severe it might be, how long it might last for and what action should be taken. The SmPC is updated throughout the life-cycle of the drug as new data emerge, and they can be accessed in the electronic Medicines Compendium (eMC).

Regulators can review a license if new information comes to light after the drug is in general use, and make further recommendations to improve the benefit-risk ratio of the drug. To support these decisions and recommendations about the drug's safety, existing and often new evidence from "clinical research" is needed.

1.1.4 Sources of Evidence on Harms

Prior to starting any clinical research, an investigator must determine the appropriate study design to answer the question at hand. Selecting the correct study type also depends on ethical considerations, disease of interest, and the resources available. A well-designed study will clearly identify an exposure and an outcome in an objective, quantifiable manner to answer a defined hypothesis. Understanding the various indications for different study designs is important not only for devising one's own study but also for critically reviewing the literature. Therefore it is important firstly to outline some of the frequently encountered study designs used in clinical research and discuss their respective strengths and limitations to making assessments about harms. We begin by discussing randomised controlled trials (RCTs), then observational studies and

the impact of systematic reviews, but will also extend to the use of data in post-marketing surveillance.

1.2 Randomised Controlled Trials (RCTs)

In an RCT, study subjects are randomly assigned to one of two groups; treatment arm, which receives the intervention, or the control arm, which receives a placebo or no treatment. Both study arms are subsequently followed in an identical manner and analyzed for differences in outcomes. The intrinsic design of an RCT allows investigators to assess causality of a treatment, rather than simply a correlation. RCTs generally have stringent selection criteria to ensure that subjects are comparable in most respects, thereby reducing confounding and isolating the effect of the intervention.

1.2.1 Issues to consider when designing RCTs with harm endpoints

Properly designed and executed RCTs are considered the "gold standard" for evaluating efficacy because they minimize potential bias. However, relying solely on published RCTs to evaluate harms can be problematic.

Most RCTs lack pre-specified hypotheses for harms; they are usually designed to evaluate beneficial effects as their primary objective, with assessment of harms being the secondary objective [8]. As a result, the quality and quantity of harms reporting in clinical trials is frequently inadequate [9, 10]. Furthermore, RCTs often lack large enough sample sizes [11] or are sometimes limited in duration to adequately assess uncommon or long-term (delayed) harms [7]. They are also explanatory, rather than pragmatic in design, i.e., they assess benefits and

harms in ideal, homogenous populations and settings [12]. Patients who are more susceptible to AEs are often underrepresented in such “efficacy” trials. Publication and selective outcome(s) bias in RCTs can lead to distorted conclusions about harms when data are unpublished, partially reported, downplayed, or omitted [13, 14].

Despite these limitations RCTs are the gold standard for demonstrating efficacy, the basis for most regulatory approvals, and claims made on behalf of drugs and other interventions. For this reason, harms data in RCTs must be addressed in detail when they are available.

1.2.2 Including unpublished harms from RCTs

In addition to evaluating results of published RCTs, results of completed or terminated but unpublished RCTs, as well as unpublished results should be included. There are a number of potential advantages for accessing unreported outcomes which can help in evaluating the risks for publication or outcome reporting bias [15], and to evaluate discrepancies in conclusions based on unpublished harms data against those based on published harms in RCTs [16].

However unpublished data from trials can be difficult to locate systematically. Recent efforts have been made by researchers for further disclosure of clinical trial results, by obtaining data and certain documentation from regulatory agencies and drug companies. These researchers were able to unveil more comprehensive data and information about a clinical trial, mainly through accessing clinical study reports (CSRs). The CSR has now made it possible to obtain further existing harms information that may not have been detailed in

the trial publication, or was unpublished in the first place. The value of the CSR was demonstrated in a recent study [17] assessing the benefits and harms of reboxetine against placebo or selective serotonin reuptake inhibitors in acute treatment of depression. The unpublished data from the manufacturer and in the CSRs suggests that Reboxetine is ineffective and potentially harmful and that the published evidence in journals is affected by publication bias.

Registry reports have also been used in the past to obtain further information and results of clinical trials. Since the release of the 2007 Food and Drug Administration (FDA) reform bill trial sponsors are now responsible for reporting results to the clinical trial results database (ClinicalTrials.gov) [18] which can be accessed by the public. Other similar schemes and databases have been set up by the World Health Organization (WHO) and certain drug companies.

1.2.3 Reporting of harms in clinical trials

When reporting harms in clinical trials it is important to discuss not only expedited and the different forms of mandatory reporting to health authorities, but also reporting guidance for published literature.

1.2.3.1 Expedited and mandatory reporting

To ensure that all new and clinically important harms are not overlooked or reported too late, health authorities in the United States, the European Union (EU), Japan, and elsewhere require that safety information be reported on both an expedited (within 7 or 15 calendar days) and period (quarterly, biannual, annual, etc.) basis. Serious adverse reactions, unexpected reactions and those reactions with a relationship to treatment must be reported to authorities

expeditiously. Health authorities also require mandatory periodic submission of safety information during clinical development and when a drug is marketed [19]. These responsibilities of safety reporting are clearly laid out within a range of key regulations and documents, including the EU clinical trials directive [20], the medicines for human use (clinical trials) regulations 2004 [21] and the International Conference on Harmonization - Good Clinical Practice (ICH-GCP) E6 [22].

There are also the different types of mandated reports that are required by health authorities that also contain further safety information:

- "Individual case safety reports (ICSRs)" which are required for SAEs and SUSARs in RCTs, and they usually provide a narrative summary of the event.
- "Investigator's brochure (IB)" contains a summary of available nonclinical and clinical study information for efficacy and safety findings from complete clinical trials, and is routinely updated.
- "Clinical study report (CSR)" which provide detailed summaries of potentially unpublished information on harms of a clinical trial.
- "Periodic safety update reports (PSURs)" which primarily summarizes safety findings of a marketed drug. The purpose is to determine if any new regulatory concerns have emerged, and if the benefit-risk profile of the drug has changed.

1.2.3.2 Reporting harms in published literature

In addition to expedited reporting and the different forms of mandatory safety reporting, it is also important for harms to be reported in the published literature. However prior analyses of published RCTs suggest suboptimal reporting of harms-related data [9, 10, 23, 24]. Prompted by such evidence, the consolidated standards of reporting in trials (CONSORT) members convened in May 2003 to generate an extension of the CONSORT recommendations regarding the appropriate reporting of harms. The panel generated a 10 recommendation checklist [8], with accompanying explanation and examples of appropriate reporting in RCTs. The reporting standards since the release of the extension have not been assessed, but will be later in this thesis.

1.2.4 Collecting harms data from patients

When collecting and recording harms from patients in clinical trials, a range of approaches can be taken; from asking the patients standard questions, keeping diary cards, developing questionnaires or checklists and recording events in case report forms (CRFs) [25].

The use of standard questions should be the usual method in all clinical trials, and should be unambiguous and asked in the same way for each patient as defined in the study protocol. Diary cards generally collect harms experienced by the patients in an unrestricted fashion on a daily basis, the questionnaire or checklist collects harms in a structured fashion, so that valid statistical comparisons can be made between treatment arms. The collection of harms through questionnaires is usually performed either with a quality of life

questionnaire, or a questionnaire designed specifically for trials with specific drugs.

The process of dealing with large amounts of data collected at investigator sites during clinical trials, including AEs reports, has been largely paper based. Data are usually recorded on paper (or electronically) in CRFs, and then reviewed and verified at investigator sites, by a study monitor or a clinical research associate from the sponsor company or a clinical research organization. CRFs record information on the AE: description, category, start/end date, outcome, severity grade, seriousness, expectedness and action taken as demonstrated in the sample CRF in Figure 2. SAEs and SUSARs are usually sent in advance of the complete CRFs and entered into a separate harms database.

Figure 2: Sample case report form for adverse events, taken from the clinical trials unit from the University of Liverpool.

HEADER: PI NAME, Protocol or IRB Number, Protocol Short Title

Subject Initials ☐☐☐ **Subject ID#** ☐☐☐☐☐ **Page** ☐☐ **of** ☐☐

Adverse Event Tracking Log

#	Date Reported	Adverse Event Description	Adverse Event Category**	Start Date	End Date	Outcome[1]	Severity/ Grade[2]	Serious (Y or N)	Expected (Y or N)	AE Treatment[3]	Action Taken[4]	Attribution[5]	PI Initials	Date of PI Initials

- AE number. "1" indicates the first adverse event documented on the form, 2 = the second, etc. If the adverse event changes in severity, enter it as a separate adverse event row on the paper form using the same AE number as the one that ended.

**look up corresponding AE Category at: http://safetyprofiler-ctep.nci.nih.gov/CTC/CTC.aspx

Outcome[1]
0 – Fatal
1 – Not recovered/not resolved
2 – Recovered w/sequelae
3 – Recovered w/o sequelae
4 – Recovering/Resolving

Severity/Grade[2]
1 – Mild
2 – Moderate
3 – Severe
4 – Life-threatening
5 – Fatal

AE Treatment[3]
0 – None
1 – Medication(s)
2 – Non-medication TX

Action Taken[4] with Study Treatment
0 – None
1 – Interrupted
2 – Discontinued
3 – Dose reduced
4 – Dose increased
5 – Not Applicable

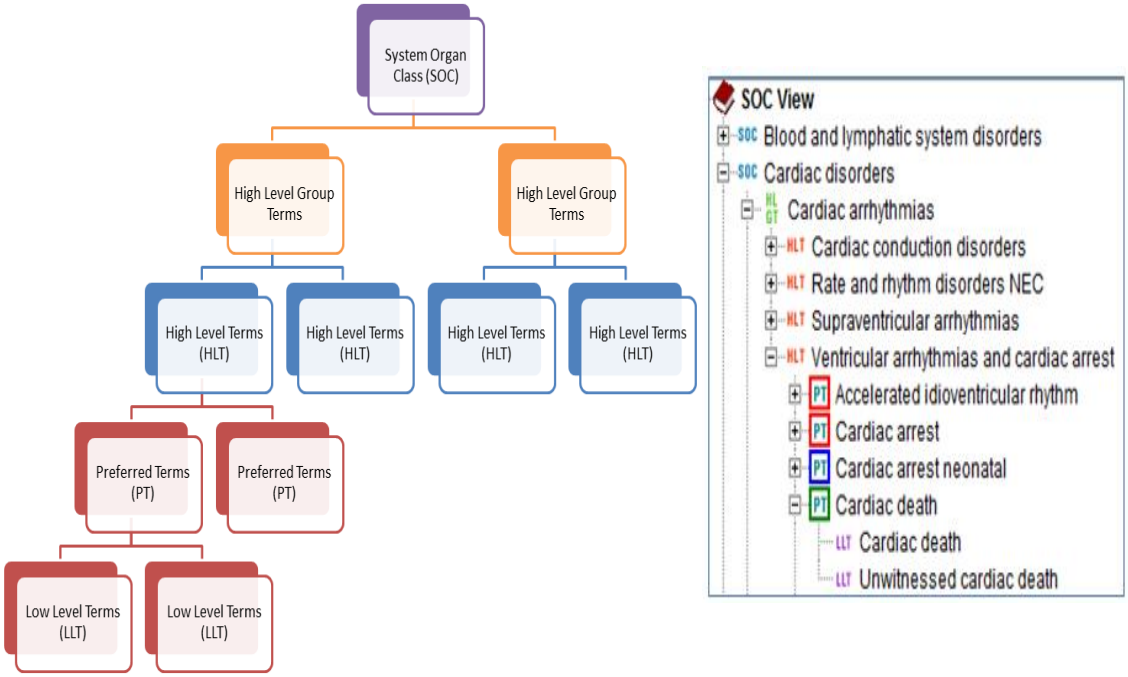
Attribution/ Relatedness[5]
0 – Definite
1 – Probable
2 – Possible
3 – Unlikely
4 – Unrelated

1.2.5 Coding harms data

Data reported and collected will later be transformed by a medical coder employed by the trial sponsor. Coding is a process whereby harms data are categorized in a standard way so the data can be pooled or combined for analysis. Coders often use a medical dictionary, which is a predefined list of possible AEs organized in a hierarchy, to code the narrative description of an AE. In the past companies have historically used many different dictionaries to code and categorize harms, such as the World Health Organization's (WHO's) Adverse Reaction Terminology (WHO-ART) [26], Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART) [27], or the International Classification of Diseases (ICD 9 and ICD 10) [28]. In 1994, the pharmaceutical industry, together with regulatory agencies, developed a standard dictionary named the Medical Dictionary for Regulatory Activities (MedDRA) [29].

MedDRA is split into a five level hierarchy (Figure 3), with lowest level terms (LLTs) at the bottom, followed by preferred terms (PTs), and with system organ class (SOC) at the top. Events are initially coded with LLTs which consist of thousands of synonyms and alternative spelling of PTs. In the earlier phases of a drugs lifecycle, MedDRA can be used, for example, for recording AEs and baseline medical histories in clinical trials, in the analysis and tabulations of data from these, and in expedited submission of SAEs to regulatory agencies. It can also be used in constructing standard product information; such as SmPCs or product labeling. After licensing, MedDRA is used in PV for continuing evaluation of drug safety.

Figure 3: The Hierarchy of MedDRA coding system with an example.



MedDRA continues to grow and develop more coded items. The benefit of more coded items means that the coding of AEs could intuitively lead to less inter-observer variation, because there will be more exact matches to the reported AE. Conversely, it might also lead to increased variation because it becomes difficult to code nonspecific terms. When there is great uncertainty on how AEs are coded and a lack of proper training, it can often lead to misclassification. It was shown in a recent review [30] assessing the inter-observer variation and other challenges of coding AEs, that the increase in MedDRA categories has potentially made detecting AEs harder, and therefore compromised the safety assessment of interventions. Comprehensive inter-observer studies are needed to overcome the issue of coding AEs.

1.3 Observational Comparative studies

Observational comparative studies draw inferences about the possible effect of a treatment on subjects, where the assignment of subjects into a treatment group against the control group is outside the control of the investigator.

To assess harms adequately observational studies are almost always necessary. The exception is when there are sufficient data from RCTs to reliably estimate harms. Even though observational studies are more susceptible to bias than well-conducted RCTs, for some comparisons there may be few or no long-term, large, head-to-head, or effectiveness RCTs [31]. Observational studies may also provide the best (or only) evidence for evaluating harms in minority or vulnerable populations (such as pregnant women, children or elderly patients) who are underrepresented in clinical trials.

1.3.1 Cohort and Case-control studies

The term observational studies is commonly used to refer to cohort, case control, and cross sectional studies [32], but can refer to a broad range of study designs, including spontaneous case reports, uncontrolled series of patients receiving interventions, and others. All can yield useful information as long as their specific limitations are understood. The choice of study designs also depends on whether investigators are seeking to determine what harms might be associated with a treatment (hypothesis generating) or whether certain harms are more likely (hypothesis testing). Well-designed and reported case-control and population-based cohort studies are well suited for testing hypotheses on whether one intervention is associated with greater risk for an AE

than is another, and for quantifying the risk [32, 33]. They also make stronger precautions against bias than other observational designs.

1.4 Impact of systematic reviews

Clear and complete reporting of harms data from RCTs and observational studies is also important for inclusion in systematic reviews. By including data from both types of studies this can influence the quality or amount of evidence regarding harms.

1.4.1 Combining data from RCTs and Observational studies

In a recent study [34] to assess the level of agreement and disagreement in estimates of harm derived from meta-analysis of RCTs as compared to meta-analysis of observational studies for 19 studies, the empirical evidence indicated that there was no difference on average in the risk estimate of adverse effects between RCTs and observational studies. The study recommends that systematic reviews of harms should not be restricted to specific study types, and instead it may be preferable for systematic reviewers of adverse effects to evaluate a broad range of studies that can help build a complete picture of any potential harm and improve the generalizability of the review.

However, in another report comparing evidence on harms in RCTs and non-randomized studies, the findings show that large observational studies usually report smaller absolute risk of harm than RCTs [35]. There was no clear tendency for RCTs or observational studies to report larger relative risks. In more than half of the comparisons, estimates of relative or absolute risk varied more than

twofold. Discrepancies between RCTs and observational studies may occur because of differences in populations, settings, or interventions; differences in study design, including criteria used to identify harms; differential effects of biases, or some combination of these factors, as summarized in Table 1 [36].

Table 1: The Key strengths and limitations to consider when synthesizing harms data from RCTs and observational studies.

Study design	Key strengths	Limitations
Randomized controlled trials (RCTs)	Randomization reduces possibility of confounding and bias Certain harms can be prospectively specified for detailed monitoring Intervention is typically well defined	Limited power to detect significant differences between groups for adverse effects, and often lacks precision Recruitment criteria may lead to exclusion of patients who are at risk of harms (i.e., children, elderly and pregnant women) Potential for biases from industry when trailing new drugs, and therefore side-effects can be ignored
Meta-analysis of controlled observational studies	Pooled analysis has greater power to detect significant differences, even with rare events Summarizes complete data set and can evaluate consistency of findings among studies Study population allows research into events reported in elderly and pregnant women	Reliant on quality of primary data Missing or unreported data on AEs is a major problem, as are the statistical techniques of pooling sparse data Potentially very small amount of data available on new interventions Susceptible to selective outcome reporting of primary studies Heterogeneity within pooled analysis

1.4.2 Limitations of systematic reviews

There are several other issues especially relevant to discuss when systematically synthesizing evidence on harms. This includes, combining studies only when they are similar enough to warrant combining particularly when evaluating rare and uncommon events, exploring potential sources of heterogeneity in meta-analysis, and adequately considering outcome reporting bias and tools for assessing risk of bias.

1.4.2.1 Rare events

Evaluating comparative risks of uncommon or rare events in systematic reviews can be particularly challenging. A frequent problem in RCTs and systematic reviews is interpreting a non-significant probability value as indicating non-significant difference in risk for a rare AE, particularly when the confidence intervals (CIs) are wide and encompass the possibility of clinically important risks.

For example, in one trial [37] investigating patients with meningitis, “treatment with dexamethasone did not result in an increased risk of AEs” compared with placebo for treatment of hyperglycemia, herpes zoster, or fungal infection because the p-values were greater than 0.20. However, the 95% CIs for relative risk estimates of these three AEs showed clinically significant increase risks. In such as case, researchers should acknowledge the lack of statistical power to assess risks adequately and should interpret the CIs.

1.4.2.2 Meta-Analysis

The exact choice of statistical methods to evaluate harms data in a systematic review will depend upon the individual context. Meta-analysis is the preferred method to synthesize evidence in a comprehensive, transparent, and reproducible manner. Though, the rarity of some serious harm outcomes, the relatively small size of some trials, and the restricted patient populations may limit the detection and full evolution of the harms of drugs in individual trials.

The assessment of statistical heterogeneity is appropriate but of lesser concern when dealing with rare but serious AEs where the primary focus is detecting the harm. Commonly employed tests for statistical heterogeneity include; Cochran's test which is considered relatively underpowered; the Peto odds ratio (OR) method with 95% CI which may provide the best CI coverage, and is more powerful and relatively less bias than random effects analysis when dealing with low event rates; and the fixed effect Mantel-Haenszel test and odds ratio which can be used to reduce confounding, and can adequately deal with zero events within the analysis [36].

1.4.2.3 Outcome Reporting Bias

Furthermore, the credibility of findings from individual trials and from summaries of trials examining a similar research question (that is, systematic reviews and meta-analyses) has been undermined by numerous reporting biases in the published medical literature. Reporting biases are often difficult to detect, but have the potential to discredit earnest efforts towards evidence-based decision making [13, 14, 38].

One of the major biases often involved when performing systematic reviews is outcome reporting bias (ORB), which refers to the selective reporting of some results but not others in trial publications. ORB acts in addition to, and in the same direction as "publication bias" of entire studies to produce inflated estimates of treatment effect. The suppression of non-significant findings could lead to the use of harmful interventions.

In a recent study [14] to determine the extent and nature of selective non-reporting of harm outcomes in a cohort study, including 92 systematic reviews of RCTs and non-RCTs, found significantly high evidence of ORB as a result of partially missing reported harms. The study proposes a classification system considering selective outcome reporting that should be appraised outside of the Cochrane risk of bias tool, which is currently being updated. The recommendations from this study are for improvements of reporting harms in both primary studies and systematic reviews.

To overcome ORB, reviewers should also attempt to identify further data from multiple sources including CSRs and clinical trial results registries like the clinicaltrials.gov, as key harms information may be missing from the published trial report.

1.4.2.4 Assessing risk of bias

The development of instruments for assessing risk of bias specifically in studies of harms is still in an early stage of development. General tools for assessing methodological quality can be used but with caution, because they may apply only to the primary focus of the study – usually the beneficial effects of the

intervention. For example, for current risk of bias tools like the McMaster Quality Assessment Scale of Harms (also known as McHarm), are designed to detect inflated treatment differences (type I error, i.e., finding of a harm that is not truly present) [39]. The McHarm tool was developed from quality rating of 15 items generated by a Delphi census review of the literature on harms and from previous quality assessment instruments. The subsequent list of the 15 quality criteria was tested for reliability and face, construct, and criterion validity. The McHarm tool is intended for use in conjunction with standardized quality-assessment tools for design-specific internal validity issues.

However due to poor monitoring, lack of clear case definitions and missing data mean that genuine adverse reactions may go undetected or be misclassified. It is therefore believed that systematic reviews of harm should explicitly assess the risk of bias toward the null (e.g., with more attention on harms with lower estimates of risk, like with rare or unexpected events) to prevent a false sense of security (type II error), whereby a drug is erroneously declared safe or not significantly different from the placebo or comparator [40].

The Cochrane handbook for systematic reviews of interventions [41] also highlights some areas of special concern: methods for monitoring and detecting harms, conflicting interests, selective outcome reporting (section 1.4.2.3) and blinding. Furthermore, the Cochrane risk of bias tool for non-randomised studies of interventions (ACROBAT-NRSI) was recently developed allowing for assessments of harms or benefits of an intervention. 1.4.3 Guidance to conducting systematic reviews of harms

Studies in the past have also identified other major challenges when developing systematic reviews of harms. This includes a poor quality of information on harms reported in original studies [9, 10, 23, 42], difficulties in identifying relevant studies on harms when using standard systematic search techniques [43, 44], and the lack of a specific guideline to perform a systematic review of harms.

To overcome some of these challenges a number of efforts have been made by collaborative groups and researchers by developing a logical framework and reporting guidelines to guide systematic reviewers.

1.4.3.1 The Cochrane Adverse Effects Methods Group

In 1993 the Cochrane collaboration [45] was formed to organize medical research information in a systematic way to facilitate the choices that health professionals, patients, policy makers and others face in health interventions according to the principles of evidence-based medicine. The group conducts systematic reviews of RCTs which it publishes in the Cochrane library. A few reviews have also studied the results of non-randomised, observational studies.

The Cochrane Adverse Effects Methods Group (AEMG) was formally registered with the Cochrane Collaboration on the 14th June 2007 [46]. The AEMG aims to develop the methods for producing high quality systematic reviews and to advise the Cochrane Collaboration on how the validity and precision of systematic reviews can be improved. A recent publication from the group has provided technical advice for a structured approach to conducting systematic reviews of harms [47], where reviewers are also given general guidance on the

assessment of study bias, data collection, analysis, presentation and interpretation of harms in a systematic review. This work will be discussed in more detail in later chapters. The group has also developed and proposed search strategies with appropriate search filters to help identify information on harms [43]. These search strategies aim to help balance the sensitivity (the ability to identify as many relevant articles as possible) with precision (the ability to exclude as many irrelevant articles as possible) when searching bibliographic databases. The AEMG also contribute chapter 14 (Adverse effects) to the Cochrane handbook for systematic reviews of interventions [41].

1.4.3.2 PRISMA Harms Guideline

Additional to the work carried out by the Cochrane AEMGs, in 2009 the Preferred Reporting Items for systematic Reviews and Meta-Analysis (PRISMA) statement [48] was developed as a revision of the Quality of Reporting of Meta-Analysis (QUOROM) statement [49]. The PRISMA statement was developed to guide researchers when conducting systematic reviews and performing meta-analysis in systematic reviews. The statement thus far has mainly focused on efficacy and not on harms. However, in a recent study [50] the quality of reporting in systematic reviews of harms were assessed using their own set of proposed items. The aims of this study were to provide valuable research in the first step of the development for the PRISMA harms extension.

1.5 Post-marketing surveillance

Monitoring the safety of a drug after it has been released on the market is also important. Data on harms after marketing mainly include spontaneous reports

and electronic health databases designed specifically for PV. These data can be used in pharmacoepidemiological studies to further research the risks of an adverse effect. A comprehensive PV program also includes evaluation of other relevant clinical findings (e.g., laboratory tests results, vital signs, cardiac or other specialized testing) that we do not address.

1.5.1 Spontaneous reporting

Spontaneous reports refer to unsolicited reports of clinical observations originating outside of a formal clinical study that are submitted to drug manufacturers or regulatory agencies. Some of the events will represent true adverse effects of treatment; many will be symptoms of disease being treated, or coincidental events that are unrelated to the diseases or treatment [51]. The most important reports are either new (i.e., not included in the drug label or SmPCs), rare, serious events associated with the drug's use, or recognized AEs occurring at a higher than anticipated rate.

Spontaneous reporting systems which collect reports centrally, can "signal" emerging problems and thereby have the potential for uncovering previously unknown adverse reactions. Since these reports are submitted by health care professionals, a great deal of time is spent analyzing individual reports and any patterns underlying these reports [52]. The limitations of spontaneous reports include substantial and unquantifiable underreporting (thus, such systems do not produce accurate estimates of incidence for a given AE) as well as lack of verification of important medical details.

Adverse events may be spontaneously reported at disproportionately high rates at various times in the drug's marketing life cycle. As a result of this, sophisticated statistical approaches to formalize the "signal generation" aspect of spontaneous reports, aimed at determining when a particular type of AE is reported disproportionately relative to other AEs associated with a given drug, have been developed. Such systems, often using Bayesian statistical methods, are used and evaluated by safety reviewers employed by regulatory authorities and drug companies. These methods may be useful as automated searching tools, especially as the number of spontaneous reports increases. However a clinical evaluation is usually required to determine the true causal effect.

1.5.2 Electronic health databases

Electronic health databases contain patient medical records and prospectively recorded information on medical events such as prescriptions, previous history, diagnosis, and test results. One widely used medical practice database for pharmacoepidemiological research is the Clinical Practice Research Data-Link (CPRD) in the UK. This database is a unique resource because it includes very detailed medical information, symptoms, and signs in a well-defined, representative, and stable population, and it is also validated (i.e., information on diagnosis and on prescriptions has been found to agree with that recorded on paper charts or provided by physicians).

However, there also exist some obvious limitations with the use of electronic health databases. The most widely used terminology for coding AEs has proved ill-suited to identifying the adverse effects of drugs, with differing coding

dictionaries used than the standard MedDRA dictionary which is predominantly used in most spontaneous reporting systems. In addition, evaluation of some accepted statistical methods (i.e., longitudinal Bayesian signal detection algorithms and other disproportionality analysis methods) have revealed systematic bias, finding statistically significant associations between drugs and events where no relationship was thought to exist. Other statistical methods also appeared unreliable [53].

In term of studies including electronic health records, self-controlled methods performed better than case controls and new user cohorts, even though the later two methods are widely used in other observational studies. The databases are also limited with respect to exposures to recently marketed drugs, and may be therefore better suited to studying older, well-established drugs or drug classes. Another issue is the duration of patient follow-up, which tends to be only a few years. Meaningful secondary care data is often not provided [54].

1.6 Thesis outline

This chapter has summarized some of the current issues and challenges involved when assessing harms in clinical trials. The thesis will cover reporting related issues by evaluating the progress of reporting guidelines in a systematic review, and the potential for exploiting further unpublished harms information contained within CSRs, which was explored in a case study. Additionally a survey was conducted to explore the current practice in clinical trial units across the UK, to understand how harms data is managed, used and analyzed. Finally, we investigate the use of signal detection methods for analyzing harms data from

clinical trials. Different scenarios of data were simulated to explore the potential for improved detection of safety signals, and to provide guidance in their use. Further descriptions of each of the chapters are discussed below.

The work in Chapter 2 has been published in the British Medical Journal (BMJ) Open [55], and I am first author. A systematic review of reviews was performed to evaluate the reporting of harms in RCTs when using the CONSORT-harms extension as a benchmark. The harms extension which was developed in 2004 includes 10 recommendations to complement in the preparation of RCT reports. Since the release of the extension there has been no indication of the current standards for reporting harms in RCTs, therefore this review will be the first to access this since its installment.

The work in Chapter 3 is currently under review for publication, and I am first author. The review in chapter 2 was restricted to assessing the reporting of harms in only publications of RCTs, therefore in this chapter we explore the value of using CSRs to exploit further information on harms. A case study of orlistat trials was conducted to assess whether, published results of harm outcomes in journal publications is consistent, with the underlying trial data within the unpublished data contained within CSRs. This was shown in an extensive meta-analysis of all harms data. This research highlights the value of CSRs and the potential for improved data transparency of clinical trial results.

Following on from chapter 3, the potential value of external sources of data beyond a RCT is explored in chapter 4 to maximize the information available when designing and analyzing trials. This involves a critical review of the

different PV systems in post-marketing safety surveillance including passive systems (spontaneous reporting systems), prescription-event monitoring, and electronic health databases which are used predominantly for hypothesis testing/strengthening in pharmacoepidemiological studies. These different data sources are to be investigated further in chapter 5, to discuss their potential value when used in clinical trial units (CTUs).

Chapter 5 investigates the current practices in CTUs by carrying out a survey across UK clinical research collaboration (CRC) registered CTUs. The aim of this survey is to understand how CTUs could improve upon the use of their existing harms data, to explore the value for using harms from external sources as discussed in chapter 4, and to understand the potential for using statistical signal detection methodologies to analyze harms data that may be available within CTUs, and within the wider CTU network. The results from this survey will be used to inform the simulation study in chapter 7.

After examining current practices in CTUs to determine the methodologies used to analyze harms data, chapter 6 will explore some of the more commonly used signal detection algorithms (SDAs) for analyzing spontaneous reported data and clinical trial data. Three SDAs based on disproportionality analysis are introduced in detail, performance related issues are evaluated and the potential for refinements also discussed. These SDAs will be explored further in chapter 7.

Chapter 7 starts with a literature review of past studies to assess their aims for investigating the use of SDAs, and determine what refinements were made if any. Then the performance of the three SDAs introduced in chapter 6, are

compared using simulated data. This includes evaluations of their performance for controlling for false discoveries, and the ability to maintain suitable levels of sensitivity by exploring the use of different thresholds. Furthermore, we determine their characteristics when detecting rare signals, and explore their potential for use in harms databases similar to CTUs.

Chapter 8 uses the work of chapters 2 to 7 to try to overcome some of the challenges that stem from the reporting, conduct, analysis and interpretation of harms in clinical trials. Recommendations for reporting in RCTs are split into discussions of reporting guidelines (chapter 2) and the potential improvements for better transparency by exploiting harms data in CSRs (chapter 3). The current practices when managing and analyzing harms in CTUs has been evaluated in the survey (chapter 5). Finally the simulation study (chapter 7) provides recommendations and guidance to using SDAs to analyze harms data in a number of different scenarios. Chapter 8 concludes with a section discussing potential further research.

Chapter 2: Reporting of Harms in RCTs – Systematic Review

In this chapter the quality of reporting harms data is explored in detail by conducting a systematic review of previous reviews. This work has been published in the British Medical Journal (BMJ) open [55] and the Cochrane AEMG has added the paper to their list of relevant publications for reporting in RCTs [46]. The paper has recently been cited in a number of other relevant published articles discussing outcome reporting bias issues associated with harms, and the endorsement of reporting guidelines for completeness of reporting [14, 56].

2.1 Introduction to the CONSORT Statement

Considering the importance of RCTs in the present world of evidence based practice, it is essential that the quality of trial findings in medical journals should be standardised in terms of the reporting rationale, methods, results and context of those results. To address these issues, two groups the standard of reporting trials (SORT) and Asilomar working group on recommendations for reporting of clinical trials in biomedical literature merged their proposal into one single, coherent evidence-based recommendation called the 'Consolidated Standards of Reporting Trials' (CONSORT) statement which was first published in 1996 [57]. The CONSORT statement provides the evidence based minimum set of recommendations for reporting RCTs, which is intended to facilitate the complete and transparent reporting of RCTs and aid their critical appraisal and

interpretation. Since 1996 the CONSORT statement has been updated twice, including the recent update in 2010 [58].

Since the publication of the CONSORT statement several healthcare journals have endorsed its use, leading to improvements in quality of reporting of RCTs. Recent systematic reviews [59, 60] comparing CONSORT-adopting and non-adopting journals resulted in a significant improvement in adherence to all items within CONSORT adopting journals. Due to the success of the standard CONSORT statement and other recognized additional complexities of particular trial designs and issues, additional extensions to the CONSORT statement, have been developed. For example for RCTs with specific designs (e.g., cluster randomized trials, non-inferiority and equivalence trials, pragmatic trials), data (e.g., harms, abstracts), and interventions (e.g., herbals, non-pharmacologic treatments, acupuncture).

As well as reporting guidelines for trial authors, networks also exist to help promote the good reporting of health research studies of RCTs. The EQUATOR (Enhancing the Quality and Transparency Of health Research) network has been established as a global hub to improve medical research and reliability of literature by promoting accurate reporting. The EQUATOR network provides training and guidance to peer reviewers and researchers when using the CONSORT guideline and extensions.

2.2 Methods

2.2.1 The CONSORT-Harms Extension

The standard CONSORT statement [61] is primarily aimed at reporting the intended, usually beneficial effects of intervention(s) with only one item (item 19) devoted to unintended AEs in the original 2001 checklist. This limitation, along with the accumulating evidence that reporting in RCTs was of poor quality with an imbalanced ratio of benefit-harms reporting [9, 10, 44], resulted in a CONSORT statement extension developed in 2004 to improve harms reporting (CONSORT-harms). The CONSORT-harms extension aims to help address perceived shortcomings in measurement, analysis, and reporting of harms data [8]. The extension consists of a ten criteria checklist to address the quality of harms reporting in all sections of an RCT journal article (title, abstract, introduction, methods, results and discussion) (Table 2). The subsequent update of the standard CONSORT statement, published in 2010 [62], now specifically refers to the additional CONSORT-harms extension but it is still unclear whether authors and journals routinely adopt the use of this extension [23, 60, 63].

2.2.2 Systematic Review

The aim of this study is to systematically review the evidence from previously conducted empirical studies that have assessed the adequacy of harms reporting in RCTs using the CONSORT-harms extension as a benchmark.

In this systematic review published and unpublished research were included, namely studies that evaluated the quality of harms reporting in RCTs against the CONSORT-harms recommendations [8]. No restriction was placed on the clinical

area or type of intervention studied. Excluded studies were those that assessed harms reporting using assessment criteria other than CONSORT-harms, and studies that assessed harms reporting using study designs for which the CONSORT guideline was not intended (e.g. observational studies).

Table 2: The 10 CONSORT-harms recommendations [8].

Recommendation	Description
1	If the study collected data on harms and benefits, the title and abstract should so state.
2	If the trial addresses both harms and benefits, the introduction should so state.
3	List addressed adverse events with definitions for each (with attention, when relevant, to grading, expected vs. unexpected events, reference to standardized and validated definitions, and description of new definitions).
4	Clarify how harms-related information was collected (mode of data collection, timing, attribution methods, intensity of ascertainment, and harms-related monitoring and stopping rules, if pertinent).
5	Describe plans for presenting and analyzing information on harms (including coding, handling of recurrent events, specification of timing issues, handling of continuous measures and any statistical analyses).
6	Describe for each arm the participant withdrawals that are due to harms and the experience with the allocated treatment.
7	Provide the denominators for analyses on harms.
8	Present the absolute risk of each adverse event (specifying type, grade, and seriousness per arm), and present appropriate metrics for recurrent events, continuous variables and scale variables, whenever pertinent.
9	Describe any subgroup analyses and exploratory analyses for harms.
10	Provide a balanced discussion of benefits and harms with emphasis on study limitations, generalizability, and other sources of information on harms.

The search strategy was developed with support from an information specialist with experience in systematic review search methodologies, and particularly

identifying studies focusing on harms reporting. The search strategy which is provided in Appendix A was then implemented in the following databases:

- Cochrane methodology register
- Database of abstracts of reviews of effects (DARE)
- Ovid MEDLINE
- Scopus
- ISI Web of Knowledge.

Conference abstracts were searched for in the web of knowledge Conference Proceedings Citation Indexes (CPCI-S or CPCI-SSH) and the Zetoc database [64]. An unpublished Masters dissertation involving one of the co-investigators was also obtained. Date filters were not used during the search criteria, although our interest lies only within studies published after 2004 (i.e. after the release of the harms extension), with the cut-off date June 2012.

The titles and abstracts of reports were identified by the search of the databases then screened with the full articles obtained for all potentially eligible studies. The screening which was done by one author was conducted through the referencing software EndNote (Version X5). Each full article was assessed independently by two investigators to determine eligibility. A copy of the full article was then obtained for all non-excluded reports, and each full article was assessed by two independent investigators to determine if it met the inclusion criteria. Any additional material about the study included as supplementary material on the journal website was also obtained.

2.2.3 Quality assessment and risk of bias

Two investigators independently assessed the methodological quality of each study using the Cochrane Risk of Bias (RoB) tool [65] as a guideline. The Cochrane collaboration's recommended tool for assessing risk of bias in RCTs is neither a scale nor a checklist. Instead, it is a domain-based evaluation, in which critical assessments are made separately for different domains. It was developed between 2005 and 2007 by a working group of methodologists, editors and review authors. It is a two part tool, addressing seven specific areas; sequence generation, allocation concealment, blinding of participants and personnel, blinding of outcome assessment, incomplete outcome data, selective outcome reporting and 'other issues'.

The purposes for the risk of bias in this study, was to assess reviews and not individual RCTs, so we adapted a risk of bias tool from the standard Cochrane tool and formulated our own criteria which are explained below.

Each study was graded as low risk, high risk or unclear as indicated below:

1. Were the trials included in the study a representative sample, e.g. unselected journals, and reasonable time scale?

   Low risk of bias: Studies included trials from a primary search of all the available literature.

   High risk of bias: Studies were highly selective of the trials included, e.g. high impact journals or specialized journals only.

   Unclear risk of bias: Not stated how studies were selected.

2. During the data extraction of CONSORT-harms criteria, were reviewers blinded to study authors, institution, journal name and sponsors?

   Low risk of bias: Reviewers were blinded.

   High risk of bias: Reviewers were not blinded.

   Unclear risk of bias: Not stated.

3. Is there evidence of selective outcome reporting in the study (i.e. were all CONSORT-harms recommendations considered and if not were suitable reasons provided)?

   Low risk of bias: Studies that considered all CONSORT-harms criteria or reasons for excluding specific criteria were transparent and justified.

   High risk of bias: Studies did not consider all CONSORT-harms criteria.

   Unclear risk of bias: Unclear whether all CONSORT-harms criteria were considered.

4. Did more than one reviewer assess the CONSORT-harms criteria for each primary RCT, with a description of how agreement was achieved?

   Low risk of bias: Data extraction was completed independently by two people or reasonable attempts were made to maximize data extraction reliability.

   High risk of bias: Data extraction not completed independently by two people.

   Unclear risk of bias: Not stated.

Lead authors were contacted when any of the criteria were deemed unclear, or not reported in the journal article.

2.2.4 Data Extraction

The data extraction was completed by two independent investigators and any discrepancies were resolved through discussion with a third investigator. The data extraction included:

- Study characteristics: Inclusion criteria including clinical area, types of interventions, databases or journals searched within the study and any search date restrictions.
- Sample size (defined by the number of RCT reports assessed for reporting quality).
- Reporting quality: inclusion of any of the ten recommendations from the 2004 CONSORT-harms checklist (Table 2).

2.2.5 Analysis methods

For each study, the percentage of included RCTs that satisfied each CONSORT-harms recommendation is presented with 95% confidence intervals (CIs). Some studies had presented data for individual items described within each of the ten criteria rather than overall data. For example the recommendation was split into sub-items of assessment; these are presented as such in tables with a caption to provide further explanation. Forest plots were used to graphically depict the levels of adherence to the CONSORT harms recommendations, this was demonstrated as the proportion of studies within each review that satisfied each criteria. So that readers can easily discern the extent of compliance and

heterogeneity between studies with the I-squared statistic (Appendix A: Figures 20). We refrained from statistically combining results from the different studies due to the differences in their study characteristics. The R software (version 3.0.2) was used to perform any meta-analysis.

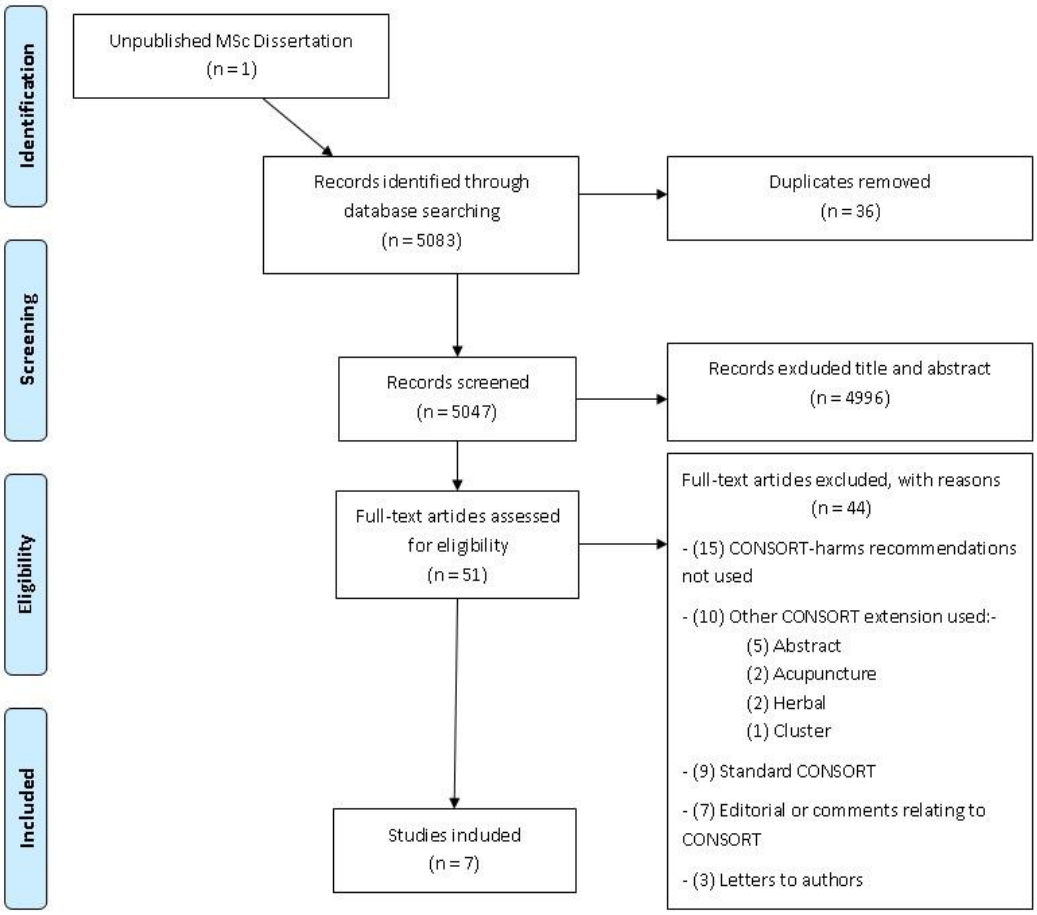
In accordance with the Cochrane Handbook, I^2 statistics were interpreted as (0% to 40%, might not be important; 30% to 60%, may represent moderate heterogeneity; 50% to 90% may represent substantial heterogeneity; 75% to 100%, considerable heterogeneity) [41].

2.3 Results

The search strategy identified 5083 potentially eligible study cohorts (including one unpublished dissertation), which were then screened at title and abstract level in Endnote (Figure 4).

There were 36 duplicates removed and 4996 citations excluded at this stage. Full papers were reviewed for the remaining 51 citations and seven articles, with one being a dissertation obtained by departmental communication that met the inclusion criteria and were included in the study. 15 were excluded since they did not use the CONSORT harms guideline, 10 used another CONSORT extension, 9 used the standard CONSORT, 7 were editorials or comments relating to CONSORT, and three were letters to authors. We identified seven studies assessing the quality of reporting across almost 800 RCTs which were included in this study.

Figure 4: Flow diagram of study identification and selection



Five studies (Bagul [66], Breau [67], Turner [68], Shukralla [69] and Capili [70]) contained trials focusing on specific clinical areas with two studies (Pitrou [71] and Haidich [72]) covering multiple clinical areas (Table 3). Four studies [66, 69, 71, 72] included trials using drug interventions, one study [70] comparing acupuncture and another alternative complementary medicines [68], the interventions were unclear in one study [67]. MEDLINE was used by four studies [69-72] to identify the relevant literature, three studies [66, 68, 69] used the Cochrane database of RCTs and three studies [67-69] searched specialised journal databases. The date restrictions used in the search strategy of each

study ranged from a one year period up to a nine year span. The studies were published after 2008, four years after the release of the harms extension with three studies [67, 69, 72] including trials that had been published before the publication of CONSORT-harms, with a pre and post-CONSORT harms assessment. Five studies [66-69, 71] excluded trials published in a non-English language.

Table 3: Characteristics of included studies.

Study Characteristic	Bagul (2012) [66]	Breau (2011) [67]	Capili (2009) [70]	Haidich (2009) [72]	Pitrou (2009) [71]	Shukralla (2011) [69]	Turner (2011) [68]
Clinical area	Hypertension	Urology	Acupuncture	Mixed	Mixed	Epilepsy	Acupuncture therapies and other Complementary Alternative Medicines (CAM)
Type of intervention(s)	Drug interventions for Hypertension	Unclear	Acupuncture (excluding studies that evaluated acupressure, laser acupuncture, and auricular acupuncture)	Drug interventions	Drug interventions	Drug interventions for epilepsy	Acupuncture, massage therapies and herbal medicines.
Journals/Databases searched	Cochrane Central Register	Journal of Urology; Urology; European Urology; BJU International	MEDLINE; Allied & Complementary Medicine; Cumulative Index to Nursing & Allied Health Literature; Evidence Based	MEDLINE (Annals of Internal Medicine, BMJ, JAMA, Lancet, NEJM)	MEDLINE via PubMed (NEJM, Lancet, JAMA; BMJ; Annals of Internal Medicine; PLoS	MEDLINE; Cochrane Library; Epilepsy Group Trial Registry	Cochrane Collaboration's CAM Field specialized register of trials

			Medicine Reviews (EBMR)		Medicine)		
Date restrictions	January 2005 to September 2010	1996 and 2004 only	2005 to 2008	2003 and 2006 only	January 1st 2006 to January 1st 2007	January 1999 to December 2008	2009
Number of RCTs included in the study	41	152	10	102	133	152	205
Inclusion criteria	Randomised controlled hypertension trials comparing two parallel arms reported in standard CONSORT endorsing journals	RCTs of therapeutic interventions published in the three selected journals for the two years.	Studies published in the English language, acupuncture for pain reduction, a method for evaluating level of pain, and randomized allocation to treatment group.	Published RCTs assessing drugs in the selected journals for the two years.	Articles were included if the study was identified as an RCT with 2 parallel arms (in selected journals).	RCTs comparing AEDs (Anti-Epileptic Drugs); RCT patient population with epilepsy; RCTs published in English.	All Cochrane Complementary Medicine RCTs pertaining to 15 CAM intervention categories

2.3.1 Risk of bias

Lead authors were contacted by email with any queries relating to the quality of their study, or CONSORT criteria; however two authors Breau [67] and Capili [70] failed to respond. The risk of bias for the seven included studies, assessed across four domains, is summarized in Table 4.

Table 4: Risk of bias assessment.

Risk of bias criteria	Bagul (2012) [66]	Breau (2011) [67]	Capili (2009) [70]	Haidich (2009) [72]	Pitrou (2009) [71]	Shukralla (2011) [69]	Turner (2011) [68]
Representativeness of sample of trials (Low if trials were searched across unselected journals and across a reasonable time period).	High	High	Low	High	High	Low	Low
Blinding of reviewers during CONSORT-harms data extraction (Low if reviewers blinded to study authors, institution, journal name and sponsors).	High	Low	Unclear	High	High	High	Low
Selective outcome reporting (Low if all CONSORT-harms criteria assessed).	Low[a]	Low[a]	High[c]	Low	Low[d]	Low[a]	Low[a,b]
Reliability of data extraction (Low if more than one reviewer assessed the CONSORT harms criteria for each review that was undertaken, with a description of how agreement was achieved).	High	Low	Low	Low	Low	Low	Low

[a] Recommendation nine was not included in these studies as subgroup analysis was either not reported in any of the included studies or considered to be irrelevant for the therapeutic area being investigated.
[b] Authors response: "Recommendation 8 has been captured elsewhere in data extraction, to report this item would be to duplicate information presented".
"Recommendation 10 was considered too vague to assess with any objectivity so we decided to leave this item, especially given that some of our primary outcomes were already reasonably subjective".
[c] Recommendations 1, 2, 7, 8, 9 and 10 were not assessed, and reasons were not detailed. We classified this as high risk because recommendations 7 (number of patients analyzed) and 8 (Results for each adverse event) were not assessed.
[d] Recommendations 2, 9 and 10 were not assessed, and reasons were not detailed. However, this study is classified as low risk because the missing items relate to introduction (recommendation 2) and discussion (recommendation 10).

Six studies [66, 67, 69-72] were classified as high risk of bias for at least one domain with one of these studies [66, 71] classified as high risk for three domains. Four studies [66, 67, 71, 72] did not include trials from a representative sample as the search had targeted specific journals rather than a full systematic database search. Blinding of assessors was only implemented in two studies [67, 68] with one study [70] unclear. Most studies used all the CONSORT harms criteria with the exception of the subgroup analysis item. One study [68] discarded the use of recommendation eight (Results for each AE), since it was captured elsewhere within the data extraction, and recommendation ten (balanced discussion), which was considered too vague to assess with any objectivity. Reporting of the assessment within three studies [67, 70, 71] was unclear and authors were contacted. The authors did not respond for two studies [67, 70] and in another study [71] a response was received but some details remained unclear. Six studies [67-72] had used two independent data extractors while one study [66] had not and was classified as high risk of bias for this domain.

2.3.2 CONSORT-Harms recommendations

The results extracted for the CONSORT-harms criteria (Table 5) demonstrate variability in the level of adherence to items. Heterogeneity is highlighted by the individual forest plots (Appendix A, Figures 20) where inflated I^2 - squared values of over 85% are represented for all recommendations, denoting considerable heterogeneity.

Table 5: CONSORT harms criteria reported across included studies.

	Bagul (2012) [66]	Breau (2011) [67]	Capili (2009) [70]	Haidich (2009) [72]	Pitrou (2009) [71]	Shukralla (2011) [69]	Turner (2011) [68]
Total number of trials included in the study	41	152	10	102	133	152	205
CONSORT Recommendation	% of trials (95% CI) that adhered to each recommendation						
(1) Title & Abstract	20 (9, 35)	12 (6, 20) 1i) 12 (6, 20) 1ii) 64 (53, 74)	NR	76 (67, 84)	71 (63, 79)	88 (81, 92)	21 (16, 27)
(2) Introduction	34 (20, 51)	54 (43, 65)	NR	48 (38, 58)	NR	74 (67, 81)	4 (2, 8)
(3) Definition of adverse events	0 (0, 9)	15 (8, 24)	10 (0, 45)	59 (49, 69)	16 (10, 23)	3a) 36 (29, 45) 3b) 32 (25, 40) 3c) 47 (39, 55) 3d) 16 (11, 23) 3e) 22 (15, 29)	6 (3, 11)
(4) Collection of harms data	10 (3, 23)	4i) 22 (14, 32) 4ii) 6 (2, 13) 4iii) 0 (0, 4)	20 (3, 56)	81 (74, 89)	89 (82, 94)	4a) 57 (49, 65) 4b) 76 (69, 83) 4c) 33 (26, 42)	17 (12, 22)
(5) Analysis of harms	0 (0, 9)	76 (66, 84)	20 (3, 56)	44 (34, 54)	12 (7, 19)	5a) 36 (28, 44) 5b) 7 (4, 13)	6 (3, 10)
(6) Withdrawals	51 (35, 67)	35 (25, 45)	70 (35, 93)	59 (50, 69)	53 (44, 61)	6a) 71 (63, 78) 6b) 72 (65, 79)	30 (24, 37)
(7) Number of patients analysed	17 (7, 32)	35 (25, 45)	NR	74 (64, 82)	84 (77, 90)	7a) 78 (72, 85) 7b) 40 (32, 48)	18 (13, 24)
(8) Results for each adverse event	39 (24, 56)	8i) 0 (0, 4) 8ii) 28 (19,38)	NR	89 (82, 95)	73 (65, 80)	8a) 35 (28, 44) 8b) 68 (60, 76) 8c) 47 (39, 56) 8d) 19 (14, 27)	-

(9) Subgroup Analysis	-	-	NR	53 (43, 63)	NR	-	-
(10) Balanced discussion	5 (1, 17)	10i) 61 (50, 71) 10ii) 14 (7, 23) 10iii) 44 (33, 55)	NR	83 (76, 91)	NR	10a) 68 (60, 76) 10b) 61 (54, 70) 10c) 41 (34, 50)	-

NR Not reported in manuscript, and no response from authors when contacted.

- **Author detailed reasons for not reporting the recommendation.**

1) (i) Harm, safety or similar term used in title; (ii) Harm addressed in abstract.

4) (i) When harm information was collected; (ii) Methods to attribute harm to intervention; (iii) Stopping rules.

8) (i) Effect sizes for harms; (ii) Stratified serious and minor harms.

10) (i) Interpret harm outcome; (ii) discuss generalizability; (iii) discuss current evidence.

3) (a) Definition of AE; (b) All or selected sample; (c) Treatment Emergent AE; (d) Validated instrument; (e) Validated dictionary.

4) (a) Mode of AE collection; (b) Timing of AE; (c) Details of attribution.

5) (a) Details of presentation and analysis; (b) Handling of recurrent AE.

6) (a) Early or late withdrawals; (b) Serious AEs or death.

7) (a) Provide denominators for AEs; (b) Provide definitions used for analysis set.

8) (a) Same analysis set used for efficacy and safety; (b) Results presented separately; (c) Severity and grading of AEs; (d) Provide both number of AEs and number of patients with AEs.

10) (a) Discusses prior AE data; (b) Discussion is balanced; (c) Discusses limitations.

Of the six studies that assess inclusion of harms in the title and abstract of their included RCTs, three [69, 71, 72] reported compliance in over 70% of RCTs, but three [66-68] reported compliance in less than 30% of RCTs. The introduction section of the included RCTs reflect an imbalance in the reporting benefit-harms, with one study [68] reporting that less than 5% of RCTs had mentioned harms in the introduction, and one study [69] reporting more than 70% of its included RCTs has satisfied this criteria.

The definition of adverse events in reports is unsatisfactory in most studies [66-69, 71] indicating that fewer than 20% of RCTs satisfy these criteria adequately. The collection of harms-related information is described by more than 80% of RCTs in two studies [71, 72], but this high level is not consistent across the other five studies with one study [66] suggesting that as few as 10% of RCTs had provided an adequate description. The analysis and coding of adverse events is poorly described, with less than 50% of RCTs satisfying this criteria across six studies [67-72] with one of these studies [66] indicating that none of the RCTs had provided an adequate description. The reporting of participant withdrawals due to harms was inconsistent within two studies [67, 68] suggesting infrequent reporting with less than 40% of RCTs mentioning withdrawals, and three studies [66, 71, 72] suggest occasional reporting with 50-60% of RCTs mentioning withdrawals, and two studies [69, 70] suggesting that reporting of withdrawals was quite common with approximately 70% of RCTs mentioning withdrawals.

When providing the denominators within trial reports, the results were also varied across studies, with three [69, 71, 72] all identifying more than 70% of trials that satisfied this criterion, but two studies [66, 68] identifying less than 20% adherence. The risk and severity grading of adverse events is detailed in more than 70% of trials across two studies [71, 72], but the reporting is inadequate in three studies [66, 67, 69]. An assessment of reporting of harms within subgroup analysis was only carried out within study [72].

Four studies [66, 67, 69, 72] assessed their included RCTs for a balanced report on the benefits and harms within their discussion: one study [66] identified a

very low percentage (<10%), two studies [67, 69] identified a moderate percentage (approximately 60%), and one study [72] identified a high percentage (over 80%) of trials that met this criterion.

2.4 Discussion

This is the first study to systematically review empirical studies assessing the quality of reporting according to the CONSORT harms guideline [8]. Data were extracted from seven studies that had each assessed the quality of reporting across almost 800 RCTs from a range of clinical specialties. Eight years have now passed since the release of the harms extension, allowing adequate time for the guideline implementation. But, this study highlights that the reporting of harms in RCTs is inconsistent, and at times very poor. Heterogeneity is easily discerned between studies for each recommendation with inflated I^2 - squared values of over 85%. Further adherence to the CONSORT harms is needed.

The standard CONSORT statement for reporting RCTs is well established in health research with increasing evidence to support the use of the guideline [23, 60]. Currently the standard CONSORT is endorsed by over 50% of the core medical journals in the abridged index Medicus on PubMed [73]. In a review [74] of 116 health research journals, 41 provided online instructions to authors. Almost half (19/41 (46%)) mentioned the standard CONSORT guideline but none referred to the CONSORT extension for harms.

Previous studies [9, 10] prior to the CONSORT-harms statement have highlighted the problems associated with the lack in quality when reporting harms across

various different interventions. For example a systematic survey [10] was conducted in 2001 to determine if reporting of ADRs in a wide selection of RCTs was in accordance with the Standards Of Reporting Trials (SORT) group recommendations. Trial reports within this survey failed to provide details of how ADRs were defined or recorded: *"48/160 (30%) did not give clear definitions to the adverse event experienced"*. This survey found further evidence of poor reporting with *"44/86 (51%) of trials didn't give details on how severity was defined, or if used which severity grading system was used"*.

In the same year, a survey [9] of safety reporting including 192 randomized drug trials for seven medical areas, found the quality and quantity of safety reporting to vary across medical areas, study design, and settings. Reporting was found to be largely inadequate: *"Only 39% of trials provided adequate reporting of clinical adverse effects and 11% of those adverse effects had partially adequate reporting"*. Furthermore reporting of discontinuations were found inadequate: *"The numbers of discontinuations due to toxicity per study arm were mentioned in 75% of trial reports, but specific reasons for these discontinuations were given only 46% of the time".*

The focus in this study was to assess the reporting according to the CONSORT harms criteria only. The included studies contained trials reported prior to the publication of the CONSORT harms guideline. However, any changes in reporting over time were not assessed in this study. Nevertheless, our results support those from previous studies [9, 10] that used various guidelines published before the release of the CONSORT harms extension. This study should be regarded as a

reflection of reporting standards in general rather than an assessment of adherence to the CONSORT harms extension.

This study was strengthened by its assessment of quality of the included studies across four key domains. With the guidance of the Cochrane review [65] a RoB tool was designed to perform a generaliseable assessment of the included studies. In this assessment only the one study [68] determined to be low risk of bias across all four of the assessment criteria. No restriction was placed on the inclusion criteria of the identified studies, meaning that the time span and clinical area were varied. Whilst this is a-strength in terms of generaliseability of results, it may also be considered as a level of heterogeneity that cannot be explored due to the limited number of studies.

Although the CONSORT harms extension provides researchers and journals with a strict guideline to follow when reporting harms, there is supporting evidence that the uptake of adopting such guidelines appears to be slow [23]. It also seems that more than just the publication of the CONSORT guideline is required to assist editors and investigators in proper conduct and reporting of harms related issues in RCTs. The standard CONSORT has seen improvements over time with great emphasis and persistence by CONSORT members and researchers. Evidence is accumulating with large systematic reviews highlighting these improvements. The CONSORT extension for harms and further developments will help in the detection of adverse reactions in health care.

Complete and accurate reporting is essential to guide decisions on advances in medical interventions. The responsibility to ensure greater balance between

reporting of both benefits and harms lies with authors of research and journals publishing that research. It is recognized that journals have limited space for the reporting of all outcomes which can lead to selective outcomes reporting [75, 76]. Nevertheless, researchers should make full use of on-line facilities to publish supplementary material to ensure that all important available information on the potential harmful effects of drugs is available in the public domain.

Further dissemination strategies should be used to ensure that trial journal editors and trial investigators are aware of the importance of adequate reporting of harms related data in RCTs. As it stands, it is unclear as to whether the problem of the poor reporting of harms data in trial publications is a result of the lack of awareness of the CONSORT for harms statement, or journals and peer reviewers not implementing this guideline. The most effective strategy would follow that of the CONSORT statement with the extension for harms comprehensively incorporated in journal requirements along with clear instructions to peer reviewers for guidelines of acceptance.

In this review it was clear from the studies included that different approaches have been taken when assessing adherence to the CONSORT-harms checklist. Therefore we recommended that systematic reviewers follow the guidance provided in this study to help support future studies that wish to use the CONSORT-harms to access the quality of harms reporting within RCTs. Our risk of bias assessment tool should be used to ensure that the study has been conducted to the highest quality by following the four criteria, but also this criteria could be extend to support reviewers with the search criteria when

locating the literature. We list some useful key-words which were used in our search strategy in Appendix A, but further guidance on the use of different bibliographic databases and search techniques (i.e., free-text and/or combined with medical subject headings (MeSH)) is needed. We also recommend that the assessment of harms reporting over time is discussed. Reviewers could perform regression modeling with the time of publication included in the model to look for any improvements in reporting over time.

Since the reporting of harms in published journals of RCTs was found to be poor and inadequate in this study, chapter 3 will investigate other avenues to fully exploit the use of existing harms data.

Chapter 3: Reporting of Harms in Clinical Study Reports – Case Study

In chapter 2 the reporting of harms in RCTs was assessed by systematic review, the results from this review suggest that journal publications of RCTs poorly reported harms according to the CONSORT-harms. In this chapter we provide a further extensive evaluation of reporting harms in RCTs, by comparing the results from a meta-analysis based on data extracted from journal publications against the corresponding meta-analysis based on data extracted from the unpublished clinical study report (CSR). The chapter begins by providing a detailed background of CSRs with supporting evidence of their use and impact in the research, and then the results from a case study are presented in section 3.2. This case study is currently under review for publication.

3.1 Introduction

There are two driving concerns that continue to grow when relying on published medical research to reflect the truth [77]. Firstly, trials often remain unpublished years after completion and the results are therefore invisible to the public. Secondly, trials often display a distorted representation, where publications present a bias or misleading description of the design, conduct, or results of a trial [15, 38].

In recent years major initiatives have been developed to prevent or at least try to overcome these growing concerns with the registration of clinical trials as a precondition for publication in the international committee of medical journal

editors (ICMJE) [78], and mandatory trial registration and reporting of methods and results in the WHO's international clinical trials registry platform (ICTRP) [79] from 2005 onwards. However, the application of these measures have been insufficient; since they do not apply to clinical trials completed before 2005.

3.1.1 Understanding the Evidence Iceberg

Various types of formats exist for reporting clinical trials of interventions. Journal publications and registry reports currently represent the main publically available information source for obtaining summaries of clinical trial data for the purposes of clinical and health policy decision making [80]. However, results in the past have found reporting in journal publications to be inadequate and inconsistent [23], and although clinical trial registries have been responsible for making major strides in improving the transparency of trial data, a recent study suggested that the results from trial registries often remain invisible [81]. Trial protocols can also provide detail on the intended methods of conducting, analyzing and reporting in the trial.

In contrast to these three formats, there also exists a realm of unpublished and often invisible source for accessing further information and data on clinical trials, including: Individual participant data (IPD), unpublished data, case report forms (CRFs) and Investigators brochure (IB) as detailed in Table 6. These sources in the past have been found valuable to inform on evidence base decisions on the efficacy and safety of clinical trials, however accessing them can often be difficult.

Table 6: Sources of unpublished and often invisible clinical trials data.

Source	Description
Individual participant data (IPD)	Data for each participant in a trial. This contrasts with aggregate or summary data, which is produced by combining data from multiple participants. Individual participant data allows for the replication of all analysis in the study reports and exploration of further analysis.
Unpublished data	Data of any type (measurements, analysis, narratives, or judgments) from a trial that have not been published, irrespective of whether the trial is published. Since trial reports in peer reviews scientific journals typically provide highly compressed summaries of trial data, large amounts of unpublished data will remain for these trials.
Case reports forms (CRFs)	The original paper or electronic forms on which individual participant's data (demographic, efficacy, safety, etc) are recorded during the clinical trial, and the data they contain are statistically analyzed only after they have been entered into an electronic database of individual patient data. Forms can vary in length, from a few pages to hundreds of pages, and each trial can have multiple forms - for example, for different visits or for the different tests or procedures the participant undergoes.
Investigators brochure (IB)	A document written by a sponsor and intended for clinical investigators interested in becoming involved in a study. It summarizes the current body of evidence about an intervention under investigation, typically based on preclinical and human studies. The document is periodically updated in light of new information.

3.1.2 Clinical Study Report

The Clinical Study Report (CSR) is another format for reporting clinical trials. The CSR is a structured document which summarizes the analysis methods and results of a clinical trial submitted for marketing authorization of an investigational medicinal product in the European Union, Japan, or the United States [82]. CSRs are an "integrated" full report which can be up to a thousand pages in length, and include extensive detailed information on the efficacy and harms of interventions. Information in these documents relating to harms, are

usually separated individually by AE and SAE terms in summary tables and listings.

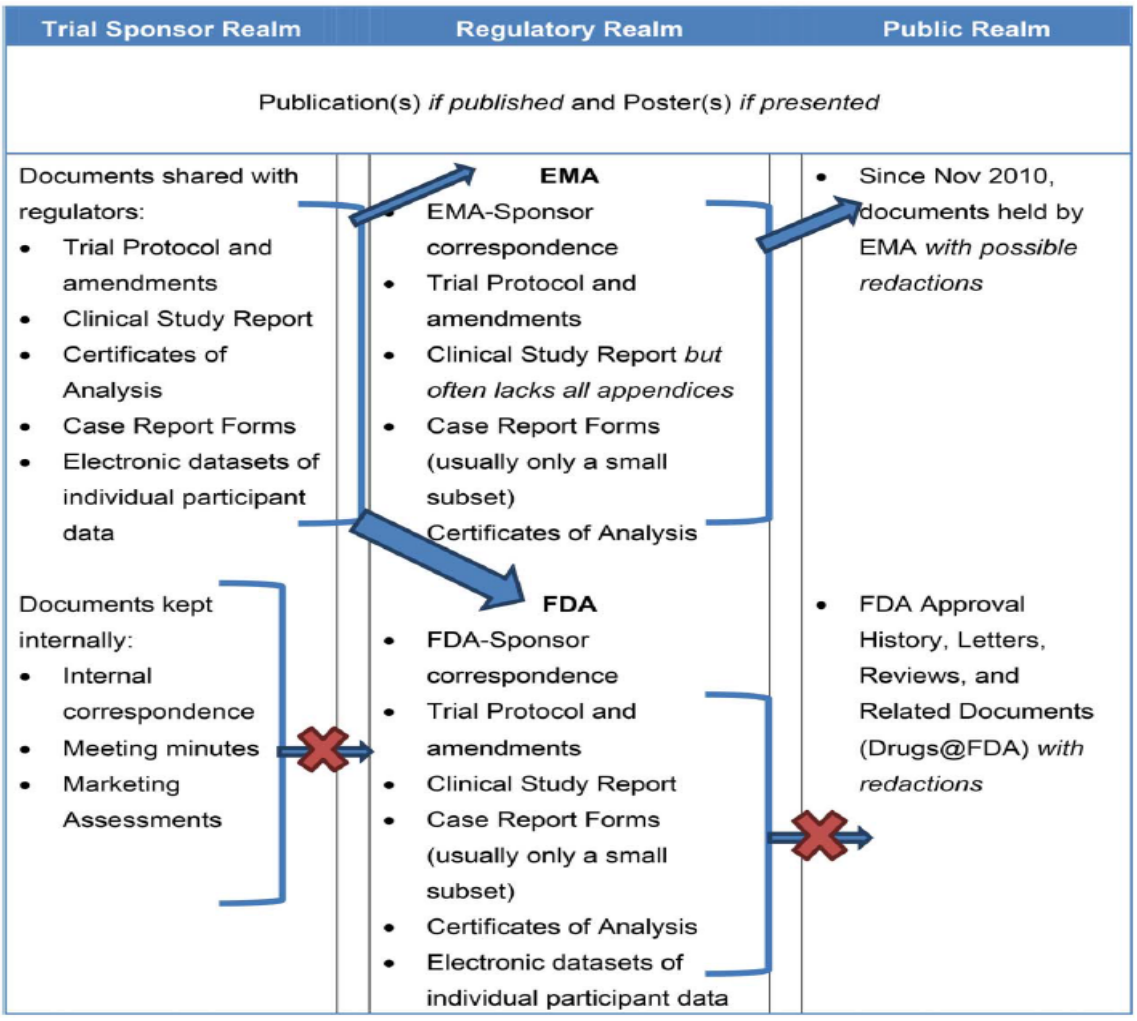
In the past researchers have made major efforts to gain access to CSRs, with the intention to inform regulatory decision-making [83]. The information contained within CSRs has proved vital when evaluating both the efficacy [84] and safety [85] of clinical interventions. Evidence from journal publications has previously been questioned, and even overturned by findings from unpublished information reported in the CSR [86].

3.1.3 Open Access to Clinical Trials Data

On December 2009 Roche was the first global health-care company to release CSRs after growing concerns over their product Tamiflu [84]. Their policy now allows researcher's access to CSRs and summary reports that have been used for regulatory purposes since 1st January 1999. In 2010 the European medicine agency (EMA) [87] became the first major regulatory agency to agree to an open access policy to confidential documents, including CSRs. However, in 2013 the EMA was forced to take a backwards step, when the general court of the European Union (EU) ordered them to limit the access to their reports due to legal cases from two drug companies [88]. The EMA has since published their final policy on access to documents and CSRs in October 2014 [89], meaning that researchers will now be able to re-assess data sets and obtain CSRs. The FDA has also set up a similar policy, although there access to such material appears much more rigorous (Figure 5). Also in 2013, the pharmaceutical company GlaxoSmithKline (GSK) [90] announced their plans to make their CSRs publically

available through their Clinical Trials Register, and also open access to requested patient level data from GSK clinical trials which are made available through an online request system.

Figure 5: Types of clinical trial data typically held within and transferred between three realms: trial sponsor, regulatory and public
(Permission obtained from Doshi, BMJ Open 2013).



3.2 A Case study

The aim of this case study is to carry out an exploratory review to determine the quality and completeness of reporting harms data within a sample of CSRs, and to compare meta-analysis of harms data from these CSRs against the meta-

analysis based on data extracted from corresponding journal articles. Roche sponsored orlistat trials were selected for this case study.

3.2.1 Roche's Policy on Data Sharing

The Roche Data Sharing Policy is a global policy for both Roche and Genentech on the sharing of clinical trials data. The policy provides the opportunity to request and receive global CSRs and other summary reports. In addition, researchers can obtain access to analyzable patient-level data from clinical trials upon request.

A Roche CSRs typically follows a set structure consisting of five modules of information:

- *Module I*: The 'core report' which includes; background and rationale, objectives, materials and methods, efficacy results, safety results, discussion, conclusion and appendices.
- *Module II*: 'Study documents' including; Protocol and amendment history, blank CRF, subject information sheet and consent form, glossaries of original and preferred terms, randomization list, reporting analysis plan, certificates of analysis, list of investigators and list of ethics committee.
- *Module III*: 'Listings of demographic and efficacy data'.
- *Module IV*: 'Listing of safety data'.
- *Module V*: 'Statistical report and appendices' - Statistical analysis and efficacy results.

3.2.2 Orlistat in obesity research

Orlistat (Trade name: Xenical) which is marketed by Roche in most countries is used in the treatment of obesity, as a selective inhibitor of gastric and pancreatic lipase [91]. Mild but unpleasant Gastrointestinal (GI) side effects are commonly reported with orlistat use. A systematic review [92] including 16 randomized placebo controlled trials of orlistat which estimated the risk of discontinuations due to AEs, reported an increase of 3% (95% CI 1-4%) in risk with the use of orlistat. The most common AEs leading to withdrawal were GI (40%); only eight (50%) trials specified the number of AEs due to GI problems. Another study [93] including 29 trials of orlistat indicated an increase in risk for events; diarrhoea, flatulence, abdominal pain and dyspepsia in orlistat treated patients compared with placebo. No SAEs were reported in these reviews. There is concern that there may also be an associated increased risk of serious hepatic events as indicated in a case series study using primary care data from the Clinical Practice Research Datalink (CPRD) [94].

3.3 Methods

We planned to identify independent trials each of which were reported within two different trial summary reports: CSRs and publically available journal publications. The aim was to compare each trial's summary reports and determine whether there were inconsistencies in quality and quantity of reporting of harms. CSRs were released by Roche (Genentech; South San Francisco, CA) and any analysis was carried out using R version 3.0.2.

3.3.1 Systematic search

A search was implemented in the Cochrane Central register (final search 6 July 2013) and Ovid MEDLINE (final search 2 July 2013) to obtain all relevant published RCTs comparing orlistat against placebo for the treatment of obesity. The search terms used are displayed in Appendix B, Table 28. Each full article was assessed independently by one investigator to determine eligibility. We included published RCTs investigating the use of orlistat. No restriction was placed on the clinical area. Excluded studies were observational studies and those that did not specify orlistat as their primary intervention.

3.3.2 Data collection and extraction

Roche were contacted and asked to provide the corresponding CSRs for each of the trial publications identified. This involved listing all relevant published literature with authors, trial ID and journal title with any additional information about research sponsors, grants etc. Roche were responsible for the 'preparation' and 'redaction' of the CSRs, which involved deleting or blanking out any patient confidential information.

For each matching document pair (CSR and journal publication) the following data were extracted:

- Content and characteristics of both document types: whether a clear primary objective of safety was defined, word count of information relating to harms in both the journal publication (including any online supplementary material) and in the CSR documents of text only (word count performed using the software AnyCount version 7.0 [95]). Missing

pages relating to safety due to redactions were noted in the results, we managed to obtain these upon further requests.

- Name of each reported adverse event (AE) and serious adverse event (SAE) term recorded for both placebo and orlistat, with the number of patients in safety population, as defined in the respective document. The AE coding system used was also detailed.
- Reporting structure of harms (CONSORT-harms [8] used as a benchmark).

One investigator extracted the data (AH), and a second investigator (CTS) checked the data extraction for two of the included trials (Chanoine [96] (Trial ID: NM16189), Halpern [97] (M37013)).

3.3.3 AEs and SAEs

For a particular trial, all harms (AEs and SAEs) reported in either journal publication or CSR were extracted and compared across the two document types. The total number of reported MedDRA preferred terms, were compared. If a MedDRA preferred term was reported in both the CSR and journal publication the numerical data were compared and any discrepancies noted.

For each MedDRA preferred term (AE and SAE) the data extracted from CSRs were pooled across trials using fixed effect meta-analysis. A corresponding meta-analysis was performed using the data extracted from journal publications. The pooled Risk Difference (RD) with 95% confidence interval [98], and the I^2 statistic were compared between CSR and journal publication based analyses [99]. We stress that these meta-analysis results are based on a subset of the

eligible trials of orlistat and are presented for the purpose of a methodological comparison rather than definitive clinical results.

3.3.4 Structured reporting of harms

Using the CONSORT-harms extension [8] as a benchmark for reporting harms data from a RCT, documents were assessed across fifteen adapted criteria (Table 7) that focus on the methods and results.

Each trial was classified as follows for each individual criteria:

BOTH - both documents report the criteria.

CSR - only reported criteria in clinical study report.

Pub - only reported criteria in trial publication.

NR - criteria not reported in either document.

The total number of criteria satisfied in each CSR and journal publication for a particular trial was calculated and expressed as a percentage of the 15 criteria.

Table 7: Fifteen criteria (adapted from the CONSORT-harms extension) assessed to evaluate the completeness of reporting methods and results of harms.

	Criteria	Description of criteria	Description of complete reporting for criteria
Methods	1	List addressed adverse events with definitions.	Listed AEs with definitions (with attention, when relevant, to grading).
	2	Mode for collecting data.	Full description of questionnaires, interviews, or tests used to collect information on harms. Detailed information on questions asked.
	3	Timing and time frame of surveillance.	Description of time frame of surveillance for AEs, with stopping period detailed.
	4	Attribution methods.	Person responsible for making attribution disclosed and whether blinding was used.
	5	Intensity of ascertainment.	Specify clearly how withdrawals are handled in the analyses.
	6	Harms related monitoring.	Plans for monitoring and rules for stopping for benefits and harms separately.
	7	Coding of AEs.	Reference to any coding system used and person responsible for the coding.
	8	Handling of recurrent events.	Specify how recurrent events are handled, detailed as separate events or as one.
	9	Timing issues.	Timing of events if recurrent explained.
	10	Plans to perform any statistical analyses and inferences.	Described how pre-specified statistical analyses are separated from post hoc analyses, and any common problems addresses.
Results	11	Withdrawals and discontinuations.	Reasons for discontinuations and separated by arm. Flow diagrams used to display withdrawals.
	12	Denominators for analyses on harms.	Analyses and definitions used and clearly stated (i.e. Intention To Treat (ITT)), and all denominators for safety population are clearly detailed.
	13	Specifying AE type.	Results presented separately by System Organ Classification type.
	14	Grading or scaling used.	Each AE type should offer appropriate metrics of absolute risk.
	15	Seriousness per arm.	Reported separately for each type of event.

When both document types reported on any particular individual criteria (i.e. BOTH), the reported information was compared and classified as follows:

CSR (+) - The CSR provides more information than the journal publication.
(E.g. full data was provided and/or is reported in text of the CSR but not in the journal publication).

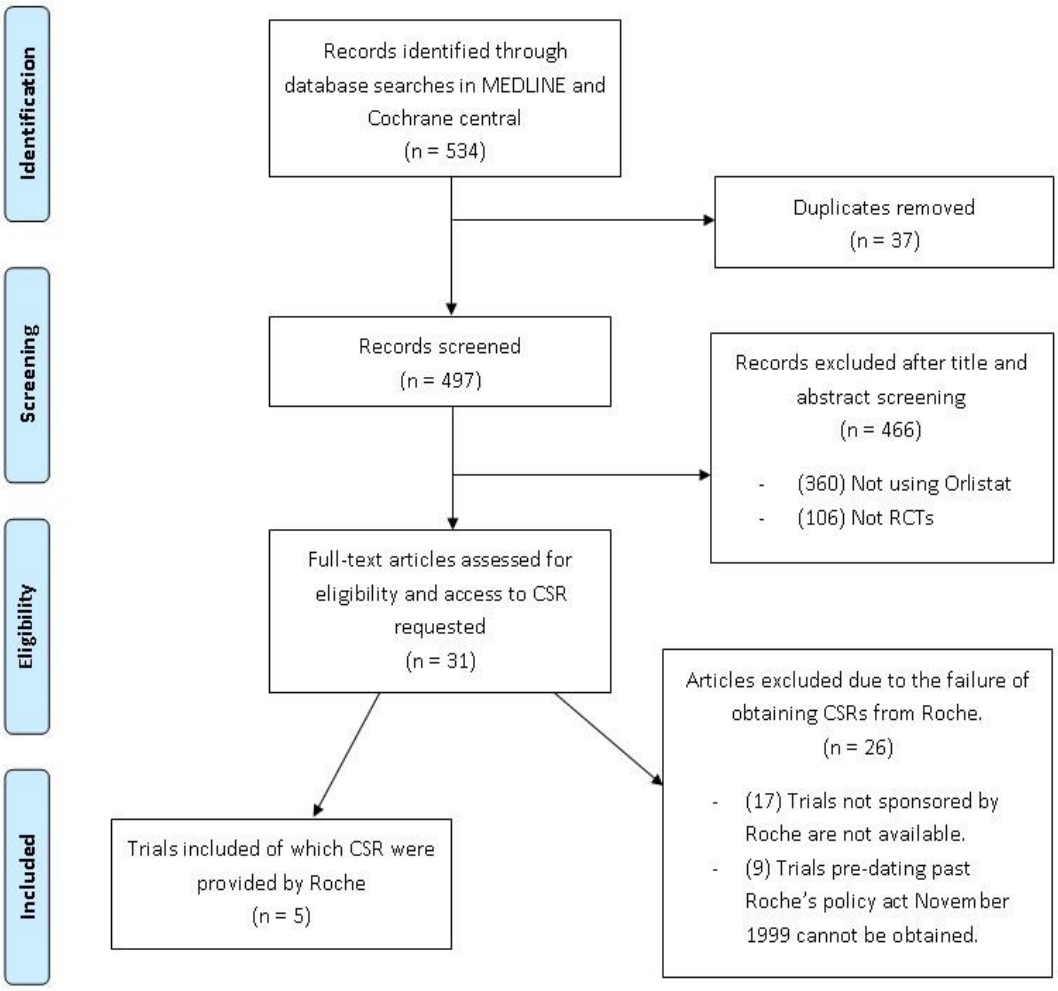
Similar (O) - Both document types provide equal and similar information.

CSR (-) - The journal publication provides more information than the CSR.

3.4 Results

Thirty-one journal publications related to 31 RCTs of orlistat were identified from the search (Figure 6). We requested access to full CSRs from Roche corresponding to each of these trials. The CSRs could not be provided for 26 of these trials: 17 trials were not Roche-sponsored, and CSRs were therefore not held by Roche and 9 trials pre-dated Roche's policy extension, which only allows access to trials dating back to the 1st January 1999.

Figure 6: Flow diagram for obtaining trial reports.



CSRs were obtained and matched with the corresponding journal publication for five trials (Chanoine [96] (Trial ID: NM16189), Halpern [97] (M37013), Hanefeld [100] (M37002), Kelley [101] (M37047) and Torgerson [102] (BM15421)). Module I of the CSR was provided for all trials. Module II was not provided for one trial (BM15421) and module V was not provided for one trial (NM16189). We contacted Roche to provide reasons for any missing sections, and they informed us that these sections contained confidential information and had to

be removed. Modules III and IV were not provided for any of the trial CSRs since they contained individual patient data listings.

Table 8 shows the content and characteristics for each trial document pair. Safety was not the primary objective for any of the five trial journal publications, but was defined as a secondary objective in three journal publications [96, 97, 102], and not specified in two journal publications [100, 101]. Two trials [97, 100] were published in the Journal of Diabetes, Obesity and Metabolism, two trials [101, 102] in the Journal of Diabetes Care, and one trial [96] in the Journal of the American Medical Association (JAMA).

The mean word count across the five trial journal publications was 7265 (Standard deviation (sd) 1894) with an average of 10% of words (mean (sd) 757 (287)) dedicated to safety. The CSRs had a mean (sd) of 163411 (96872) words across all trials, with approximately 3% (mean (sd) 4663 (1446)) related to safety. The mean difference between the CSR and journal publication was 3906 (95% CI (1756, 6056)) words.

Table 8: Content and characteristics of trial documents.

Trial ID	**NM16189**		**M37013**		**M37002**		**M37047**		**BM15421**	
Safety primary objective of trial?	No†		No†		No¥		No¥		No†	
Author, journal of publication and year	Chanoine JAMA (2005)		Halpern Diabetes, Obesity and Metabolism (2003)		Hanefeld Diabetes, Obesity and Metabolism (2002)		Kelley Diabetes Care (2002)		Torgerson Diabetes Care (2004)	
CSR Research report no. (date of CSR)	1011426 (2003)		1002688 (2000)		1003882 (2001)		1002743 (2001)		1008213 (2002)	
Volume of both trial documents										
Trial document	Pub	CSR	Pub	CSR	Pub	CSR	Pub	CSR	Pub	CSR
Total number of words in document	10568	146801	6371	45464	6382	140166	7090	170347	5915	314277
Total number of words relating to safety (% of total)	1147 (10.9)	4883 (3.3)	908 (14.3)	2664 (5.9)	638 (10)	4964 (3.5)	707 (10)	4150 (2.4)	387 (6.5)	6653 (2.1)
CSR Module$^{\Phi}$ supplied by Roche										
I	✓		✓		✓		✓		✓	
II	✓		✓		✓		✓		*	
III	*		*		*		*		*	
IV	*		*		*		*		*	
V	*		✓		✓		✓		✓	

CSR; Clinical Study Report, Pub; Journal publication; † Safety secondary objective; ¥ Objective to assess improvements in glycaemic control, and cardiovascular disease risk; $^{\Phi}$Modules explained in section 3.2.2.

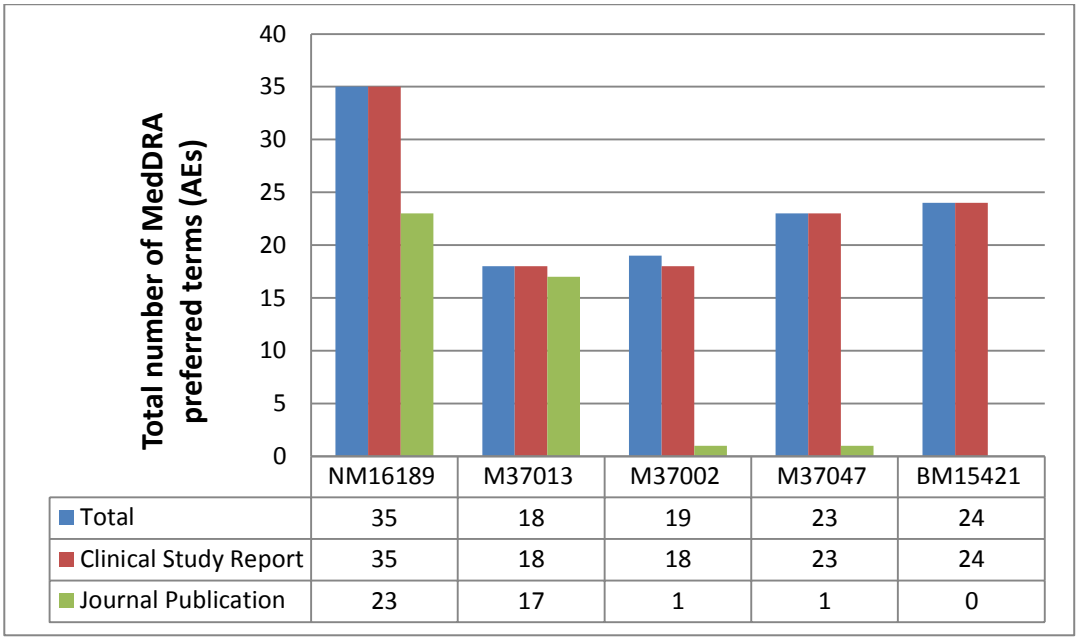
3.4.1 Comparison of reported adverse event and serious adverse event data

MedDRA version 2.3 had been used to code AEs and SAEs in all five trials.

3.4.1.1 Adverse Events

The total number of MedDRA preferred terms for adverse events varied across trials (Figure 7).

Figure 7: The total number of MedDRA preferred term (Adverse Events) reported in CSRs and Journal publications across all five trials.



	NM16189	M37013	M37002	M37047	BM15421
Total	35	18	19	23	24
Clinical Study Report	35	18	18	23	24
Journal Publication	23	17	1	1	0

Journal publications did not always report the complete list of identified MedDRA preferred terms that appeared in the CSR (Table 9). One trial M37013 [97] showed very good consistency between the CSR and journal publication with 18 MedDRA preferred terms for AEs in total, 18 (100%) of which were listed in the CSR and 17 (94%) within the journal publication. One trial NM16189 [96] reported 35 MedDRA preferred terms across the CSR and publication combined, with only 23 (66%) of these listed in the journal publication.

Table 9: Reporting of adverse events in Clinical Study Reports (CSRs) and journal articles for Olistat trials.

Trial ID	NM16189	M37013	M37002	M37047	BM15421
Listed adverse events - MedDRA preferred term					
Gastrointestinal disorders					
Fatty/oil stool	BOTH+	BOTH+	CSR	CSR	CSR
Increased defecation	BOTH+	BOTH+	CSR	CSR	CSR
Liquid stools	BOTH+	BOTH+	NR	CSR	CSR
Oily spotting	BOTH+	NR	CSR	CSR	CSR
Oily evacuation	BOTH+	BOTH+	NR	CSR	CSR
Flatus with discharge	BOTH+	BOTH+	NR	CSR	CSR
Abdominal pain	BOTH+	BOTH-*	CSR	CSR	CSR
Colic Abdominal	NR	CSR	NR	NR	NR
Abdominal pain upper	NR	NR	NR	NR	CSR
Solid stools	NR	BOTH+	NR	NR	NR
Soft stool	BOTH+	NR	NR	CSR	CSR
Flatulence	BOTH+	BOTH+	CSR	CSR	CSR
Decreased defecation	NR	BOTH+	NR	NR	NR
Abdominal distension	NR	BOTH+	CSR	NR	NR
Faecal incontinence	BOTH+	BOTH+	NR	CSR	CSR
Faecal urgency	BOTH+	BOTH+	CSR	CSR	CSR
Faeces discoloured	CSR	NR	NR	CSR	CSR
Gastritis	NR	BOTH+	NR	NR	CSR
Nausea	BOTH+	BOTH+	NR	NR	NR
Pellets	NR	BOTH+	NR	NR	NR

Vomiting	NR	BOTH+	NR	NR	NR
Dry mouth	NR	BOTH+	NR	NR	NR
Dyspepsia	CSR	NR	NR	NR	NR
Haemorrhoids	NR	NR	NR	NR	CSR
Infections and Infestations					
Nasopharyngitis	BOTH+	NR	CSR	CSR	NR
Upper respiratory tract infection	BOTH+	NR	CSR	CSR	NR
Influenza	NR	NR	CSR	NR	CSR
Gastroenteritis viral	NR	NR	NR	CSR	NR
Ear Infection	NR	NR	NR	CSR	NR
Tooth abscess	NR	NR	NR	CSR	NR
Sore throat	BOTH+	NR	NR	NR	CSR
Sinusitis	BOTH+	NR	NR	NR	CSR
Gastroenteritis	BOTH+	NR	NR	NR	NR
Bronchitis	CSR	NR	NR	NR	NR
Pharyngitis	CSR	NR	CSR	NR	NR
Viral infection	CSR	NR	NR	NR	NR
Upper respiratory tract infection viral	CSR	NR	NR	NR	NR
Urinary tract infection	NR	NR	CSR	NR	NR
Metabolism and nutrition disorders					
Hypoglycaemia	NR	NR	Pub	BOTH+	NR
General disorders and administration site conditions					
Influenza like illness	NR	NR	NR	CSR	NR
Pyrexia	NR	NR	NR	NR	CSR

Nervous System disorders					
Headache	BOTH+	NR	NR	NR	NR
Migraine	CSR	NR	NR	NR	NR
Dizziness (Exc. Vertigo)	NR	NR	CSR	CSR	NR
Psychiatric disorders					
Depression	CSR	NR	NR	CSR	CSR
Respiratory, thoracic and medicinal disorders					
Rhinitis seasonal	BOTH+	NR	NR	CSR	CSR
Nasal congestion	BOTH+	NR	NR	NR	NR
Asthma	CSR	NR	NR	NR	NR
Cough	NR	NR	NR	NR	CSR
Skin & Subcutaneous tissue disorders					
Dry skin	NR	NR	NR	CSR	NR
Dermatitis	CSR	NR	NR	NR	NR
Ecchymosis	NR	NR	CSR	NR	NR
Injury and positioning					
Joint sprain	BOTH+	NR	NR	NR	CSR
Back pain	BOTH+	NR	NR	NR	NR
Limb injury	BOTH+	NR	NR	NR	NR
Laceration	CSR	NR	NR	NR	NR
Immune system disorders					
Hypersensitivity	CSR	NR	NR	NR	NR
Musculoskeletal, connective tissue and bone disorders					
Spinal disorder	NR	NR	CSR	NR	NR
Sciatica	NR	NR	CSR	NR	CSR
Pain in limb	NR	NR	CSR	NR	NR

Vascular disorders					
Varicose veins	NR	NR	CSR	NR	NR
Total number of adverse events terms reported across CSR and publication	35	18	19	23	24
Total number of adverse event terms reported in CSR (% of total)	35 (100)	18 (100)	18 (95)	23 (100)	24 (100)
Total number of adverse event terms reported in journal publication (% of total)	23 (66)	17 (94)	1 (5)	1 (4)	0 (0)

BOTH+ = 'reported in CSR and the corresponding journal publication and agreed in data'; BOTH- = 'reported in CSR and the corresponding full academic journal publication but disagreed in numerical value'; CSR = 'only reported within the CSR'; Pub = 'only reported in the journal publication (specific events may have been subsumed under other categories in summary reports)'; NR = 'neither reported in the CSR or journal publication'.

* Results in CSR: 10 (5.9% of patients), Journal publication: 13 (7.7% of patients) in orlistat treatment arm.

There was very poor consistency for three trials (M37002 [100]; M37047 [101]; BM15421 [102]) with 5% or fewer of the total MedDRA preferred terms being reported in the journal publication (M37002: 1 (5%); M37047: 1 (4%); BM15421: 0 (0%)).

When a MedDRA preferred term was listed in both the CSR and journal publication, there was complete agreement in the numerical results (Table 9) except for one case in trial M37013 [97]. Where there were 3 additional patients with abdominal pain on orlistat identified within the journal publication.

3.4.1.2 Meta-analysis for AEs

In total 61 individual MedDRA preferred terms for AEs were reported in either the CSR or journal publication across the five trials (Table 10). 30 (49%) of these terms were reported in the CSR and corresponding journal publication for at least one trial allowing a comparison of pooled results. In all 30 meta-analysis (MA) comparisons there was agreement in the direction of effect of pooled results. However, in 6 (20%) MA comparisons the magnitude of effect differed (the 95% CI for the pooled risk difference (RD) did not overlap between the CSR and journal publication results). In particular for the MedDRA preferred terms of 'increased defecation', 'oily spotting', 'oily evacuation', and 'faecal incontinence' the pooled RD from journal publications was greater than CSRs (highlighted in red) whereas for 'soft stools' and 'faecal urgency' the pooled RD from CSRs was greater than from journal publications (highlighted in blue).

Table 10: Meta-analysis for adverse events (AEs) reported at least once in CSR and corresponding journal publication.

These meta-analysis results are based on a subset of the eligible trials of orlistat and are presented for the purpose of methodological comparison rather than definitive clinical results.

						Pooled Estimates		Heterogeneity	
Event	**Document**	**Orlistat events**	**Total randomised to orlistat**	**Placebo events**	**Total randomised to placebo**	**Risk Difference (RD) (accurate to 2 dp's)**	**95% CI (accurate to 2 dp's)**	**I^2 (%)**	**P - Value**
Gastrointestinal disorders									
Fatty/oil stool	Publication	254	526	33	350	0.38	(0.33, 0.44)	56.4	0.1299
	CSR	1133	2586	144	2358	0.37	(0.35, 0.39)	73.2	0.0048
Increased defecation	Publication	116	526	35	350	0.14	(0.10, 0.19)	95.3	<0.0001
	CSR	570	2586	344	2358	0.08	(0.05, 0.10)	86.7	<0.0001
Liquid stools	Publication	17	174	8	168	0.05	(0, 0.10)	NA	NA
	CSR	430	2081	323	2075	0.05	(0.03, 0.07)	0	0.09799
Oily spotting	Publication	102	352	7	181	0.25	(0.20, 0.31)	NA	NA
	CSR	345	2412	62	2189	0.10	(0.08, 0.11)	96.2	<0.0001
Oily evacuation	Publication	99	526	3	181	0.22	(0.17, 0.26)	98.2	<0.0001
	CSR	178	2442	9	1836	0.07	(0.06, 0.08)	98.3	<0.0001
Flatus with discharge	Publication	83	526	8	350	0.12	(0.09, 0.16)	92.1	0.0004
	CSR	301	2442	29	2274	0.11	(0.09, 0.12)	75.6	0.0064
Abdominal pain	Publication	90	526	26	350	0.08	(0.04, 0.12)	71.5	0.0613
	CSR	360	2586	236	2358	0.04	(0.02, 0.05)	37	0.1746
Colic Abdominal	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	3	174	0	169	0.02	(-0.01, 0.04)	NA	NA
Abdominal pain upper	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	160	1649	157	1655	0	(-0.02, 0.02)	NA	NA

Solid stools	Publication	11	174	9	169	0.01	(-0.04, 0.06)	NA	NA
	CSR	11	174	9	169	0.01	(-0.04, 0.06)	NA	NA
Soft stools	Publication	53	352	19	181	0.05	(-0.01, 0.10)	NA	NA
	CSR	606	2268	327	2105	0.12	(0.10, 0.14)	80.7	0.0056
Flatulence	Publication	37	526	13	350	0.03	(0, 0.06)	68.4	0.0752
	CSR	486	2586	383	2358	0.04	(0.01, 0.06)	51.4	0.0837
Decreased defecation	Publication	4	174	20	169	-0.1	(-0.15, -0.04)	NA	NA
	CSR	4	174	20	169	-0.1	(-0.15, -0.04)	NA	NA
Abdominal distension	Publication	3	174	3	169	-0.01	(-0.03, 0.03)	NA	NA
	CSR	56	318	35	253	-0.01	(-0.06, 0.05)	0	0.7236
Faecal incontinence	Publication	33	526	1	350	0.08	(0.05, 0.11)	NA	NA
	CSR	119	2442	11	2105	0.04	(0.04, 0.05)	84.3	0.0017
Faecal urgency	Publication	75	526	24	350	0.05	(0.01, 0.09)	95.5	<0.0001
	CSR	474	2586	138	2358	0.12	(0.10, 0.13)	95.9	<0.0001
Faeces discoloured	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	65	2268	15	2105	0.02	(0.01, 0.03)	0	0.4391
Gastritis	Publication	2	174	0	169	-	-	NA	NA
	CSR	85	1823	82	1824	0	(-0.01, 0.02)	NA	NA
Nausea	Publication	54	526	28	350	0	(-0.03, 0.04)	58.9	0.1187
	CSR	54	526	28	350	0	(-0.03, 0.04)	58.9	0.1187
Pellets	Publication	2	174	6	169	-0.02	(-0.06, 0.01)	NA	NA
	CSR	2	174	6	169	-0.02	(-0.06, 0.01)	NA	NA
Vomiting	Publication	2	174	2	169	0	(-0.02, 0.02)	NA	NA
	CSR	2	174	2	169	0	(-0.02, 0.02)	NA	NA
Dry mouth	Publication	0	174	2	169	-0.01	(-0.03, 0.01)	NA	NA
	CSR	0	174	2	169	-0.01	(-0.03, 0.01)	NA	NA
Dyspepsia	Publication	NR	NR	NR	NR	NR	NR	NR	NR

	CSR	12	352	5	181	0.01	(-0.02, 0.04)	NA	NA
Haemorrhoids	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	16	1649	38	1655	-0.01	(-0.02, 0)	NA	NA
Infections and Infestations									
Nasopharyngitis	Publication	99	352	46	181	0.03	(-0.05, 0.11)	NA	<0.0001
	CSR	164	763	97	534	0.02	(-0.02, 0.07)	0	0.9902
Upper respiratory tract infection	Publication	114	352	48	181	0.06	(-0.02, 0.14)	NA	<0.0001
	CSR	120	496	52	265	0.04	(-0.02, 0.10)	62.7	0.1014
Influenza	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	360	2060	325	2008	0.02	(-0.01, 0.04)	0	0.7201
Gastroenteritis Viral	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	10	267	10	269	0	(-0.03, 0.03)	NA	NA
Ear infection	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	8	267	2	269	0.02	(0, 0.05)	NA	<0.0001
Tooth abscess	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	7	267	6	269	0	(-0.02, 0.03)	NA	NA
Sore throat	Publication	59	352	29	181	0.01	(-0.06, 0.07)	NA	NA
	CSR	111	2001	74	1836	0	(-0.01, 0.02)	0	0.9183
Sinusitis	Publication	40	352	19	181	0.01	(-0.05, 0.06)	NA	NA
	CSR	111	2001	74	1836	0.01	(0, 0.02)	0	0.9665
Gastroenteritis	Publication	23	352	8	181	0.02	(-0.02, 0.06)	NA	NA
	CSR	23	352	8	181	0.02	(-0.02, 0.06)	NA	NA
Bronchitis	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	17	352	6	181	0.02	(-0.02, 0.05)	NA	NA
Pharyngitis	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	11	352	4	181	0.01	(-0.02, 0.04)	NA	NA
Viral infection	Publication	NR	NR	NR	NR	NR	NR	NR	NR

	CSR	10	352	3	181	0.01	(-0.01, 0.04)	NA	NA
Upper respiratory tract infection viral	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	8	352	1	181	0.02	(0, 0.04)	NA	NA
Urinary tract infection	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	11	144	8	84	-0.02	(-0.1, 0.06)	NA	NA
Metabolism and nutrition disorders									
Hypoglycaemia	Publication	47	411	30	353	0.04	(0, 0.09)	90.1	0.0015
	CSR	45	267	26	269	0.07	(0.01, 0.13)	NA	NA
General disorders and administration site conditions									
Influenza like illness	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	13	267	9	269	0.02	(-0.02, 0.05)	NA	NA
Pyrexia	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	53	1649	47	1655	0	(-0.01, 0.02)	NA	NA
Nervous System disorders									
Headache	Publication	134	352	56	181	0.07	(-0.01, 0.16)	NA	NA
	CSR	134	352	56	181	0.07	(-0.01, 0.16)	NA	NA
Migraine	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	9	352	2	181	0.01	(-0.01, 0.04)	NA	NA
Dizziness (Exc. Vertigo)	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	18	411	8	353	0.02	(0, 0.05)	0	0.8653
Psychiatric disorders									
Depression	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	62	2268	43	2105	0.01	(0, 0.02)	0	0.9346
Respiratory, thoracic and medicinal disorders									
Rhinitis seasonal	Publication	21	352	9	181	0.01	(-0.03, 0.05)	NA	NA
	CSR	69	2268	36	2105	0.01	(0, 0.02)	0	0.9572
Nasal congestion	Publication	31	352	11	181	0.03	(-0.02, 0.07)	NA	<0.0001

	CSR	31	352	11	181	0.03	(-0.02, 0.07)	NA	<0.0001
Asthma	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	8	352	3	181	0.01	(-0.02, 0.03)	NA	<0.0001
Cough	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	35	1649	27	1655	0	(0, 0.01)	NA	NA
Skin & subcutaneous tissue disorders									
Dry skin	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	7	267	0	269	0.03	(0.01, 0.05)	NA	NA
Dermatitis	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	16	352	3	181	0.03	(0, 0.06)	NA	<0.0001
Ecchymosis	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	5	144	3	84	0	(-0.05, 0.05)	NA	NA
Injury and positioning									
Joint sprain	Publication	35	352	17	181	0.01	(-0.05, 0.06)	NA	NA
	CSR	78	2001	59	1836	0	(-0.01, 0.01)	0	0.8484
Back pain	Publication	28	352	11	181	0.02	(-0.03, 0.06)	NA	NA
	CSR	28	352	11	181	0.02	(-0.03, 0.06)	NA	NA
Limb injury	Publication	18	352	5	181	0.02	(-0.01, 0.06)	NA	NA
	CSR	18	352	5	181	0.02	(-0.01, 0.06)	NA	NA
Laceration	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	12	352	4	181	0.01	(-0.02, 0.04)	NA	NA
Musculoskeletal, connective tissue and bone disorders									
Hypersensitivity	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	11	352	2	181	0.02	(0, 0.04)	NA	<0.0001
Spinal disorder	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	12	144	11	84	-0.05	(-0.13, 0.04)	NA	NA
Sciatica	Publication	NR	NR	NR	NR	NR	NR	NR	NR

	CSR	45	411	43	353	0	(-0.05, 0.04)	0	0.517
Pain in limb	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	4	144	3	84	-0.01	(-0.06, 0.04)	NA	NA
Vascular disorders									
Varicose Veins	Publication	NR	NR	NR	NR	NR	NR	NR	NR
	CSR	4	267	0	84	0.03	(0, 0.06)	NA	NA

P-value represents significance of the heterogeneity between studies reporting the event within CSR or journal publication.

NA – Only one study reports the event and therefore can't calculate heterogeneity or P-value.

For the 31 MedDRA preferred terms that had only been reported in a CSR, 23 (74%) analyses suggested an increased risk of an adverse event on orlistat, 2 (6%) of which were statistically significant (faeces discolouration and dry skin). For 4 (13%) MedDRA preferred terms there was no difference between orlistat and placebo and for a further 4 (13%) MedDRA preferred terms there was a suggestion of an increased risk of an event with placebo, 1 (3%) of which was statistically significant (haemorrhoids). The one MedDRA preferred term hypoglycaemia was reported only in the journal publication for trial M37047 [101].

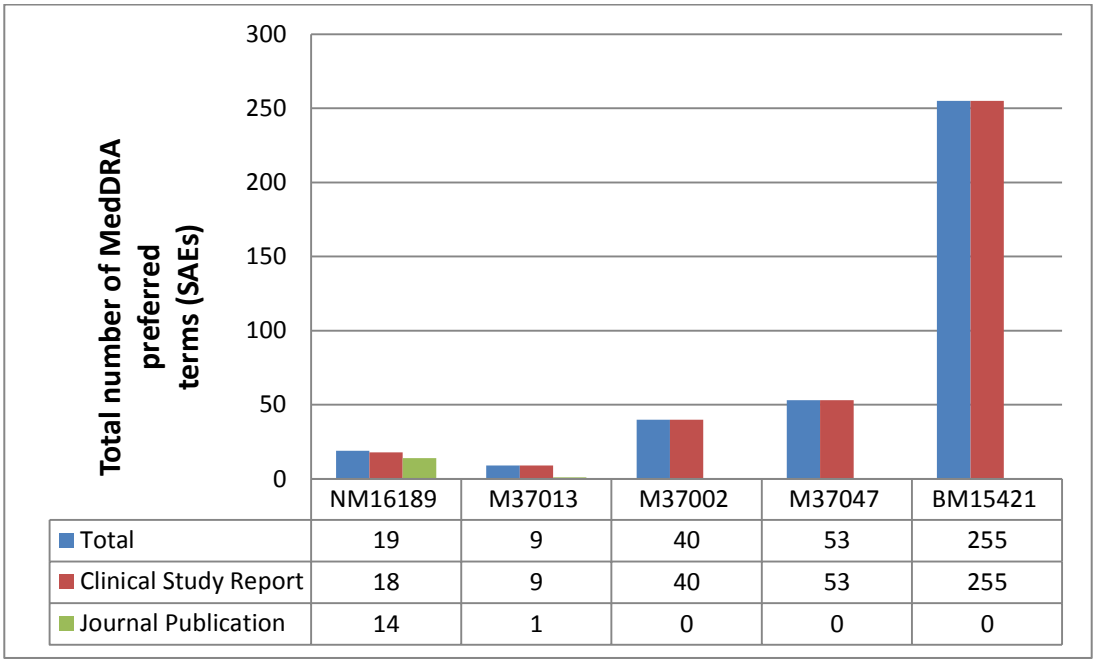
3.4.1.3 Serious Adverse Events

The total number of MedDRA preferred terms for SAEs varied across trials (Figure 8). One trial NM16189 [96] showed good consistency between the CSR and journal publication with 19 MedDRA preferred terms for SAEs in total, 18 (95%) of which were listed in the CSR and 14 (74%) within the journal publication. There was very poor consistency for four trials (M37013 [97], M37002 [100], M37047 [101], BM15421 [102]) with 11% or fewer of the total MedDRA preferred terms being reported in the journal publication (M37013: 1 (11%); M37002: 0 (0%); M37047: 0 (0%); BM15421: 0 (0%)).

In trial NM16189 [96] there were 19 SAEs terms reported across the CSR and journal publication. 13 of these were reported in both documents, either with full numerical agreement (12 SAE terms), or with disagreement in numerical results (1 depression SAE on orlistat reported in the CSR and 2 depression SAEs reported in the journal publication) (See Appendix B, Table 29). Five SAE terms

were only reported within the CSR (demyelination (1) and bronchospasm aggravated (1) on placebo, and convulsions (1), suicidal ideation (1) and liquid stools (1) on orlistat). Encephalomyelitis SAE was reported for placebo within the publication but not the CSR.

Figure 8: The total number of serious adverse events reported in CSRs and Journal publications across all five trials.



	NM16189	M37013	M37002	M37047	BM15421
Total	19	9	40	53	255
Clinical Study Report	18	9	40	53	255
Journal Publication	14	1	0	0	0

Trial M37013 [97] reports 9 SAEs with only “diarrhoea and dehydration” on orlistat reported in both documents. The remaining 8 SAEs were only reported in the CSR; death (1), diabetes mellitus (1), hysterectomy and perineoplasty (1), mitral lesion (1) on placebo and cholaeistiny due to chronic cholelithiasis (1), nephrectomy due to previous renal carcinoma (1), nephrectomy and lithotripsy due to previous nephrolithiasis (1), ovary carcinoma and ascites (1) on orlistat. The three remaining trials (M37002 [100], M37047 (21) and BM15421 [102]) report a high number of SAEs (40, 53 and 255) within the CSR that have not been reported in the corresponding journal publication.

3.4.1.2 Meta-analysis of SAEs

In total 326 MedDRA preferred terms for SAEs were reported in either the CSR or journal publication across the five trials (Appendix B, Tables 30, 31 and 32). 14 (4%) of these terms were reported in the CSR and corresponding journal publication for at least one trial allowing a comparison of the pooled results. However, in 1 (7%) MA comparison the magnitude of effect differed (the 95% CI for the pooled risk difference (RD) did not overlap between the CSR and journal publication results). In particular for the MedDRA preferred term 'depression' the pooled RD from the journal publication was greater than the CSR (Table 30). For the 311 (95%) MedDRA preferred terms that had only been reported in a CSR, 16 (5%) analyses suggested an increased risk of a SAE on orlistat, 2 (13%) of which were statistically significant (carotid artery stenosis, varicose veins) (Table 31). The MedDRA preferred term 'encephalomyelitis' which was only reported in the journal publication, was non-significant (Table 32).

3.4.2 Structured Reporting

The quality of reporting harms related information, as assessed against the 15 criteria adapted from the CONSORT-harms checklist, are displayed in Table 11.

Table 11: Comparison of 15 harms criteria (CONSORT-harms used as a benchmark).

			Trial ID				
	Criteria	**Description of item**	**NM16189**	**M37013**	**M37002**	**M37047**	**BM15421**
Methods Criteria	**1**	List addressed adverse events with definitions.	CSR	CSR	CSR	CSR	CSR
	2	Mode of collecting harms data.	BOTH $_{O}$	BOTH $_{O}$	BOTH $_{O}$	CSR	BOTH $_{+}$
	3	Timing and time frame of surveillance for adverse events.	BOTH $_{O}$	Pub	CSR	NR	BOTH $_{+}$
	4	Attribution methods.	CSR	NR	CSR	NR	NR
	5	Intensity of ascertainment.	CSR	BOTH $_{O}$	CSR	CSR	CSR
	6	Harms related monitoring.	CSR	BOTH $_{O}$	CSR	CSR	CSR
	7	Coding of AEs.	CSR	CSR	BOTH $_{+}$	CSR	CSR
	8	Handling of recurrent events.	NR	CSR	NR	CSR	NR
	9	Timing issues.	CSR	CSR	CSR	NR	CSR
	10	Plans to perform any statistical analyses and inferences.	CSR	BOTH $_{+}$	BOTH $_{+}$	BOTH $_{+}$	BOTH $_{+}$
	Total items satisfied for methods criteria in CSR (% of total 10 items assessed)		9 (90)	8 (80)	9 (90)	7 (70)	8 (80)
	Total items satisfied for methods criteria in publication (% of total 10 items assessed)		2 (20)	5 (50)	3 (30)	1 (10)	3 (30)
Results criteria	**11**	Withdrawals and discontinuations.	BOTH $_{+}$	BOTH $_{+}$	BOTH $_{+}$	BOTH $_{+}$	CSR
	12	Denominators for analyses on harms.	BOTH $_{O}$	BOTH $_{O}$	BOTH $_{+}$	CSR	BOTH $_{O}$
	13	Specifying AE type.	BOTH $_{+}$	BOTH $_{+}$	BOTH $_{+}$	BOTH $_{+}$	BOTH $_{+}$
	14	Grading or scaling used.	NR	BOTH $_{+}$	BOTH $_{+}$	BOTH $_{+}$	BOTH $_{+}$
	15	Seriousness per arm.	BOTH $_{+}$	BOTH $_{+}$	BOTH $_{+}$	BOTH $_{+}$	BOTH $_{+}$
	Total items satisfied for results criteria in CSR (% of total 5 items assessed)		4 (80)	5 (100)	5 (100)	5 (100)	5 (100)
	Total items satisfied for results criteria in publication (% of total 5 items assessed)		4 (80)	5 (100)	5 (100)	4 (80)	4 (80)
Total items satisfied in CSR (% of total 15 items assessed)			13 (87)	13 (87)	14 (93)	12 (80)	13 (87)
Total items satisfied in publication (% of total 15 items assessed)			6 (40)	10 (67)	8 (53)	5 (33)	7 (47)

BOTH = 'reported in CSR and the corresponding journal publication'; CSR = 'only reported within the CSR'; Pub = 'only reported in journal publication'; NR = 'neither reported in the CSR or journals publication'. Completeness of data where agreement (BOTH) is made coded as: + 'More complete in CSR'; O 'Similar quality for both documents'; - 'less complete in the CSR'.

The CSRs satisfied 70-90% of the methods related criteria across the 5 trials compared to the journal publications that satisfied between 10-50%. CSRs consistently provided much greater detail regarding planned analyses than the journal publication and on only one occasion did the journal publication provide greater detail than the CSR (trial M37013 [97]; item 3: timing and time frame of surveillance for AEs). Both CSRs and journal publications satisfied 80-100% of criteria within their results sections, but greater detail was generally provided in the CSR. This included full summary tables of AEs and SAEs data, including withdrawals due to harm, severity grading and denominators for the numbers included in the safety population.

3.5 Discussion

Our analysis showed differences in the completeness and quality of reporting harms related information between journal publications and CSRs. A substantial amount of information on patient-relevant harm outcomes, including SAEs, required for unbiased trial evaluation was missing from the publicly available journal article. Including the extra data reported in CSRs altered the magnitude of pooled risk difference estimates in a few cases. Furthermore, there were several MedDRA preferred terms which had never been reported in the corresponding journal publications for this subset of trials. Therefore, restricting evidence synthesis to journal publications would effectively miss these potential harms.

Our meta-analyses were based on a subset of the eligible trials of orlistat and are presented for the purpose of methodological comparison rather than

definitive clinical results. However the results from journal publications in this study are similar to findings from past studies [92, 93] assessing the safety of orlistat in a more detailed meta-analysis (restricted only to journal publications) including more trials. The most commonly reported AEs related to gastrointestinal effects, with increased risks of flatulence, abdominal pain and dyspepsia in orlistat treated patients compared with placebo.

Where there was agreement for reporting on certain harms criteria related to methods and results, information in the publication lacked detail and completeness compared with the CSR. Journal publications are often impeded by word count restrictions, which results in inadequate reporting of harms data. This is still noticeable even after the release of the CONSORT-harms extension [8], as the findings from our recent review [55] suggest. In contrast CSRs have no such word restrictions imposed and theoretically all relevant information should be included. Our study shows that the content of safety information available in the CSR is superior.

A recent study [80] which compared the information gained from CSRs as compared with publically available sources (journal publications and registry reports), reported that CSRs provided considerably more information on harm outcomes. Over 86% of all harm outcomes (AEs and SAEs) were available from the CSRs, compared to only 26% from the journal publications. Combining harms data from registry reports and journal publications increased the proportion of outcomes to 43%. Furthermore, withdrawals due to AEs were detailed completely in 91% of CSRs, with only 51% of journal publications providing

complete information. In another study [16] inadequate safety reporting was shown in the Medtronic manufactured product, recombinant human bone morphogenetic protein 2 (rhBMP-2) used in spinal fusion surgery. Harms data were found to be missing from the publications, with considerably more data found in confidential reports, including the corresponding trial CSRs.

Further evidence of inadequate reporting of benefits and harms were found in a more recent study investigating the product duloxetine in patients with major depressive disorder [103]. The CSRs were found to contain extensive data on major harms that were unavailable in journal publications and in trial registry reports. The study also reports inconsistencies between protocols and CSRs and within CSRs. The value of this missing data could have a major impact on the safety of the product in a systematic review of adverse effects based solely on publically available data from journal publications.

In our study we performed meta-analysis on all reported harms data which allowed us to obtain results that would have been available from restricting analyses to journal publications as might be done in a traditional evidence synthesis, and compare those against results incorporating all the available evidence from CSRs for the 5 included trials. To our knowledge such a methodological comparison has not been published previously. However, the meta-analysis results do not provide comprehensive unbiased clinical results as they are based only on a subset of the 5 orlistat trials. Therefore a broader selection of trials would be necessary to address the standards of harm reports in general. The 26 remaining trials were excluded from this methodological

comparison because they were not sponsored by Roche or pre-dated Roche's policy act, and therefore CSRs could not be provided.

Of the five CSRs obtained from Roche we did not receive a full CSR for any trial. Some of the reports failed to include any information from modules II, III, IV and V, and some CSRs had missing pages with information of AEs removed. Therefore results in this study were based only on the information available, though we were able to analyze all reported harms data in this methodological comparison. We contacted Roche to provide reasons for these missing pages and they explaining that confidential patient listings were detailed, and therefore had to be redacted. Additionally Modules III and IV within the Roche CSRs also contained confidential patient data and were therefore redacted. Orlistat was granted approval by the European Medicines Agency (EMA) on 21 January 2009, however due to the legal proceedings and the limited access over the last year we were unable to obtain further reports via the EMA.

We did not undertake detailed clinical assessments of the causality and relatedness of the AEs and SAEs that had not been reported in journal publications. The CSR does state that most events were either unrelated or remotely related and so it could be that the journal publication authors decided not to report all events, or were limited due to restricted journal space. The assessment of relatedness needs to be carried out. In addition, none of the journal publications mentioned that they had only reported a subset of possible harms data, and none had described a rationale for this decision. The CSRs also indicate that only commonly observed AEs (defined as those events with

incidence rate in orlistat group of ≥ 5%) were summarized, meaning that there are potentially more AEs unreported even in the CSR. Clear definition of SAEs was not provided, particularly for those missing from the journal publication. We also did not study the effect of grade and attributions might have on the omission and inconsistency of reporting. Sensitivity analysis considering each of these key points should be performed.

Furthermore the MAs were conducted without any adjustments for multiplicity, meaning that the results could be misleading when discussing statistically significant differences between orlistat and placebo. The risk difference was used as a measure of inconsistency between document types, however the RD can often be biased and misleading when detecting rare events [104]. Therefore other statistical measures should be considered. Nevertheless, this methodological comparison showed statistically significant differences for certain AEs and SAEs only reported in the CSRs; their also appears to be a systematic trend with suppressed results from journal publications as being more detrimental to orlistat. However some have shown suppressed trends in the opposite direction which should be investigated in further work. CSRs are only developed by commercial companies when submitting applications for marketing approval and so this investigation is focused on the completeness of reporting of harms in journal publications of commercial trials. Similar issues could be apparent in non-commercial trials but this could not be explored here.

Our findings suggest that CSRs produce more complete and robust information on harms data collected in clinical trials compared to publically available journal

publications. However, inconsistencies of harms reporting in this case study were not sufficient enough to raise any serious concerns about the use of orlistat, and therefore including unpublished data from the CSRs did not alter the magnitude of the results in the meta-analysis. Signals of potential harm for a product have been raised in systematic reviews [16] of published literature when the numbers of events are suspected to be too small or even missing, giving rise to considerable uncertainty and inconclusive findings. CSRs should be considered in similar cases whenever there is uncertainty about the efficacy and safety of a product.

Given some of the major pitfalls involved when accessing CSRs, this will likely dissuade systematic reviewers to even consider their potential inclusion in evidence synthesis of harms. Perhaps a more viable solution appears to be that journals should require more thorough reporting of harms via online supplements (e.g., CSRs, de-identified case report forms (CRFs), study protocols and complete tables of AE related information). Also reviewing CSRs can be difficult, as they are extremely lengthy documents and therefore represent a considerable challenge to researchers. Therefore there is a need to develop tools and methodological approaches that will reduce the workload and still allow researchers to use them in an accurate and efficient manner.

Alternatively, where CSRs may not be available upon requests, trial registry reports can sometimes provide additional information from journal publications. However, as highlighted in two recent studies [80, 103] access to these reports is not an adequate alternative to access to CSRs. In addition to CSRs, reviewers

may also consider the complete case report forms (CRFs) to support a synthesis of harms [105]. Though like CSRs, CRFs are usually restrictive and are held by sponsors or regulatory agencies.

The debate around disclosure and clinical trial data release will undoubtedly continue with various stakeholders including funders, academics, industry, publishers and regulators supporting the move towards greater transparency. The new EU clinical trial regulation [106] published on 27th May 2014 supports this claim under section (67). The guideline states that trial data should be publically accessible and presented in an easily searchable format, with related data and documents (including trial protocol and CSR) linked together by the EU trial number. The BMJ also stated that it will no longer publish trials of drugs or devices where the authors do not commit to making the relevant anonymised patient level data available, this is due to be extended to all submitted clinical trials from the 1st of July. The EMA have now adopted the new policy making clinical trials data more accessible [89]. Roche should also be commended for voluntarily submitting their data and allowing further access to their CSRs. Our research provides further empirical evidence supporting the potential value of the CSR.

Further efforts are also needed to improve trial reporting in journal publications, including training for authors and peer reviewers. The EQUATOR network (Enhancing the QUAlity and Transparency Of health Research) aims to promote the use of reporting guidelines and good research reporting practices which should act as the first step to help improve reporting [107]. We also recommend

that authors should make it clear in the journal publication when reporting a subset of harms, and justify why they are doing this, and where the full information can be obtained.

Chapter 4: Sources for Identifying Information about Harms

Chapter 2 has highlighted inadequacies of reporting harms in RCTs and chapter 3 explored the use of CSRs as an alternative approach to obtaining additional detailed information about existing harms data. But, as discussed, CSRs will only provide harms data for a subset of possible trials and there are of course many other sources of harms data that could be exploited. This could be of particular value for the purpose of evidence synthesis and for designing new RCTs where there may be limited, or inadequate, information available from traditional journal publications. For example, we may wish to summarise existing evidence to inform a sample size calculation if the primary outcome of a new RCT is based on harms. We may wish to summarise the existing information about harmful effects of treatments to help guide the safety monitoring of a new RCT, or we may wish to update the existing evidence with harms data collected in a new RCT.

In this chapter, work previously discussed by Loke et al. [108] and the Cochrane AEMG [41, 47] for guidance on selecting and retrieving information about harms to include in evidence synthesis will be outlined in section 4.1, and then an in-depth overview of available data sources that provide information about harms will be discussed in section 4.2.

4.1 Why is a structured approach needed?

Systematic reviews often rely on searches of electronic databases of published articles. Identifying and selecting relevant harms of treatment and quantifying the risk associated with them, however, often require a broader range and more comprehensive assessment of different data sources. In addition, the types of studies included in a systematic review may influence the quality or amount of evidence regarding harms.

4.1.1 Importance of the research question

The Cochrane AEMG recently proposed a framework for a structured approach to conducting systematic reviews of harms [41, 47]. The starting point of the evaluation and subsequent synthesis of harms data in this framework are guided entirely by the research question, which can be "broad" or "narrow" in scope. For example, a review with a broad scope might ask "what harms are associated with antidiabetic drugs commonly prescribed to treat patients diagnosed with type 2 diabetes?" Or, a more narrowly focused review might examine the risk of heart failure in patients with type 2 diabetes who take antidiabetic drugs. The advantages and disadvantages of addressing broad and narrow questions are discussed in Table 12.

4.1.2 A Framework based on the Research Question

As outlined in Loke [108] the scope of the research question i.e., broad or narrow, will determine whether a '*hypothesis generating*' or '*hypothesis testing or strengthening*' approach is needed to select and identify harms data. Mann

[109] has also proposed a similar approach when conducting studies in pharmacoepidemiology research, which we also incorporate into this approach.

Table 12: Advantages and Disadvantages of selecting a broad versus narrow research question for a systematic review of harms.

Scope of question	Advantages	Disadvantages
Broad Example: What common harms might a patient diagnosed with type 2 diabetes experience when taking antidiabetic drugs?	Wider coverage and can evaluate new harms that we may not have previously been aware of. Can also be used preliminary to a narrow approach, to identify specific harms of interest to investigate further.	Danger of being swamped by vast quantities of heterogeneous data and of inappropriate pooling. Can be resource intensive and may yield a diverse amount of information from which it is difficult to draw any meaningful conclusions.
Narrow, usually evaluating only a selected harm outcome in detail. Example: Does the antidiabetic drug Rosiglitazone increase the risk of heart disease or heart failure in patients diagnosed with type 2 diabetes?	Easiest approach, especially with regard to data collection. Hypothesis-testing design allows reviews to focus on important harms and reach conclusions about treatment decision.	Conclusions are limited to specific harms, and do not provide complete picture of the overall safety profile. Only appropriate for harms known in advance.

4.1.2.1 Hypothesis Generating

In hypothesis generating the researcher will investigate a broad overview of safety problems associated with a particular intervention. The first step would be to check summary products characteristics (SmPCs), drug analysis prints (DAPs) and published case reports. RCTs and observational studies can then be

used to help identify harms in published literature. Regardless of the data source used, a generated hypothesis often relates to an association which is considered important to investigate further, meaning that there could be a possible causal relationship between an adverse reaction and a drug.

4.1.2.2 Hypothesis Testing or Strengthening

Hypothesis-testing studies aim to prove whether any suspicions that may have been raised in the hypothesis generation stages are justified [109]. That is to determine whether a specific harm is likely to have been caused by the drug, or whether bias or confounding is likely. This will typically involve calculating the magnitude of risk (relative risk or odds ratio) and degree of uncertainty (95% confidence interval). The selection of the most appropriate study designs (RCT or observational study) in hypothesis testing studies can vary depending on the characteristics of the specific adverse effect. If time and resources are limited, the simplest approach is to check all relevant RCTs first, and if no reliable estimates are available, then it is sensible to proceed with observational studies.

Alternatively a more comprehensive but research intensive approach is to compare findings from both study designs and consider whether appropriate to combine together [14, 110]. For example in one study [111], data from both observational studies and RCTs were combined to present a single estimate of mortality associated with chronic usage of non-steroidal anti-inflammatory drugs (NSAIDs). For some reviews it may only be appropriate to quantitatively combine results from one or some study designs (e.g., RCTs and cohort studies)

and synthesise data from other types of studies (e.g., case series and case reports) using a narrative approach.

Hypothesis strengthening studies aim to determine whether the occurrence of an AE has any relationship with dose, duration of treatment, and characteristics of the patients [112]. This may involve assembling a cohort of published cases and/or spontaneous reports; however retrieving observational studies that have formally estimated the risk of harms is the best approach.

Before proceeding in the synthesizing of data, it is important to discuss the complexities surrounding the three key areas of review methodology that include: the study designs that are most likely to yield robust data on harms, a search strategy for locating and identifying the studies, and considering the diverse range of data sources available when researching the characteristics of the adverse effect fully.

4.1.3 What types of studies to include?

The types of studies included in a systematic review may influence the quality or amount of evidence regarding harms. Type II errors (wrongly concluding that there was no significant difference in harms between drug and placebo, and the drug is erroneously judged as safe) in reviews of harms are of most concern, as opposed to type I error which is of main focus in efficacy studies to prevent ineffective drugs being prescribed to patients. Type II errors can stem from under-reporting [113], inadequate sample sizes to measure uncommon or rare events [11], limited follow-up duration [7], difficulties in defining unexpected outcomes, exclusion of patients with risk factors for AEs, and lumping AEs into

many subcategories [30]. It is therefore important for reviewers to identify specific study designs that are most likely to yield robust harms data, rather than rely on studies that cannot detect harm, and may lead to a type II error.

Reliable detection and reporting in studies varies with predictability of the adverse effect. Uncommon events, with striking or distinct clinical features are likely to be captured through spontaneous reporting, case reports or case series, either within clinical trials or PV systems. Although spontaneous reports may provide a signal, more detailed information on the magnitude of associated risk of rare events is better sourced from case-control designs. For a quantitative analysis of relative risk the background incidence of the harm outcome, onset (timing) of the AE relative to the drug exposure, and anticipated magnitude of increase in risk with the drug should all be considered carefully [108].

4.1.4 Search strategy

To identify the relevant studies a search strategy should be developed around the research question considering the population involved, intervention being used and the outcome. In general there are two main approaches that have been discussed previously [44, 114] by either searching electronic databases using indexed terms (i.e., Medical subject headings (MeSH)) or by using free-text terms used by authors in title and abstract. Each should be combined to maximize the sensitivity for finding relevant literature.

4.1.5 Data sources

There are a wide range of sources that can be used for exploiting further information and data on harms, including: medicines information sheets (SmPCs

and PILs), pharmaceutical companies, regulatory agencies, academia projects, bibliographic databases, online registries and PV systems. However each has its own distinct limitations that should also be considered carefully, as discussed in Table 13. Data from PV systems and their potential use in observational studies will be discussed in more detail throughout this chapter.

Other review methodology issues that we do not discuss here, like assessment of bias, collecting data, analyzing and presentation and interpreting results have been outlined previously [47].

Table 13: Major sources available for information on harms.

Source	Type of information (Where obtained from?)	Limitations
Medicines information: SmPCs & PILs	Section 4.8 of SmPCs includes listings of suspected adverse effects with incidence of risk recorded (eMC).	Generally will vary across countries. Uncertain whether selections of data and the analysis are systematic, and source data usually unavailable. Restricted to license drug only.
Pharmaceutical companies	CSRs and IPD available under request (e.g., GlaxoSmithKline, Eli Lilly, Johnson & Johnson (J&J)).	Policy details and cut-off date periods can restrict access to certain trials as explained in chapter 3. Open access to IPD that could be used to re-identify patients so free-text fields that could provide important information about harms maybe redacted.
Regulatory agencies	Drug safety updates (MHRA, FDA), Drug Analysis Prints (MHRA), CSRs (EMA), and Spontaneous reports (EMA, MHRA and FDA)	Drug safety updates and drug analysis prints only provide summary information. CSRs encounter the same issues as pharmaceutical companies, although the EMA tend to be more coherent under data requests, also covers all EU centrally licensed drugs.
Academia driven projects	YODA project holds data and documents from Medtronic and J&J.	Reliant upon agreement with pharmaceutical companies for providing data. Plus limited number of trials included.
Bibliographic databases	Published journals with case reports, randomized studies and non-randomized studies (MEDLINE, SCOPUS, ISI Web of Knowledge and DARE etc).	Susceptibility to publication bias and selective outcome reporting with often only favorable outcomes reported. Meaning that harms data is often missing.
Online registry and results database	ClinicalTrials.gov contains information about medical studies in human volunteers. Includes protocol or plan, inclusion/exclusion criteria and summary of results.	Information can be limited with no AE listings, or is often very sparse. Trials are in accordance with the FDA amendments act of 2007, therefore trials finishing before this date are not included.
Pharmacovigilance systems	Spontaneous reporting systems (MHRA, WHO, EMA and FDA), Prescription event monitoring schemes (UK and New Zealand), and Health databases (CPRD, THIN and MEMO).	Spontaneous reports susceptible to bias and often lack detail in report. PEM are expensive to operate and limited to specific drugs. Data from Health databases are often too excessive to even consider.

SmPCs - summary product characteristics; PIL - product information leaflet; IPD - individual participant data; YODA - Yale University open data access; eMC – electronic medicines compendium; CPRD – clinical practice research datalink; THIN – the health improvement network; MEMO – medicine monitoring unit.

4.2 Pharmacovigilance systems

It was not until the disaster caused by thalidomide in 1961 that the first systematic international effects were initiated to address drug safety issues. At that time many thousands of congenitally deformed infants were born as the result of exposure in utero to an unsafe medicine promoted for use by pregnant mothers [115]. After the thalidomide disaster, PV systems were developed in member states for the collection of individual case histories of adverse drug reactions (ADRs).

These systems use spontaneous reporting or other pharmacoepidemiological methods to systematically analyse AEs associated with the use of drugs, identify signals or emerging problems, and communicate how to minimize or prevent harm. These systems have provided evidence in the past that can be used to institute regulatory action to protect public health and avoid further disasters [116]. However, these processes are not always perfect as recently experienced with a case involving the type 2 diabetes drug rosiglitazone with associated risk of myocardial infarction adverse effects. A meta-analysis of 42 trials of rosiglitazone was published in May 2007, showing an increased risk of myocardial infarction and death from cardiovascular causes [117]. However, the spontaneous reporting systems (FDA and EudraVigilance) were found too insensitive to detect increased risks in common events like myocardial infarction in diabetics. This case highlights the need for reviewers to examine different sources of evidence on harms.

The two main types of systems for surveillance are either passive or active in nature (Table 13). Each will now be discussed individually.

4.2.1 Passive systems

Passive surveillance means that no active measures are taken to look for adverse reactions other than the encouragement of health professionals and others to report safety concerns. Reporting is entirely dependent on the initiative and motivation of the potential reporters [118]. This is the most common form of PV, and is often referred to as "spontaneous" or "voluntary" reporting. Currently safety signals are mainly detected from spontaneously reported data, or the publication of case reports in the literature. Spontaneous reporting of clinical concerns by empirical observation of drugs has led to the detection of previously unsuspected side effects [119].

4.2.1.1 Yellow Card Scheme - A Spontaneous Reporting System in UK

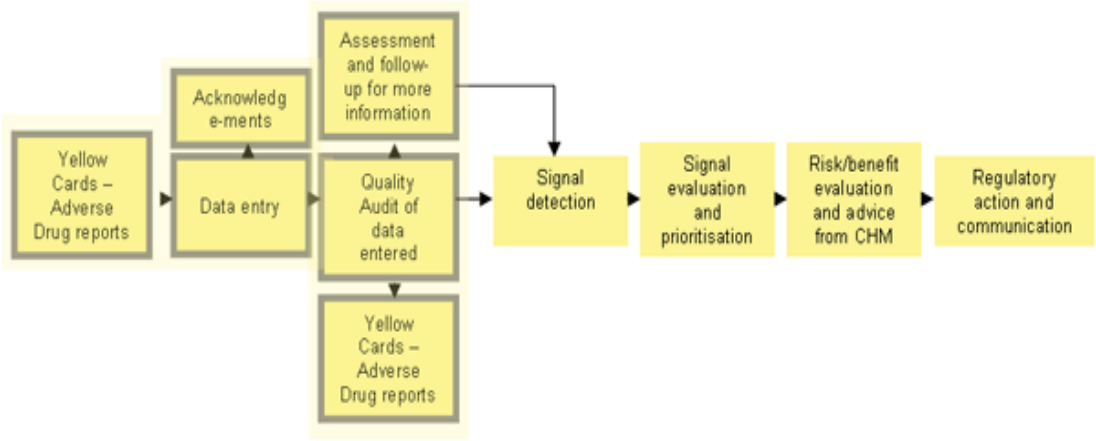
At present, most PV efforts are mainly spearheaded and coordinated by national PV centres, such as the MHRA in the UK. National centres play a major role in regulating matters pertaining to drug safety and efficacy, increasing public awareness and development of clinical practice and public health policy.

The thalidomide disaster in 1961 highlighted the necessity for the licensing and safety monitoring of drugs used in humans [120]. This signified the birth of the – *'Yellow Card Scheme'* as it later became known, because the reply-paid cards used by doctors and dentists to report adverse effects were printed on yellow paper [121]. Reporting increased following the inclusion of a yellow page in GP prescriptions pads reminding GPs to report effects, and again in 1986 following

the inclusion of the yellow card in the British National Formulary (BNF) [122]. In January 1976, the Black Triangle (▼) Scheme was introduced to highlight certain medicines (predominantly newly licensed), for which intensive monitoring was required. Any suspected side effects involving black triangle drugs must be reported.

Yellow Card reports which were originally held within the Adverse Drug Reactions Online Information Tracking (ADROIT) database have now been transferred to the 'Sentinel database'. The sentinel database carries out certain operations after receiving a yellow card form, which usually includes the signal detection, signal prioritization and evaluation, risk/benefit evaluation and regulatory action and communication stages as shown in Figure 9. The signal detection, prioritization and evaluation stages will be discussed later in this thesis.

Figure 9: Yellow card scheme operations (adapted from MHRA website)



Information collected by the yellow card is vital and will help in establishing a suspected ADR and any causal relationship with a drug, as well as allowing

contact tracing to the reporter for clarification or enquiries on further required details if needed.

Data collected by the yellow card includes:

- Suspect drug - route of administration, daily dose and dates of administration
- Suspect reaction - include diagnosis if relevant, whether the reaction was serious and the reason why, any treatment given for the reaction and its outcome
- Patient sex, age at time of reaction, patient's weight and local identification number
- Reporter details
- All drugs currently being taken by the patient and drug history for the last three months prior to reaction
- Any information on drug re-challenge
- Relevant medical history including allergies
- Any other information that the reporter considers relevant.

4.2.1.2 World Health Organisation - Programme on International Drug Monitoring

It was recognised more than 40 years ago that maintaining an international database of ADR case reports, and creating a network of institutions and scientists concerned with drug safety issues provided an enormous benefit. The World Health Organisation (WHO) Programme on International Drug Monitoring was born on such a belief in 1968 and is based at the WHO Collaborating Centre

for International Drug Monitoring, the *'Uppsala Monitoring Centre (UMC)'*, in Sweden [123].

Each participating country has a national centre which communicates directly with the UMC and is responsible for collecting spontaneous reports of ADR suspicions. The UMC transforms the case reports into specific WHO format before it is entered into the *'VigiBase'* (WHO Adverse Drug Report database) [124]. As of April 2015, the database reportedly contains over 10 million case reports with more than 120 countries having joined the programme, with 29 countries being considered as 'associate members'. When warranted, signals are written up and published in medical journals to be reviewed and to initiate necessary actions. The role of the system is to concentrate on the rare (with incidence < 1:1000) but clinically significant reactions.

4.2.1.3 EudraVigilance

A similar scheme to the WHO international drug monitoring program, is the 'European union drug regulating authorities pharmacovigilance (EudraVigilance)' system [125]. EudraVigilance is a central database management system, created on December 2001, and maintained by the EMA. Spontaneous ADR reports received from the EEA health regulatory agencies and pharmaceutical companies are stored in the EudraVigilance post-authorization module, and from May 2004 the EudraVigilance also receives SUSAR reports from clinical trials which are stored separately as part of the EudraVigilance clinical trial module. SUSARs submitted to the MHRA for EU licensed drugs, are also transferred to the EudraVigilance clinical trials module.

The aim of EudraVigilance is to create a common EU reporting procedure for adverse reactions and side-effects for drugs marketed in the EU, and to support the public by making safety information available for scientific assessment. However, at the moment only health regulatory authorities from the EU and pharmaceutical companies have access to this database, although steps are being undertaken to allow public access and academics to certain elements within the database.

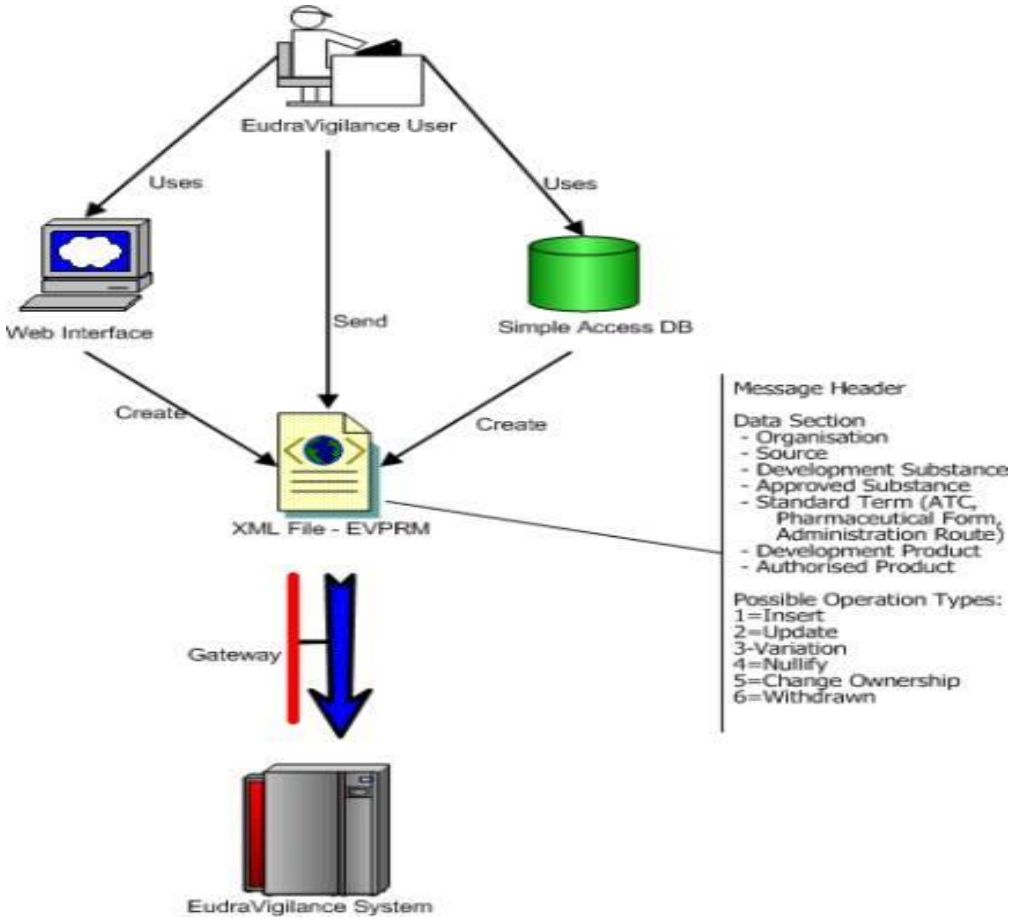
Reports can be submitted via the EudraVigilance Gateway which is an electronic regulatory submission environment. The Gateway allows drug companies, applicants and sponsors of clinical trials and health regulatory authorities to report through a common reporting point, and allows for standardization of data and elimination of data transcription errors. Medicinal product data is owned by the sender organisation that entered the information into EudraVigilance. They can add, remove or alter any information added at any time by accessing the Gateway. Registered national health authorities can view all information on EudraVigilance, but other organisations can only view and make changes to data entered by the organisation itself (Figure 10).

4.2.1.4 Strengths and Weaknesses of Passive systems

Passive systems have clear strengths, with a system that covers all drugs and the whole patient population, including subtypes such as elderly. They are regarded as non-interventional with respect to prescribing habits, and thus include the reporting of events that cannot be readily studied for ethical reasons, such as overdoses or inappropriate co-medication. Passive systems are able to monitor

all drugs in use, and remain the only system that is able to monitor drugs which are not widely used. Additionally they are able to detect a wide spectrum of ADRs (including severe or rare), interactions and other problems (e.g. pharmaceutical defects). They are also effective, rapid, continuous, and comparatively inexpensive.

Figure 10: Eudravigilance data collection process (adapted from EudraVigilance website)



There also exist a number of inherent weaknesses when using passive systems. Often there is only limited information on reports and secondary case evaluation is not always possible. Not all events that occur will be recognised as drug

induced by a healthcare professional, and even those that are suspected will not necessarily be reported to the relevant authority. The “under reporting” effect leads to decreased sensitivity of the system, which may also be vulnerable to selective reporting, e.g. reporting rates for established centres are frequently less than 10% for serious reactions. Linking data between systems is not encouraged, since duplicate reports may appear from multiple systems. The number of reports received may depend on numerous factors; the inherent acute toxicity of the drug, the usage of the drug, how long the drug has been on the market, the year of its introduction and whether there has been any publicity about the drug. The control information is not collected as part of passive systems (i.e. drug use is not known, and thus one has no direct information on incidences or denominators), hypothesis testing studies are usually needed to confirm safety signals, and would therefore cause a delay in the issuance of an appropriate warning. The data can also be expensive to access.

4.2.2 Active systems

Active (or proactive) safety surveillance means that active measures are taken to detect adverse reactions [126]. This is managed by active follow-up after treatment and the events may be detected by asking patients directly or screening patient records. This surveillance is best done prospectively. The most comprehensive method is cohort event monitoring (CEM), commonly referred to as prescription-event monitoring (PEM) and is currently carried out in the UK and New Zealand. Other methods of active monitoring can include the use of

registers, record link-age and screening of laboratory results in medical laboratories which will be discussed later in the chapter.

4.2.2.1 Modified-Prescription-Event Monitoring

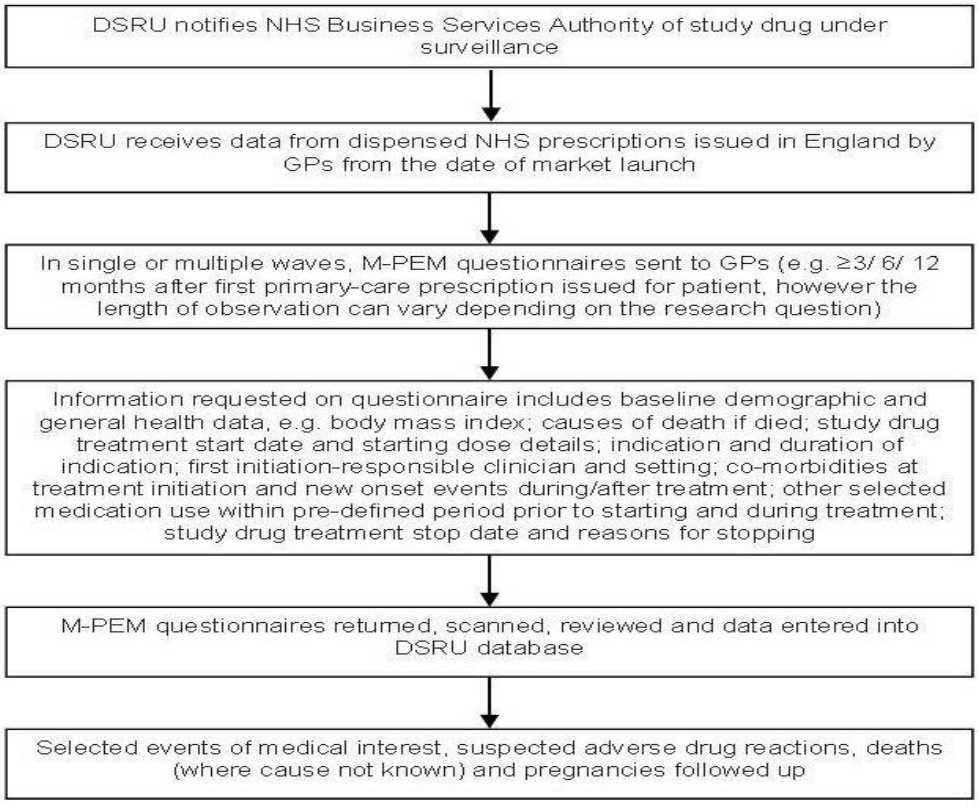
Recognising the importance of monitoring drug use in 'real life' and the theoretical basis for establishing a system to monitor events regardless of relatedness to drug exposure, led Professor W.H.W. Inman to establish the system of PEM at the Drug Safety Research Unit (DSRU) in Southampton in 1981 [127]. The DSRU is an independent registered medical charity but is extensively supported by donations from the pharmaceutical industry whose work is principally concerned with PV associated with newly marketed drugs. DSRU operates outside the MHRA or any government office.

PEM is a non-interventional observational cohort form of PV, and generates signals which through pharmacoepidemiology can be investigated to determine relevant concerns regarding drug safety. It is a hypothesis generating technique with a large database of 900,000 patients, and currently has computerised clinical data on over 100 newly marketed medicines with an average cohort size of over 10,000 patients. From 2011 onwards the DSUR no longer conducts standard PEM studies. In parallel with scientific developments in pharmacoepidemiology and regulatory requirements in PV, the technique has evolved in becoming a more targeted safety study known as Modified PEM (M-PEM) [128].

M-PEM relies upon the collection of NHS prescribing information from individual prescriptions once they have been issued to a patient and dispensed by a

pharmacist (Figure 11). The prescriptions are dispatched by the pharmacist to NHS prescription services for reimbursements. The NHS will then send a copy of the relevant prescriptions from relevant GPs in UK to the DSRU. DSRU then collect details of the prescriptions for the drugs it is monitoring, and records information on the first 20,000 - 30,000 patients prescribed a new drug.

Figure 11: Database structure (adapted and modified from Drug Safety Research Unit Website).



After a suitable interval of 3-12 months, the doctors who prescribed the drug being monitored are sent green form questionnaires on which they are asked to record events reported by the patient subsequent to the prescription. Although the aim is to acquire information on medical events for all the prescriptions, no more than 4 forms are sent to each doctor in any 1 month. A medical event is

defined as any new illness, change in an existing illness or reason for medical consultation (e.g. in relation to pregnancy), regardless of whether it was thought to be related to treatment. All reported events are followed up to determine the outcome and the cause of all deaths are established. Events are then investigated for causal relationship and the incidence density (number of reports/number of patient-months of exposure x 1000) of a particular event/adverse effect is calculated. Signals are generated by an event having unusually high incidence density.

4.2.2.2 Strengths and Weaknesses of M-PEM

M-PEM has clear strengths, including the existence of a large database with the data containing over 900,000 patients (including the whole of England). It also enables large cohorts to be assembled over time and usually experiences a high return rate of the forms with on average more than 55%. The M-PEM system is the only form of post-marketing surveillance which prompts all doctors using new drugs to report the events which follow their use in the UK. It represents the 'real world' use of newly marketed medicines with no patients excluded, and asks about events and not ADRs, which could prevent GPs from returning the forms due to doubts. Therefore there is a possibility to detect side-effects which no doctor(s) has suspected. Numerator and denominators are provided in M-PEM which is collected within a known time frame.

M-PEM also poses some obvious weaknesses. It is only for new drugs intended for long term widespread use within the primary care system, and does not extend into hospital monitoring. Therefore the use of monitored drugs initiated

or stopped in hospitals will not be detected, and the process involved when choosing which drug to monitor is not clear. The delayed onset makes it difficult to detect harms, with some never being detected. Population selection bias may exist in M-PEM, by excluding certain populations such as children and elderly. This may occur when drugs chosen to be studied may not be used by all populations, or doctors may choose to return green forms only for the sub-population consisting of the majority of prescriptions, i.e. returning forms only for adult prescriptions and ignoring the few prescriptions for children. Limitations on the number of green forms for each doctor may also create selection bias and conceal bias from doctors which fail to return the forms. Finally, M-PEM is relatively costly and requires a huge amount of resources.

4.2.3 Health Databases

Health databases or health record-linkage databases in active surveillance are used for drug safety observational studies, primarily cohort or case-control. Record linkage is the systematic combining of records of individuals in a population stored separately, and has made significant contribution to PV by linking drug exposure to outcome data. The primary aim of observational studies using this type of data is hypothesis testing or strengthening, of a known or suspected side effect [129].

4.2.3.1 Clinical Practice Research DataLink

The General Practice Research Database (GPRD) was established in June 1987 as the VAMP Research Databank. Participating GPs received practice computers, and the VAMP medical text-based practice management system in return for

undertaking data quality training and submitting anonymised patient data for research purposes [130]. In November 1993, Reuter's health information acquired VAMP Ltd and one year later Reuters donated the database to the department of health where the database was renamed GPRD. In 1995, Reuters launched Vision, a Windows-based practice management software application used by GPs in the GPRD scheme. In 1999, the Medicines Control Agency - MCA (which became part of the newly created MHRA in April 2003) took over management of the GPRD, and initiated a redevelopment programme to enable broader research usage of the database. The database has since been renamed the Clinical Practice Research DataLink (CPRD) as of 2012 [131].

The CPRD is the world's largest computerised database of anonymised longitudinal medical records from both primary and secondary care settings. Data as of the year 2009 consist of over 20 million active patients from approximately 600 primary care practices throughout the UK (approximately 6% of UK population) providing 46 million patient years of high quality validated data. CPRD is operated on a self-financing not for-profit basis and data are licensed exclusively for medical and health research purposes. It is used to support medical and public health research in the following areas: Clinical research planning; Drug utilization; Studies of treatment patterns; Clinical epidemiology; Drug safety; Health outcomes; Pharmacoeconomics or Health service planning [132].

The participating practices supply CPRD with a wide range of information covering all aspects of patient care, including:

- Demographics, including age and sex, practice location.
- Medical symptoms, signs and diagnoses, including comments and co-morbidity, medical history.
- Therapy (medicines, vaccines, devices) – includes co-prescription, dosage details, off-label prescription, medical procedures, repeat prescriptions.
- Treatment outcomes.
- Events leading to withdrawal of a drug or treatment – includes ADRs (certainty and severity assessments).
- Immunisation details including status, stage, and type, route of administration, reason and batch number.
- Referrals to hospitals or specialists.
- Laboratory tests, pathology results.
- Lifestyle factors (height, weight, BMI, smoking and alcohol consumption).
- Patient registration, practice and consultation details.

4.2.3.2 The Health Improvement Network

The Health Improvement Network (THIN) database represents a collaboration between two companies; In Practice Systems Ltd (INPS) who were responsible for the development of the vision software used by GPs in the UK to manage patient data, and Cegedim Strategic Data Medical Research UK (CSD MR UK) who then provided access to the data for use in medical research [133].

Since THIN data collection began in 2003, over 500 vision practices have joined the scheme. The database is used worldwide by researchers for medical studies in drug safety, epidemiology and health outcomes. The staff responsible in the

development of the CPRD has spent over 20 years facilitating the database, and therefore the data provided is formatted very similarly to CPRD data.

4.2.3.3 Medicine Monitoring Unit

The Medicine Monitoring Unit (MEMO) is a University of Dundee based research collaboration that undertakes research into the safe, effective and cost effective use of medicines and devices as well as helping to improve the understanding of disease, all using anonymised healthcare data [134]. MEMO was originally set up to undertake hypothesis testing PV studies using three original datasets: dispensed prescribing, hospitalization and death certification. These datasets remain the backbone of MEMO research. Currently, MEMO is enhanced by access to other datasets such as laboratory information and primary care data. MEMO only covers the Tayside NHS population (approximately 400,000) based register. Case note validation is possible and undertaken where coding validity or additional information is required. MEMO also works in conjunction with the information services division (ISD) to record link dispensed prescribing to hospitalizations.

4.2.3.4 Strengths and Weaknesses of Healthcare databases

Health databases possess a number of clear strengths. Clinical data is available at individual patient level both in primary and secondary care settings, currently these systems are the largest and most comprehensive source of data of its kind worldwide. The sample population is large enough for PV targeting rare diseases and special populations (e.g. pediatric and the elderly); with a considerable statistical power for cohort and case-control studies, including long-term cohort

studies. The clinical data contained within are regularly updated and validated, and available as recorded by the GP, with event mapping to MedDRA terminology and prescription data mapped to the British national formulary (BNF) classification system.

One of the main drawbacks of health databases is the susceptibility for incomplete information inputted by participating practices. If the quality of data provided is low, data from the practice will not be accepted. The data collected in health databases can also have limited record linkage capability. The high cost of accessing the data, which ranges from £7,000 to £60,000 to cover a single research study, can be too excessive for most research groups particularly public sector, although limited grants for access are sometimes available [135, 136].

4.3 Observational studies in practice

Observational studies have become increasingly accepted for use in PV. They offer a 'real world' surveillance of drug use and its complications. Observational studies can usually be divided into three main studies in regards to PV: 'pharmacoepidemiologic studies', 'registries' and 'surveys'.

4.3.1 Pharmacoepidemiologic studies

Pharmacoepidemiologic studies encompass various study designs including cohort (retrospective or prospective), case-control, observational studies and others [109]. They may use a wide variety of data sources including prospective 'real world' data (e.g., hospitalized data, clinical trial data and health databases), and are designed to test a pre-specified hypothesis. Outcomes include estimation of relative risk associated with a drug, and may even provide

estimates of risk (incidence rate) for cohort studies. Although observational studies in general are not placed highly in the hierarchy of evidence due to bias and their reduced ability to address confounding factors, they remain the only practical choice to study uncommon or delayed adverse effects. Observational studies are also gaining acceptance for use in hypothesis testing especially when more than one study is used to test the same hypothesis, therefore strengthening the result outcomes.

4.3.2 Registries

A registry according to the US FDA is - an organized system for the collection, storage, retrieval, analysis and dissemination of information on individual persons exposed to a specific medical intervention, who have either a particular disease, a condition that predisposes to the occurrence of a health related event or prior exposure to substances or circumstances known or suspected to cause adverse health effects. The creation and analysis of registries is particularly useful for examining outcome information not available in large automated databases from multiple sources. The collection of spontaneous case reports either reported or published detailing specific adverse effects are among the common application of registries and is commonly used to complement signal detection by national PV centers.

4.3.3 Surveys

Surveys such as questionnaire studies are being increasingly used as a tool in PV. Surveys are frequently used to gather and assess information on various issues such as:

1. Evaluating a safety signal

2. Evaluating knowledge about adverse reactions, AEs and various other knowledge and attitudes of/towards PV among health practitioners and the public [137, 138].

3. Assess the use of products/drugs in regards to safety, efficacy, quality and adherence to guidelines.

4. Gathering information or data regarding a specific area of interest.

Surveys are subject to a number of biases and confounding factors, with low participation being their main weakness. Various methods are used to encourage participation including payment or providing certain benefits for respondents; however this practice in itself may lead to bias and could be ethically challenging. Surveys are best validated or piloted before implementation to give credence and an idea of what to expect, as well as to identify any shortcomings that may need to be addressed. A well planned and piloted survey will often yield high-quality results.

4.4 Discussion

In this chapter we have summarized a structured framework approach proposed by the Cochrane AEMG [47], for conducting systematic reviews that include harms. In this framework, the starting point for structuring the review is determined entirely by the scope of research question: a broad overview of safety problems associated with the drug ('hypothesis generating'), or to evaluate the magnitude of risk and clarify the characteristics of the adverse effect ('hypothesis testing/strengthening'). Each of these approaches requires

careful consideration to determine which study designs and data sources to include in the systematic review.

Systematic reviews of RCTs to assess harms are usually most common. However RCTs are usually insufficiently powered, or too brief, to detect rare but serious adverse effects or modest but important increases in the risk of common disease outcomes that can have a major population impact in absolute terms. Most RCTs also tend to exclude the elderly, patients with co-morbidity or pregnancy, and this reduces the generalizability of these data. Therefore, at the time of product launch, there are often limited harms data of any new drug, in both the short- and longer-term which is directly applicable to that of the target population. Drugs in use therefore need to remain under constant surveillance (Post-marketing) and studied by observation in PV systems to identify safety signals and thus serve to generate hypothesis. PV systems however possess many strengths and weaknesses as summarized in Table 14.

Spontaneous reporting is the principle PV system in use worldwide with proven effectiveness and a good track record resulting in the avoidance of many potential disasters and the identification of new or previously unknown drug related adverse effects. They encompass the main advantages including a wide population, relatively low costs and resource utilization, and well established methodology. However spontaneous reporting systems depend on voluntary reporting of health care professionals, hence the reporting rate or under-reporting rate becomes the limiting factor.

Table 14: Key strengths and limitations for each PV system.

Type	Passive systems	Active systems	Heath databases
Key Strengths	Proven with good track record Wide population coverage National PV reporting mechanism for most countries. Relatively easy to implement Low resource utilization and cost Covers all population, drugs and health settings	Does not depend upon voluntary reporting-able to capture high incidence of harms Able to shorten lag time from marketing of drug to detection of new ADR Provide a numerator and denominator Proven with good track record	Contains huge amounts of health data with wide population coverage-considerable statistical power Does not depend on voluntary reporting – able to capture high incidence of ADRs Data are frequently update and validated Maybe able to follow through to secondary care if data linkage is available Low resource and cost requirement once setup
Limitations	Depends on voluntary reporting Reporting rate is very low even in developed countries Does not provide a denominator Long lag time between marketing of drug to detection of new ADR	High resource and cost requirement Implementation limited to selected drugs only Population selection bias and conceal bias of doctors reporting may occur Validation mechanism is unclear/difficult	Utilization is relatively new and not proven Implementation is not possible if national large health databases is not available Incomplete information input by data managers Does not cover population/drugs/setting where information is not collected High cost to access data

To help overcome the limitations of spontaneous reporting, active systems have been introduced. Among the main active system in the UK is the M-PEM, which requires participation by healthcare professionals, although participation is encouraged by providing payments to reporters. Therein lies the main weakness

of M-PEM, the high cost and resource required for implementation. Monitoring a single drug using M-PEM requires tens of thousands of green forms to be sent out to GPs, and the cost can therefore be considerable. Therefore, M-PEM has to carefully select drugs which it intends to monitor or investigate, thus limiting its usefulness as a fully-fledged PV system.

The advent of large anonymized health databases brings forth other possible systems. The arrival of more comprehensive patient health databases containing individual demographic data, health records, prescription records and even laboratory results and other associated health information have increased the usefulness of these data sets for PV purposes. The use of health databases is not only confined to just hypothesis generation, but also hypothesis testing. With promise of huge population coverage, complete prescribing and health event records from UK primary care practices and hospitals, quick access to information, the elimination of voluntary reporting by health professionals and low cost. It is easy to see why many are excited and hopeful for the use of health databases for PV, with examples in the past of studies using data from the CPRD/GPRD to inform regulatory decisions [139]. In many cases, such studies have provided reassurance about the safety of medicines, though studies are also used to triage safety signals identified through spontaneous reporting schemes by providing ready background incidence rates of diseases and drug exposure (denominator) data. Health databases are still in its infancy with many deficiencies including validation of data, linkage between databases, cost of public accessing the data, and the privacy of data which needs to be ironed out.

In the past decade, development in the field of PV has progressed tremendously with many governments highlighting it as a priority area. There is recognition that early identification of unknown serious adverse reactions for all drugs is impossible, adverse reactions cause high morbidity and mortality, represent a burden to the national cost of healthcare and continuous monitoring for adverse effects for all drugs is essential. A recent study [140] was conducted to determine the nature of evidence used to support the withdrawal of marketing authorization of drug products for safety reasons throughout the EU between 2002 and 2011. The study reports that the level of evidence used to support drug withdrawal has improved during the past 10 years, with an increased use of case-control studies, cohort studies, RCTs and meta-analysis. The research demonstrates that such studies have contributed to decision-making in almost two-thirds of cases. Previously, only one-third of decisions used evidence from observational studies or clinical trials [141].

There is also recognition of the many limitations of current PV monitoring systems that must be improved. Among the main issues for current PV systems are; 1) Increase coverage of population (including special populations such as children and the elderly) and drugs monitored, 2) Reduced cost and resource requirements, 3) Increased participation from health professionals, 4) Reduction in the lag time between drug launch, detection of adverse effects and the issue of appropriate warnings or appropriate regulatory actions, and 5) overlap between databases. In Europe, the Eudravigilance system consists of one common electronic reporting point within the EU that is advanced. This

harmonized system is compliant with ICH E2 standards. The advantage of this system is the ease of use and fast reporting (pre and post-authorization) mechanisms both from reporters, but also between health authorities. Unfortunately, despite significant globalization of pharmaceutical companies and many of the same drugs being available in the main territories, harms data including SUSARs are not shared routinely between territories. Further efforts are needed to improve access to such systems like the Eudravigilance.

Pharmacovigilance will not function without a good monitoring system or available data sources, and will lose its effectiveness with long lag times and will not be feasible if the cost and resources required are too high. Consequently, in spite of more than 50 years of PV, efforts are still currently in place to improve upon existing systems and to develop new systems. Weaknesses in PV systems are being addressed with encouragement from national monitoring bodies. Developments to address a deficiency in a PV system frequently generate further new issues. Creating the perfect PV system may not be possible, however new systems must continue to be developed and improvements upon the current ones in place to reduce the recognized limitations and deficiencies when detecting harms in the future.

Chapter 5: A Survey of current practices in Clinical Trial Units

In chapter 4 we identified spontaneous reporting systems, M-PEM and health databases as potential resources for accessing existing harms data. However, although these sources could provide valuable additional information from new RCTs and systematic reviews, the limitations that were discussed will most likely prevent their use in practice and limit their utility. Though, little is known about their use in practice and so this chapter describes a survey to investigate whether and how UK clinical trial units (CTUs) conduct harms related safety monitoring and to understand the value of the different resources available for exploiting harms external to the trial. In addition, the results in section 5.4.1.2 will be used to inform on the design of the simulation study in proceeding chapters.

5.1 Introduction

In recent years pharmacovigilance in the public sector has become an essential part of clinical trial conduct, especially across EU member states following the implementation of the EU Clinical Trials Directive [20] and its transposition into UK law by The Medicines for Human Use (Clinical Trials) Regulations 2004 [21]. The responsibilities for PV have also been laid out previously within the ICH-GCP E6 [22]. The resulting outcome of these documents now means that sponsors and clinical investigators of any clinical trial have a responsibility to adhere to these regulations and report any safety concerns where necessary. These

responsibilities can often only be fulfilled by creating a robust reporting system backed up with clear oversight of the processes involved.

The reporting system will support the preparation and submission of annual safety reports in the form of a development safety update report to regulatory authorities or research ethics committees (RECs), and facilitate direct reporting of SUSARs to the regulatory agencies [142, 143]. Oversight of the reporting processes involved are usually translated into standard operating procedures (SOPs) to break down each of the component parts individually and provide a road map of the procedures that should be followed [119].

Assessments of any harms during the trial can be evaluated as detailed in the SOP, usually by referring to the trial protocol, safety reference documents (SmPCs and IBs) or trial specific procedure for unblinding if required by the data monitoring committee [144]. Then, if significant ethical or safety concerns arise, or there is unequivocal statistical evidence of benefit prior to the completion of the study, decisions for discontinuation of the study can be made [145]. However, these decisions are rarely straightforward, and there is often a different threshold for stopping a trial in the case of potential harm than in the case of benefit [146]. More comprehensive evaluations of harms are often needed, which may require exploiting other sources for harms as discussed in chapter 4.

5.2 A Survey of Clinical Trial Units

A national survey was carried out to gain further insights into some of the practices involved within UK clinical research collaboration (UKCRC) registered CTUs [147]. The specific aims of the survey were to:

1. Investigate the advantages of using existing harms data that are data-based centrally within the CTUs.
2. Investigate the potential use of existing harms data across CTUs, and identify relevant sources external to the trial (as explored in chapter 4) which could be used to inform trial conduct.
3. Explore the methods being used to mine harms data collected centrally across trials so that safety signals can be detected more efficiently.

5.2.1 UKCRC registered CTUs

The UKCRC registered CTUs are specialist units which have been set up with a specific remit to design, conduct, analyse and publish clinical trials and other well-designed studies. They also have the capability to provide specialist expert statistical and other methodological advice and coordination to undertake successful clinical trials. In addition, most CTUs will have expertise in the coordination of trials involving investigational medicinal products (IMPs) which must be conducted in compliance with the UK Regulations governing the conduct of clinical trials resulting from the EU Directive for Clinical Trials [20].

The UKCRC consists of a network of 45 registered CTUs which have provided evidence to an international panel of experts of their capability to centrally coordinate multi-centre clinical trials (i.e. having overall responsibility for the

design, development, recruitment, reporting, data management, publicity and analysis of a portfolio of trials), and of robust systems to ensure conduct and delivery of clinical trials to the highest quality standards. Oversight and management of pharmacovigilance is of high importance for the CTUs also.

5.3 Methods

The survey questionnaire was developed and transcribed to the online data capture tool SurveyMonkey for completion during the period July 2014 to September 2014. A copy of the survey is provided in Appendix C. Pilot testing of the survey was performed and the survey was revised where necessary.

5.3.1 Population and Sampling

The survey was announced via email inviting CTU directors, co-directors and/or experience trial statisticians to participate. At least two members from each CTU were chosen, and members of whom we already had contacts for were included. A link included in the email provided individual access to the survey, so each participant could respond only once and reminders could be sent. The link also allowed the participants to forward the email on to other CTU members, where deemed necessary. A final reminder was sent out by email with an electronic copy attachment of the survey. The survey was stopped on the 29th September 2014 after approximately three months of the survey being active.

5.3.2 Structure of the questions

The survey consisted of 11 short questions (Appendix C) covering the three aims of interest as detailed in section 5.2. The survey was anticipated to take no longer than 5-10 minutes to complete. Question types included multiple choices,

free text and comments. ‘Other (specify)’ responses were offered to capture a full range of possible answers.

The first aim was to investigate how existing harms (including AEs and SAEs) are *data-based* within the UKCRC CTUs; options included, by ‘single trials’ individually, or by ‘multiple trials’ stored centrally. Harms from multiple trials can either be data-based by a range of diseases, conditions or treatments, or alternatively by a ‘diverse’ range of diseases, conditions or treatments. Of particular interest, was to determine some of the inherent advantages and disadvantages for using a central database to store harms. For those CTUs data-basing harms individually by singular trials, participants were asked to give opinions on the potential for developing a central system in the future.

Secondly we aim to identify some of the commonly used external data sources for exploiting further harms data, and discuss the potential value of their use. An array of potential data sources discussed in chapter 4 were listed as options; e.g., use of own central database, published reports and systematic reviews, health databases (CPRD/GPRD [131], THIN [133] MEMO [134]), and yellow card data from the MHRA [121]. However participants were encouraged to detail on other sources of data used.

Finally, for CTUs who data-base harms in a central reporting system questions were asked about statistical methodologies that were being used to analyze and detect safety signals. For the purposes of this survey and ongoing chapters in this thesis, participants were encouraged to provide as much insight as possible on their methodologies used, to analyze centrally stored harms data.

5.3.3 Data Analysis

Owing to the nature of the study and data collected, descriptive statistics were used to analyze quantitative responses including number(s), frequencies, percentages, with some results displayed via graphical representations where appropriate. Research Ethics Committee (Internal Review Board) approval was not required for this survey, as it did not relate to personal medical information, did not involve patients or healthcare professionals (other than in their roles held within the CTU) and participation was entirely voluntary.

5.4 Results

The survey was active over the period 15th July 2014 to 29th September 2014, and was distributed five times. The mailing list was refreshed on 28th August 2014 adding in new contacts for CTUs that were non-responsive.

A response was received from 22 (49%) UKCRC registered CTUs. Five (23%) of the survey responses were from the directors of the CTU, and remaining responses were from senior trial statisticians. The survey responders had at least five years experience working in clinical trial research, and some had up to 30 years. Multiple responses were obtained from two different members of two CTUs; these results were combined together as one response.

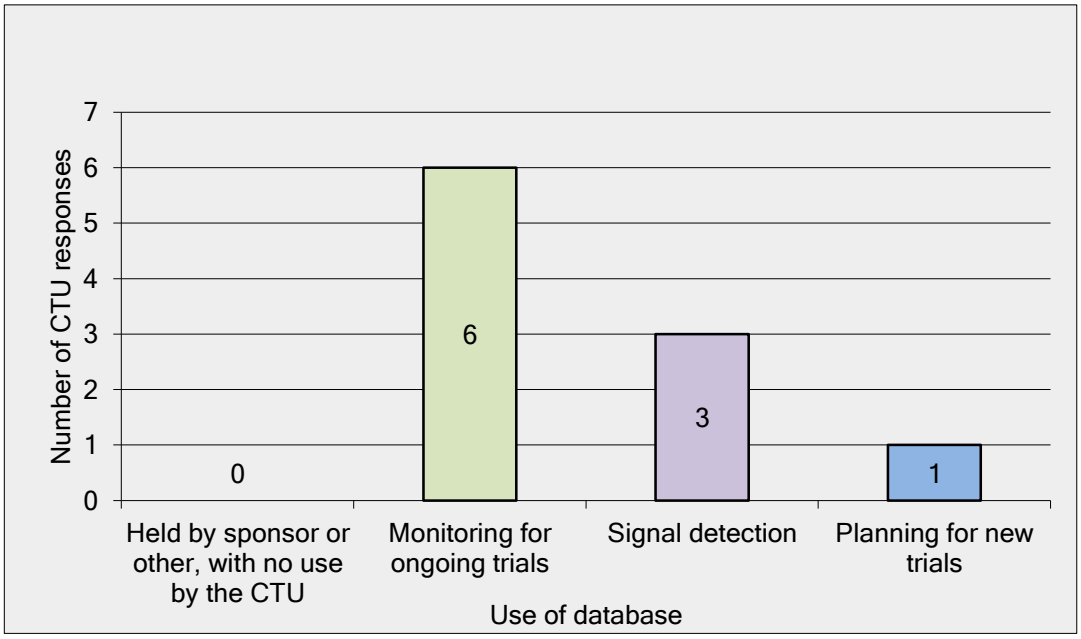
5.4.1 Collecting harms data in CTUs?

Of the 22 responding CTUs, 16 (73%) currently collect harms data in separate individual trial specific databases. Six (27%) CTUs currently collected and stored harms data using a central database including data from multiple trials including a diverse range of diseases, conditions, or treatments.

5.4.1.1 Functionality of the central database

Figure 12 displays how existing harms data is used in the six CTUs with central databases.

Figure 12: The operations in central databases within the CTU.



On-site trial monitoring appeared to be the most common purpose of use, as indicated by all CTUs with a central harms database. Three (50%) CTUs performed signal detection, and one (17%) used the database for the planning of new trials.

Table 15, details the responses from four CTUs, discussing the potential advantages for having a central database. In response 1 the CTU covers a diverse range of trials for different conditions including cancer, cardiovascular disease, stroke, obesity and diabetes. In responses 2 and 3 the CTUs predominantly conduct phase II and III cancer research, and in response 4 the CTU conducts surgical trials.

5.4.1.2 Size of Central Database

One central database contained 12 trials (for SAEs only) with over 100 individual SAEs terms. Two central databases contained 20 trials with a few hundred AEs terms in one and the other failed to provide an estimate. A further two CTUs contained 40 and 42 trials with one reporting approximately 200 AEs terms, and the other with 33 SAEs terms (AEs were not contained in the database) respectively. The remaining CTU contained 34 trials with approximately 140 AEs. The results from this section will be used to inform on the parameters in a simulation study later in this thesis (section 7.3).

Table 15: Advantages and disadvantages of central harms databases as quoted from four different CTUs.

Response	Advantages	Disadvantages
1	"Same generic data collection methods and expertise centrally in process and review for internal reporting and forward reporting". "Easy reporting for DSURS and PSURS".	None
2	"Better cover for trials, better tracking of events including rare AEs". "Coverage of whole patient population".	None
3	"Easy to compare workload for PV for incoming SAEs". "Are considerably effective and inexpensive to maintain".	"Difficult to archive specific trial SAEs".
4	"Adverse events stored in the same way for all trials". "Potential to determine drug-drug interactions and/or drug related syndromes easier".	None

5.4.1.3 Requirement for a Central Database?

Finally sixteen CTUs gave opinions on whether they would ever consider implementing a central database to store harms data from multiple trials. Two (13%) CTUs could see no benefit of a central database, three (19%) said they have further plans to develop and implement such a system, and 11 (69%) were not aware of the possibility of considering a central database.

Further comments were provided by nine CTUs. Four of these comments came from respondents who were 'not aware of considering a central database':

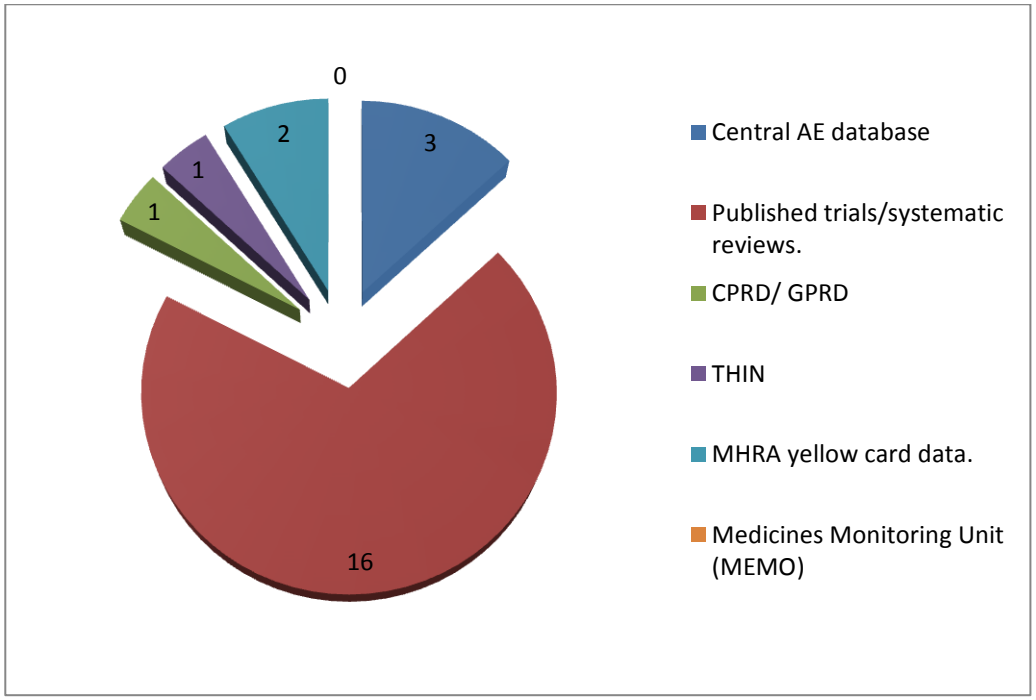
- *"Most of the trials in our unit conducted are non-CTIMP, and AEs are unlikely and less of a concern than in trials of CTIMPs".*
- *"We store SAEs on a central database which is split into trial specific sections. However we only use these separately and not for combined analysis".*
- *"The collection of AEs and SAEs is standardised and collected in a consistent manner across all trials".*
- *"Never considered this, as far as I'm aware. But we are a general trials unit: our portfolio is approximately 15 trials in 12 different conditions, and the interventions are mostly low risk (e.g., behavioral interventions)".*

One responder who indicated that they have considered a central database but could see no benefit, commented with the following: *"AEs were too unwieldy and difficult to archive"*. The remaining four comments were uninformative and therefore not listed.

5.4.2 External Sources for existing Harms

In this section we present the results from the survey discussing the advantages of using external sources of harms during the trial's safety monitoring. Figure 13 provides a breakdown of the sources currently in use, as detailed by 18 (82%) CTUs. The four (18%) remaining responding CTUs failed to provide a response.

Figure 13: Sources for external harms data used in the safety monitoring of trials.



The majority of CTUs (16/18 (89%) responding CTUs) use published trial reports and studies including systematic reviews as their main external source of data about existing harms. Three (17%) use their own CTU central database, and two (11%) have used 'yellow card data' from the MHRA. Observational data from health databases was used by two (11%), including data from the CPRD/GPRD and THIN. The information services division (ISD) based in Scotland, the health and social care information centre (HSCIC), MHRA safety updates/drug alerts,

and downloads for summary products characteristics (SmPCs) from the electronic medicines compendium (eMC) include other sources of data used.

Detailed comments were provided by 10 CTUs, discussing the value and potential limitations of using these external sources as listed in Table 16. Most of these CTUs had used published trials and systematic reviews; though others had used THIN data (1), MHRA yellow card data (1) and an array of other relevant data (1).

Table 16: The responses on value of using existing harms from external sources.

Response	Comments as quoted from responders	Source of data used
1	"The use of using routinely collected data to validate SAEs has been invaluable and will continue to be a valuable resource in clinical trials".	Central AE database and published trials/systematic reviews
2	"Trials do not operate in a vacuum, nor should they. It is important to take note of signals elsewhere, since most trials are too small to detect harm".	Published trials/systematic reviews
3	"It is required to assess ongoing safety e.g., see FDA guidance".	Array of relevant data not limited external sources listed
4	"The use of external data to inform stopping rules".	Published trials/systematic reviews and THIN data
5	"I used a cohort study to help inform a decision on an IDSMC for an external CTIMP trial; whilst acknowledging limitations. We have SOPs on adverse event reporting but they do not mention use of external data sources, and I'm unaware in the small number of CTIMP trials we support use external data sources to inform safety monitoring. This would be something agreed between the trial team and DMC so the unit may not be privy to such arrangements".	Published trials/systematic reviews
6	"Ensure information is current and, receive updates of safety information".	Published trials/systematic reviews
7	"Published results of similar agents are often presented to data monitoring committees as supporting information".	Published trials/systematic reviews
8	"Helpful for preparation of the DSUR".	Published trials/systematic reviews
9	"Review these data for IDSMCs".	Published trials/systematic reviews and MHRA yellow card data
10	"To date, we have based our safety reports solely on emerging literature. I can see the value of cross-linking safety data, but given the general nature of our trials it's less applicable to us".	Published trials/systematic reviews

5.4.3 Methods to Detect Safety Signals

When analyzing harms data, the reporting odds ratio (ROR) statistical signal detection algorithm is used by one CTU which collected their data from multiple trials in a central database. A number of the other CTUs did have considerations towards the use of statistical signal detection methods, as outlined below:

- *"Our studies are largely late phase; also the deployment of signal detection methods can involve a number of issues particularly with multiple-testing".*
- *"We use more orthodox alpha spending approaches (sequential methods) based upon safety and benefits".*
- *"In terms of monitoring safety, it depends on the trial specific data monitoring committee (DMC). If the DMC request that formalized tests be used to compare e.g., SAEs between treatment arms, then this will be incorporated into a safety report (the frequency of which is also trial dependent). However, multiple testing needs to be considered here".*

The use of statistical signal detection methods in clinical trials will be explored later in this thesis.

5.5 Discussion

The data from this comprehensive survey highlights that few UKCRC CTUs currently data-base their existing harms from multiple trials centrally, and is more common for harms to be stored separately by specific trials. Many of the CTUs indicated that they were not aware of considering the need for a central

database, though some have considered the implementation of one in the future. Those with a central database contained harms data from cancer trials or surgical trials; however one did contain trials across a diverse range of conditions. The databases were used predominantly for monitoring ongoing trials, although there was indication that they can be useful for a number of other purposes like signal detection and planning of new trials. They also enable a better coverage of trials and tracking of AEs, and comparing workloads for PV of incoming SAEs is made easier, since all events are stored in the same way for all trials.

Our results also highlight the value for using existing harms obtained externally from the trial. Published trials and systematic reviews were most commonly used, though a number of CTUs have also conducted research using observational data from health databases, like the CPRD/GPRD and THIN. Other freely accessible data sources like the ISD, HSCIC and eMC SmPCs updates were often used. One participant suggested that it is not a compulsory requirement as stated in the SOP to use such data, although many respondents emphasize the value of external harms as being an important part in the decision making for DMCs. It was unclear whether signal detection methods could be used in central databases within CTUs, and multiple testing appeared to be a common concern. Further research is needed to explore the potential of these methods.

A recent study [142] of one UKCRC registered CTU has discussed some of their own challenges experienced when implementing a central PV database. They encountered a number of complexities which included the re-training of staff

members to manage and maintain the database, and there was a requirement for new processes to be translated into SOPs once they have been agreed by all stakeholders involved. Despite these limitations, the central database provided a number of improvements to the data management of the trials, with accurate generations of line listings which were used for the production of reports required for the sponsor or management oversight, and easier reviewing for DMC or submission of annual safety reports to regulatory authorities or RECs.

Due to the general lack of information available on the current safety monitoring practices involved in UKCRC registered CTUs, this national survey therefore aims to provide some valuable insight into the management, use and analysis of existing harms data. A moderate response rate of 49% was achieved over a short period of time. The responses were from directors and statisticians with many years experience working within clinical trial research. Some responses from CTUs consisted of a number of members working within multi-disciplinary teams, which enabled a wider diverse range of opinions from specialist across the CTU.

The voluntary nature of the survey meant that some questions within the survey provided few or no comments, with many participants opting not to elaborate in further detail. This was particularity the case for the open-ended questions determining the advantages or disadvantages for a central database, the opinions of using the database and the values of using external harms. For example, it was clear that some CTUs did use pharmacoepidemiology and systematic review data, though it was unclear why and how the data was used.

This was a limitation in the way in which the question may have been worded, and perhaps an alternative format for this question may have requested that responder's detail why they used the data via a multiple choice's option. This would also encourage further expansions. Alternatively a more appropriate way to determine more accurate information on the use of these data would be to follow-up with interviews.

We restricted the survey to an active period of approximately three months, meaning that we were not able to obtain responses from the 23 remaining CTUs. Therefore there is huge potential for obtaining much more valuable information that would add to the outcome of this survey. Also as part of our strategy for distributing the surveys we did not include the option of mailing hard copies by post, although we did send an attachment copy to the participants directly.

The survey has shown that most CTUs currently data-base existing harms from trials individually, and very few have considered the need to implement a centralized system to monitor harms. For some CTUs they may only collect few AEs reports in a systematic and detailed fashion which is qualitatively different from spontaneous reporting. Hence this may limit the full value and demand for a central PV database. The use of existing harms from external sources is common amongst researchers working in CTUs. These data sources often provide more valuable insight of the adverse effect, and contribute to facilitating the DMCs for the ongoing review of trials and preparation of safety documents required for regulators like the DSUR.

To support the trial safety monitoring of EU medicines studied in clinical trials prior to authorisation, the EudraVigilance Clinical Trials Module (EVCTM) [148] from 2004 began collecting SUSAR reports. Work sharing with the EVCTM can be implemented through their Gateway system, for regular safety monitoring of ongoing clinical trials or when making evaluations of DSURs through aggregate reports. However access to the EVCTM by healthcare professionals, research organizations and the general public is currently restricted, meaning that SUSARs reports cannot be accessed or shared amongst UKCRC CTUs. Though, regulators and sponsors have full access to this data.

Restricted access to this kind of data could be a major impediment for CTUs, who already have limited resources. Therefore, it may be more effective for the CTUs to consider developing their own specific centralized database for collecting AE reports across the wider CTU network. This would allow for easier work sharing capabilities amongst the CTUs so that they can learn from each other, but also support during the trial when reviewing (with published literature) and triaging SAEs to help identify any SUSARs. In addition, such a system could supply advice before the trial with protocol design and study specific reporting requirements.

However, cost consideration is always a high priority in the public sector, and therefore training staff for oversight and management of the system will play a major role. Also the active time for translating the processes into SOPs, which would have to be consistent across all CTUs, is another limiting factor for developing a central PV database. Collecting AE data across CTUs involving a

diverse range of different trials and diseases will often result in increased heterogeneity. Some trials will report very few AEs, but others like cancer trials will likely report high numbers of AEs. Therefore developing a central system based on a specific disease area might be more advantageous to CTUs.

Finally, our results indicated that only one CTU with a central database was in use of a disproportionality signal detection method. However the potential advantages for using these methods across the wider CTU network is unclear. For example if harms data were stored centrally across CTUs or by specific clinical areas (i.e., Cancer trials) hence, resulting in a larger volume of data, then these methods would be of use to researchers. Further research is needed to fully explore the potential of these tools when analysing harms data in clinical trial settings, which will be discussed in the next chapter.

Chapter 6: Tools for Enhanced Signal Detection Analysis

In chapter 5 a survey was carried out to explore the current practices of safety monitoring in CTUs. One of the aims of the survey was to understand what methods of analysis could potentially be used to detect safety signals within and across CTU databases. Therefore in this chapter the focus will be to review the current methods used to systematically explore safety data in PV systems, but also extending to databases of a smaller scale similar to CTUs. Some of these methods will then be used in chapter 7.

6.1 Introduction

The detection and evaluation of signals is crucial for understanding the safety of medicines and for preventing harm in patients. Not only is it necessary to detect new signals, but the principles and practice of PV apply to the surveillance of a wide range of medicinal products [25].

The concept of a drug "safety signal" has been the cornerstone of PV activities for about forty years. However, as more medicines are authorized for marketing each year, and as increasing numbers of persons are taking medicines, this has resulted in an increase in the number of AEs reported to manufacturers and to regulators [123]. Manual reviewing of paper-based reports which provided the foundation of early productive PV systems is simply no longer practical. Modern PV systems, which receive several hundred thousand reports each year, and which have databases containing several million AE reports, are now required to

detect, prioritise, and evaluate safety signals in an efficient and proactive manner. This often requires a systematic approach that couples statistical and analytic methods with sound clinical judgment.

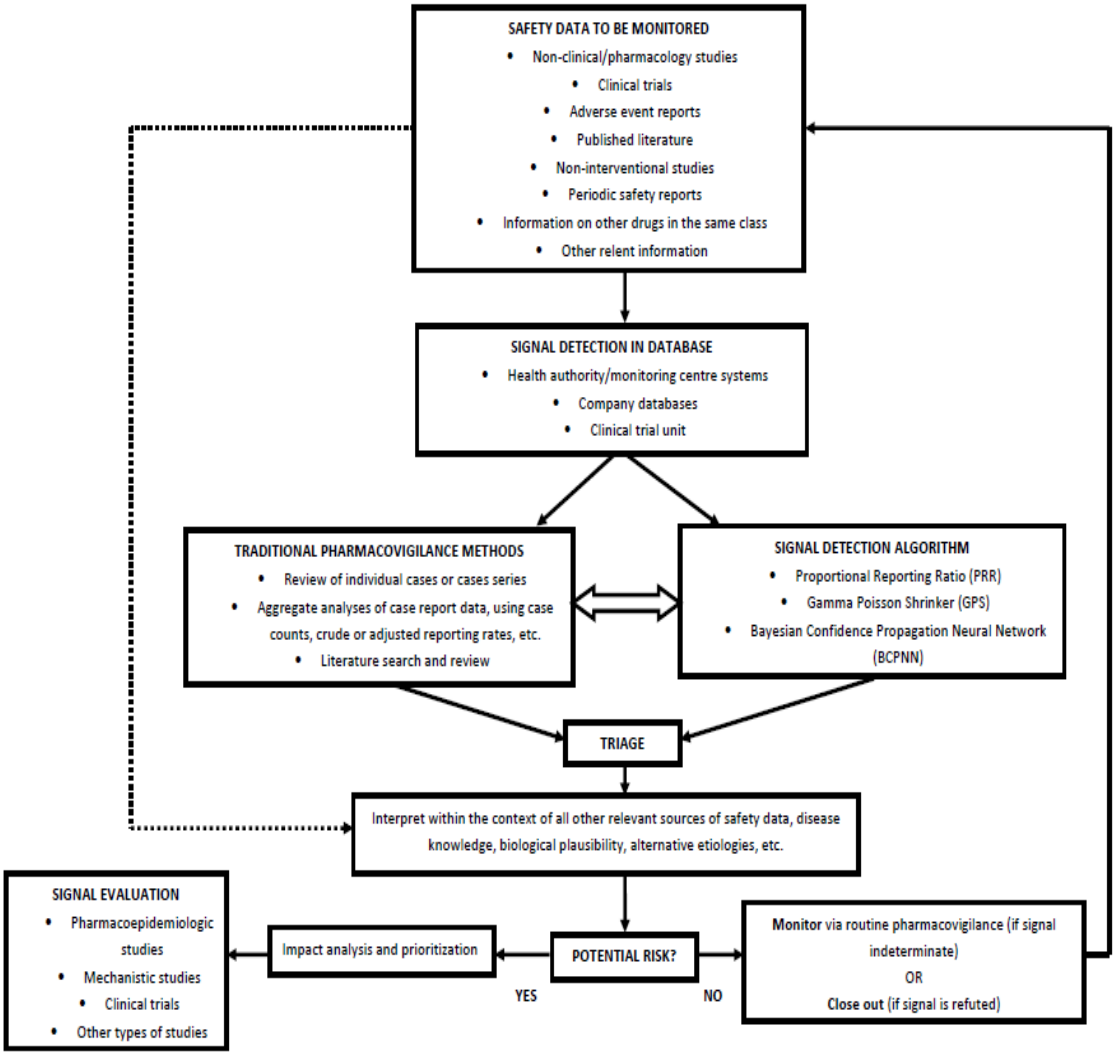
To date in the field of PV, these systematic approaches have been applied most widely to post-marketing approved signal management using passive surveillance systems of spontaneous ADR reports. Though some attempts have been made to extend from the current methods used on data from interventional clinical trials, to developing certain statistical techniques in databases holding clinical trials data [149].

6.2 A Signal Management Framework

Most companies with a central PV system define a signal management framework as the set of activities performed to determine whether there are new risks associated with an active substance or a medicinal product, or whether known risks have changed [150].

A typical signal management framework consists of a flow of sequential steps of signal detection, prioritization, and evaluation (Figure 14) as well as its linkage to risk management activities. PV and drug safety departments at drug companies may be organized differently, but many follow their adaptations of this framework explicitly or implicitly. This chapter will discuss each of these steps individually; however the primary focus will be on signal detection.

Figure 14: A typical signal management framework, adapted from the CIOMS working group VIII



6.2.1 Primary sources of safety evidence

The sources of data for identifying new safety signals can be diverse, but are often detected from monitoring *'individual case safety reports (ICSRs)'*. The ICSR is a health level seven standard (i.e., a set of international standards for transfer of clinical data between software applications used by various healthcare providers) for the capture of the information required to support the reporting of an AE, product problems or consumer complaints of the product.

Accumulation of ICSRs can occur from multiple places during the post-approval phase according to the International conference on harmonization (ICH) E2D guideline 2003, as detailed in Table 17 [151].

Table 17: Sources for the accumulation of ICSRs during the post-approval phase.

Sources of ICSR	Description of sources
I. Unsolicited sources	Spontaneous reporting; literature; internet; other sources.
II. Solicited Sources	Any organized collection of data (outcomes research, clinical trials, registries, surveys, billing databases etc).
III. Contractual agreements	Inter-company exchange of safety data.
IV. Regulatory authority	Any ICSR originating from the regulatory authority submitted to a company, e.g. Suspected Unexpected Serious Adverse Reactions (SUSARs).

6.3 Signal Detection

In recent years, statistical methods for systematically sifting through large amounts of reported AE data have been developed, mainly due to an increase in the volume of spontaneous reports. These tools and methods have collectively been termed *"signal detection algorithms (SDAs)"*. When considering the introduction of these new analytical approaches, an organization should place them, along with other existing traditional PV approaches and statistical tests, in an integrated framework of a signal detection program.

6.3.1 Traditional Signal Detection Methods

Traditional PV methods for identifying new signals and exploring safety issues to generate hypotheses generally include [152]:

- *A 'review of individual cases'* or *'case series'* in a PV database or in published medical or scientific literature, as detailed in section 6.3.1.1.
- *'Aggregate analysis of case reports'* using absolute case counts, simple reporting rates or exposure-adjusted reporting rates, as detailed in section 6.3.1.2.

These approaches are particularly important in the assessment of designated medical events (DMEs) [29] or rare events for which clinical evaluation of an individual tends to carry a larger weight, and for which there may be an especially high premium on sensitivity over specificity.

6.3.1.1 Case and Case Series Review

The "index case" or "striking case" method is probably the most commonly used technique in traditional PV [153]. Trained product safety specialists detect signals while routinely reviewing submitted information, often during the initial intake assessment of ICSRs (clinical trials, spontaneous AE reports, or cases published in the literature). The identification of even one well-documented ICSR with an unusual "striking" feature can sometimes be interpreted as a signal, even though in practice, in most situations, strong suspicions about possible drug-event associations are usually based on a series of cases with similar reported features (clustering). Admittedly, such manual reviews are subjective and benefit from a thorough familiarity of the reviewer with the product pharmacology and the condition(s) for which it is indicated.

6.3.1.2 Aggregate analysis and Period reports

Aggregate reporting involves the compilation of safety data for a drug over a prolonged period of time (months or years), as opposed to single-case reporting which, by definition, involves only individual AE reports. The advantage of aggregate reporting is that it provides a broader view of the safety profile of a drug. Worldwide the most important aggregate report is the *'Periodic Safety Update Report (PSUR)'* [154]. This is a document that is submitted to drug regulatory agencies in Europe, the US and Japan (ICH countries). In these documents marketing authorisation holders are expected to provide succinct summary information together with a critical evaluation of the risk-benefit balance of the product in light of new or changing information. In the EU there is also a link between the periodic reporting and the EU Risk management plans introduced at the end of 2005.

6.3.2 Quantitative Signal Detection Methods

In comparison to the traditional signal detection methods, SDAs are currently and routinely used by PV experts for quantitative signal detection. SDAs can often be considered an activity related to *"knowledge discovery in databases"*, i.e., the process of extracting information form a large database [155]. The purposes of quantitative signal detection are many-fold and may vary depending on the local habit of PV experts. For instance, they can be used as an aid to the traditional case-by-case assessment as a screening tool to periodically generate a list of signals required for more in depth investigations (i.e., to prioritize signals) or, on an *ad-hoc* basis to detect complex data dependencies, which are

difficult to manually detect (e.g., drug-drug interactions or drug-related syndromes) [156].

There are two main types of SDAs; those based on *"disproportionality analysis (DPA)"* and those based on *"multivariate modeling techniques"* such as logistic regression (LR). The use of these statistical SDAs differs from their conventional use in that there is no prior hypothesis or null hypothesis of any specific drug-event association, and power calculations are not performed. The application of SDAs and particularly the concept of DPA methods will be discussed in detail in section 6.3.3, whilst multivariate modeling techniques and Bayesian hierarchical modeling methods for use in clinical trials are discussed in later sections.

6.3.2.1 When is the Database Large Enough?

Before considering the use of SDAs, the question of interest that often arises is, 'when can a safety database be classified as *"large enough"*?' This phenomenon can be thought of as function of the product and/or event incidence in the population, although to date there is a lack of explicit guidance on the specific population size. The Council for international organizations of medical sciences (CIOMs) working groups [150] and academic members in the past [157] have produced the recommendations detailed in Table 18, for the implications of PV signal detection based on population size; with real examples of ADRs detected with the specific approach and method used.

Table 18: Proposed population size for sampling in signal detection.

Event incidence in product takers*	Background incidence of event*	Example of event due to taking product	Ease of proving an association (method)	Approach used
Common	Rare	Phocomelia due to Thalidomide	Easy (clinical observation)	ICSR or Periodic Review
Rare	Rare	Reye's syndrome and Aspirin	Less easy (clinical observation)	ICSR or Periodic Review
Common	Common	Cough and ACE inhibitors	Difficult (large observational trials/data)	SDA
Uncommon	Common to Rare	Breast carcinoma and Hormone Replacement Therapies	Very difficult (large clinical trials)	SDA
Rare	Common	None Known	Virtually impossible	Virtually impossible

***Frequency of ADR as defined by CIOMs: very common (≥ 1/10 (≥10%)); common (≥1/100 and <1/10 (≥1% and <10%)); uncommon (≥1/1000 and <1/100 (≥0.1% and <1%)); Rare (≥1/10,000 and <1/1000 (≥0.01% and <0.1%)); very rare (<1/10,000 (<0.01%)).**

6.3.3 Disproportionality Analysis

There are many statistical methods to examine disproportionality, each with advantages and disadvantages. However, all methods have the main aim of demonstrating a difference between observed and expected reporting of events [52]. The DPA methods for signal detection as currently applied are purely statistical methods which do not include any recognition or adjustments for pharmacological, biological, clinical or demographic determinants of ADRs.

DPA is based on 2 x 2 contingency tables (Table 19), showing figures for ADRs (i) with the drug (j) taken, ADRs without the drug, and ADRs in the whole database with and without the drug:

Table 19: Two by two contingency table used in disproportionality analysis.

	Drug of interest (j)	Other Drugs	
ADR of interest (i)	n_{ij}	$n_{i\bar{j}}$	$n_{i.}$
Other ADRs	$n_{\bar{i}j}$	$n_{\bar{i}\bar{j}}$	$n_{\bar{i}.}$
	$n_{.j}$	$n_{.\bar{j}}$	n

- n_{ij} : Number of reports involving ADR_i for the $drug_j$.
- $n_{i.}$: Marginal count involving ADR_i
- $n_{.j}$: Marginal count involving $Drug_j$
- n : Total number of reports in database.

DPA can generally be divided into the two categories of frequentist and Bayesian, both relying on the aforementioned 2 x 2 contingency table. The most popular frequentist method is the Proportional Reporting Ratio (PRR) [158], whilst the Bayesian Confidence Propagation Neural Network (BCPNN) [159] and the Gamma-Poisson Shrinker (GPS) [160] are the most prominent and widely used techniques within a Bayesian framework. In most of the AE reporting databases, there is no valid exposure information or information for the total number of subjects taking a particular drug, therefore, DPA methods are all developed for investigating the relative reporting rate instead of relative risk. The PRR method is computationally straight forward and the relative reporting rate estimated from this method is easy to interpret. The BCPNN and GPS methods require more complex computations along with the elicitation of prior hyper-parameters using expert opinions or estimation from the data in an empirical Bayesian setup.

6.3.3.1 Proportional Reporting Ratio (PRR)

The Proportional Reporting Ratio (PRR) is given by Evans et al. [158] as:

$$PRR = \frac{n_{ij}/n_{i.}}{n_{\bar{i}j}/n_{\bar{i}.}}$$

This measure is actually the relative reporting rate of drug j for ADR i versus other ADRs. The standard error of ln(*PRR*), and lower and upper bound of the 95% two-sided confidence interval are usually obtained via an approximation of the normal distribution as:

$$SE\ (\ln\ PRR) = \sqrt{\left(\frac{1}{n_{ij}} - \frac{1}{n_{i.}} + \frac{1}{n_{\bar{i}j}} - \frac{1}{n_{\bar{i}.}}\right)}$$

$$95\%\ CI = e^{\ln(PRR) \pm 1.96\ SE\ (\ln\ PRR)}$$

The PRR is calculated for every drug-ADR combination. Each PRR can be either a true signal or a falsely discovered signal (false-positive), which is determined based on the lower bound of the 95% CI being above a threshold value of 1 [161]. When the PRR is calculated, the results tend to become unstable when the number of events (n_{ij}) is small, resulting in large estimates with wide confidence intervals [162]. This will often lead to many false-positive signals for very rare events. To uncover these false-positive signals, for instance, the biological plausibility has to be examined and/or confirmatory studies to re-assess the found signals using additional data sources have to be conducted. Though other statistical methods usually applied in cross-classification tables can also be exploited to resolve this issue [161], such as the χ^2 - test with one degree of freedom (with or without Yates's correction).

The instability of the PRR when applied to low drug-event counts, led to the development of the more advanced "Bayesian shrinkage" techniques. The two methods mainly used today are the BCPNN, which is applied by the Uppsala monitoring committee to analyze the WHO database, and the GPS which is applied to the adverse events reporting system of the FDA.

6.3.3.2 Bayesian Confidence Prorogation Neural Network (BCPNN)

The Bayesian approach proposed by Bate et al. [159] is used to evaluate apparent dependencies in a dataset. The measure of disproportionality used in the BCPNN model, is referred to as the "Information Component (IC)" [156]. Assume that the number of reports n_{ij}, and the marginal totals $n_{i.}$and $n_{.j}$ follow independent binomial models with Beta priors as follows:

$$n_{ij}|p_{ij} \sim Binomial(n, p_{ij}); \text{ with}$$

$$p_{ij} \sim Beta(\alpha_{ij}, \beta_{ij}), \alpha_{ij} = 1,$$

$$\beta_{ij} = \frac{1}{E(p_{i.}|n_{i.})E(p_{.j}|n_{.j})} - 1,$$

$$n_{i.}|p_{i.} \sim Binomial(n, p_{i.}); \text{ with}$$

$$p_{i.} \sim Beta(\alpha_{i.}, \beta_{i.}), \alpha_{i.} = 1, \beta_{i.} = 1,$$

$$n_{.j}|p_{.j} \sim Binomial(n, p_{.j}); \text{ with}$$

$$p_{.j} \sim Beta(\alpha_{.j}, \beta_{.j}), \alpha_{.j} = 1, \beta_{.j} = 1,$$

where p_{ij}, $p_{i.}$, and $p_{.j}$ denote the probability of the occurrence of the number of reports n_{ij}, and marginal counts $n_{i.}$ and $n_{.j}$. The priors for the marginal

probabilities ($p_{i.}$and $p_{.j}$) are actually uniform [0, 1] (non-informative). The parameter β_{ij} is determined using the relation that $E(p_{ij}) = E(p_{i.}|data) \times E(p_{.j}|data)$; that is, the prior mean of p_{ij} is equal to its posterior mean under independence, which is a product of the posterior means of the marginal probabilities $p_{i.}$and $p_{.j}$. Thus, β_{ij} is data dependent.

Bate et al. [159] defined the IC as:

$$IC_{ij} = \log_2\left(\frac{p_{ij}}{p_{i.} \times p_{.j}}\right)$$

Using delta method, and the fact that the posterior distributions of p_{ij}, $p_{i.}$and $p_{.j}$ are independent Beta distributions with updated parameters, the posterior mean and variance of the IC_{ij} are given by [51].

$$E(IC_{ij}) = \log_2 \frac{(n_{ij}+1)(n+2)^2}{(n+2)^2 + (1+n_{i.})(1+n_{.j})(n)},$$

$$Var(IC_{ij}) = \frac{1}{(\log 2)^2}\left[\frac{n-n_{ij}+\gamma-1}{(n_{ij}+1)(1+n+\gamma)} + \frac{n-n_{i.}+1}{(n_{i.}+1)(n_{.j}+3)} + \frac{n-n_{.j}+1}{(n_{.j}+1)(n+3)}\right]$$

where

$$\gamma = \left(\frac{n+2}{n_{i.}+1}\right)\left(\frac{n+2}{n_{.j}+1}\right).$$

Assuming normal approximation for the distribution of IC_{ij}, the 95% CI for IC is given as [51]

$$E(IC_{ij}) \pm 1.96 \times \sqrt{Var(IC_{ij})}.$$

A signal is defined if the lower bound of the 95% CI is greater than 0.

An updated version of the BCPNN has been presented by in Norén et al. [163], where the prior distribution is based on the joint Dirichlet distribution for the model parameters instead of independent beta distributions. Then an estimate for the 95% CI of the IC is achieved by Monte-Carlo simulations, which helps for better computational stability.

6.3.3.3 Gamma Poisson Shrinker

As an alternative, DuMouchel proposed the so-called Gamma-Poisson Shrinker (GPS) algorithm [160]. Here, the occurrence of the target drug-event combination is considered as a rare event, such that the observed drug-event combination count n_{ij} may be assumed as a realization of a Poisson-distribution random variable.

$$n_{ij} \sim Po(\lambda_{ij} E_{ij})$$

Where $E_{ij} = \frac{n_{i.} \cdot n_{.j}}{n}$ is the expected number of reports, and λ_{ij}'s are the parameters denoting the relative reporting rates. The λ_{ij}'s are assumed to share a common prior distribution, which is a mixture of two gamma distributions given by

$$\lambda_{ij} \sim \hat{\omega}\, Ga(\hat{\alpha}_1, \hat{\beta}_1) + (1 - \hat{\omega}) Ga(\hat{\alpha}_2, \hat{\beta}_2)$$

of which the five hyper-parameters are determined by maximizing the marginal likelihood of the n_{ij}'s. The posterior distribution of λ_{ij} is also distributed according to a mixture of two gamma distributions:

$$\lambda_{ij}^{*} \sim \mathrm{w}_{ij}\, Ga(\hat{\alpha}_1 + n_{ij}, \hat{\beta}_1 + E_{ij}) + (1 - \mathrm{w}_{ij}) Ga(\hat{\alpha}_2 + n_{ij}, \hat{\beta}_2 + E_{ij})$$

Initially, the association measure of interest proposed by DuMouchel [160] was based on the posterior expectation of the logarithm of the risk ratio λ_{ij}. However now signal detection is based on the fifth percentile of the posterior distribution of λ_{ij}, denoted as GPS_{05}, and a signal is generated if GPS_{05} is greater than 2 [164]. This Bayesian estimator gives more conservative risk estimates when event counts are small; risk estimates are considerably smaller and the CIs narrower, hence the denomination "shrinkage estimate". While this shrinkage might obfuscate a real signal by reducing it to a non-conspicuous level, it helps to eliminate false-positive signals, which otherwise would have to be adjudicated subsequently.

6.3.3.4 Threshold criteria

Currently, none of these signal detection methods (PRR, IC and GPS) is considered a reference method, and one or another of them is used routinely by monitoring agencies for national or transnational PV databases. The PRR is used for screening the MHRA sentinel and Eudravigilance databases [158, 165]. The GPS is used by the FDA for the US adverse-event reporting system [164], and the IC is used by the UMC for the WHO database [155]. When systematically screening the safety data within these PV databases, specific thresholds on the criteria have been proposed by the regulatory agencies, as detailed in Table 20.

Table 20: Defined criteria used in disproportionality analysis for the major stakeholders.

Signal detection algorithm (SDA)	Stakeholder in use of SDA	Criterion for signalising	Threshold on Criterion	Advantages	Limitations
Proportional Reporting Ratio (PRR)	MHRA Sentinel and EMA Eudravigilance	$PRR_{02.5}$ (Lower 5^{th} percentile of the relative risk reporting ratio distribution)	$PRR_{02.5}$ >1	Easily applicable, easily interpretable, more sensitive as compared to Bayesian method	Cannot be calculated for all drug-event combinations. Lower specificity
Information Component (IC)	World Health Organisation (WHO) - Uppsala monitoring centre.	$IC_{02.5}$ (2.5% quantile of posterior distribution of IC)	$IC_{02.5}$ > 0	Always Applicable, more specific as compared to frequentist method, can be used for pattern recognition in higher dimension	Relatively non-transparent for people non-familiar with Bayesian statistics. Lower sensitivity
Gamma Poisson Shrinker (GPS)	Food and Drug Administration (FDA) Adverse Event Reporting System (AERS)	EB05 which we will refer to as the GPS_{05} (lower 5^{th} percentile of posterior observed-to-expected distribution)	GPS_{05} > 2	Always applicable More specific as compared to frequentist method	Relatively non-transparent for people non-familiar with Bayesian statistics. Lower sensitivity

These thresholds are still in use to date, however the EMA good PV guideline states that the threshold criteria for detecting signals can be adjusted [166]. It is also suggested that this may vary depending on the "severity" of the AE and "size of the dataset". For example in one study [167] using the multi-item gamma Poisson shrinker (MGPS) with threshold $MGPS_{05}$ > 2 in the FDA database, it was suggested that serious events such as hyperkalaemia, pancreatitis, and rhabdomyolysis were often undetected. Therefore it is recommended that a

much stricter signaling threshold (e.g., MGPS > 1) is required. Though this study is limited, and would require further exploration of other threshold values. The use of other threshold values (i.e., 1.5, 2, 4, 8) for the EB05 with the GPS method were explored in one study [164], and differences in sensitivity and specificity of signal elicitation through time when the various signal thresholds are used was investigated. This study also determined that lower threshold values improve the detection of more severe AEs, however the authors did not consider implications of false discoveries (type I error) when adjusting the threshold value.

Most published evaluations of these techniques are mainly limited to large regulatory databases, but their performance characteristics may differ in smaller safety databases of drug developers. In a recent study [168] the database size and power to detect safety signals were compared across the three safety databases (GlaxoSmithKline, FDA and WHO) where a random subset of drugs was selected. In this study it was shown that the power to detect was highest in the database with most AE reports. In general a database with the most drug-specific data will achieve the highest power. Larger database systems will also enhance the potential of early safety signal detection. However this study was limited to only investigating regulatory and large pharmaceutical company databases, therefore further investigations are need in smaller company databases.

At present, there exists no specific guidance for using different threshold criteria when considering the severity of the AE and the database size. It has been noted

in previously published literature that smaller based drug companies may adjust the threshold value to improve the sensitivity when detecting signals [169], although this will likely affect the specificity and hence increase the number of false signals detected. This was not considered in the conclusions of this study, and hence requires further investigation.

6.3.3.5 Performance characteristics

Recently, the application of PV signal detection through DPA has been subject to debate and criticism [170]. Some benefits and strengths of using SDAs are undisputed. They are generally quick and inexpensive methodologies routinely performed by regulators and researchers for drug safety evaluation [171]. A major disadvantage of signal detection and the methods of DPA is that they detect too many signals for drug-event combinations that are falsely discovered. There have also been various investigations [172-174] examining the characteristics of these methods and the appropriate criteria for each method, but to date no clear guidelines or gold standards have been established.

Although both drug companies and regulatory agencies require information on AEs in the same manner, their circumstances and objectives are different. The AE databases that are used by pharmaceutical companies generally consist of fewer drugs and have fewer reported events than the spontaneous reporting systems used by regulatory agencies [119]. Pharmaceutical company databases tend to compile data from related drugs into the 'all other drugs' category ("$n_{i\bar{j}}$" in Table 19), which can conceal significant drug-event relationships due to the high frequency of events associated with other drugs.

Pharmaceutical companies have various means, such as pharmacological examination or scrutiny of clinical data, for examining whether a signal is an ADR or not. Therefore, the balance between sensitivity and specificity requirements may differ between pharmaceutical companies and regulatory agencies. For example the SDA used by a pharmaceutical company may be required to maintain specificity at an acceptable level (e.g. ≥ 95%) while providing the greatest possible sensitivity. Due to the sensitivity issues associated with the current thresholds for the SDAs, these methods are considered to be inappropriate for use by pharmaceutical companies and smaller organizations like CTUs without suitable modification. Therefore an 'adjustment to the threshold value' may be required to make them suitable for the characteristics of the AE reporting databases to enable them to provide the performance required.

6.3.3.6 Caveats

Different groups of healthcare professionals might report suspected ADRs: nurses, pharmacists, dentists, hospital doctors and outpatient doctors [175]. Additionally consumers may wish to report. The reporter type may systematically affect the type of data collected. The method may therefore need to be adapted depending on the reporter, as the proportion of serious reactions reported may well vary between reporting groups.

For all the SDAs a comparison is made to the generalizability of the database. However if two drugs cause the same adverse reaction at the same incidence but one drug also causes many other adverse reactions, then despite the 'n_{ij}'

value being the same for both drugs, the '$n_{\bar{\imath}j}$' value will be much higher for the drug that causes lots of other adverse reactions. Thus, for the drug with a uniquely reported ADR, will result in a higher measure of disproportionality than the drug reported with many different ADRs, despite the true incidence of the adverse reaction being the same for both drugs [156].

The terminology used for coding ADRs can have a large impact on the signal detection system. If a drug causes an adverse reaction, but no specific adverse reaction term exists in the dictionary used for coding that ADR report, then the signal may be missed [176]. The structure of hierarchical terminologies used for AE classification makes their potential for signal detection on a group level unclear, when several different yet similar AE terms might be used to code a specific pharmacological effect. Thus often resulting in misclassification and the potential lumping of AEs into inappropriate subgroups as highlighted in past research [30].

6.3.3.7 Refinements to Signal Detection – What could be done?

SDAs have accepted limitations but there is a growing appreciation that such approaches are needed to make the most of large repositories of reported AE data. Therefore there are a number of important considerations for potential refinements when using SDAs.

The acceptable rate of 'false positives (type I errors)' and 'false negatives (type II errors)', will depend on the specific function of the signal detection system. Whether to highlight with high risk signals very early, or whether to be later but more confident is the key question. Repeated false alarms for signals lead to

constant clinical evaluation and work needed. Eventually the alarm will be disregarded and a true signal might be ignored. However having a limited number of false positives is preferable to missing genuine safety signals. Hence there is a general need for more powerful methods using the '*False Discovery Rate (FDR)*' as a measure of error. FDR is now regularly used for multiple testing in the genomic analysis field; however PV signal detection also involves multiple testing between drug and ADR combinations in large volumes. Therefore the DPAs methods were recently revisited in a multiple testing framework and are now able to obtain an estimate of the FDR. These methods will be discussed in more detail in chapter 7.

Research has been applied to the use of trend analysis in signal detection in the past. There are limitations in doing this work since the irregularity of reporting for some systems, and the onset date of the adverse reaction is often missing from reports. Trends are important, and their investigation leads to new insights about the methodology as well as interesting PV information [163]. For example the WHO recently examined the association between Captopril and the ADR term coughing to determine whether the recent changes to IC analysis would delay or expedite the highlighting of this signal. The association was highlighted earlier by observing the change over time in number of cases reported using IC analysis. Moreover, the choice of baseline (i.e. for estimating the expected), level of terminology (i.e. for coding AEs), method used, and stratification variables do affect which combinations are highlighted. It is as important to see what is not highlighted (and to what degree), as is to see what is highlighted.

6.3.3.8 Real-world value of Signal Detection Algorithms

Adverse drug effects are manifold and heterogeneous. Many situations may hamper the signal detection (i.e., the detection of early warning signs) of adverse effects and new signals often differ from previous experiences. Signals have qualitative and quantitative aspects. Different categories of adverse effects need different methods and resources for detection. Current PV is predominantly based on spontaneous reporting which is mainly helpful in detecting type B effects (those effects that are often allergic or idiosyncratic reactions, characteristically occurring in only a minority of patients and usually unrelated to dosage and that are serious, unexpected and unpredictable) and unusual type A effects (those effects that are related to the pharmacological effects of the drug and are dosage-related), though other sources of signal detection may also include PEM and large automated data resources on morbidity and drug use (including record linkage). Type C effects (those effects related to an increased frequency of 'spontaneous' disease) are difficult to study, however, and continue to pose a pharmacoepidemiological challenge on resources [177].

The appropriate frequency (i.e., numbers needed) of data review for signal detection is determined by, among other factors, the risk inherent in the product and may be specified in the PSUR and/or Risk Management Plan (RMP) (if applicable). Some common determinants of frequency of data review to consider are: Number of AEs/ADRs received per year, potential public health impact of AE (e.g., patient exposure data), maturity of the product (e.g., number of years on the market) and the safety profile of the product and whether there

are events that are being actively monitored [178]. However one of the main limitations of spontaneous reporting systems is their inability to provide the denominator (i.e., the number of patients actually consuming the drug of interest), which has a major impact when determining the numbers needed to use signal detection methods.

In past literature it has been suggested that SDAs may be unreliable when the number of reports for a drug-event association is less than 3 [179], which shows their general inadequacy and fallacy when detecting uncommon and rare events. It has been also suggested that spontaneous reporting systems may not be suitable when detecting adverse effects with frequency (>1/10), and therefore clinical trials are preferable [180].

6.3.3.9 Signal Detection Algorithms use in Electronic Health Databases

SDAs are now also being used on longitudinal electronic health databases for post-marketing surveillance. A recent study [53] has critically reviewed the use of these methods in observational electronic health care claims and administrative data settings. This study highlighted some of the potential pitfalls, indicating that some of the methods are susceptible to systematic bias like the longitudinal GPS method, whilst other frequentist methods (PRR and ROR) appear unreliable [181]. When electronic health database studies detect no drug risk, there are often no robust and accepted standards to judge a causal effect or whether the study was incapable of detecting it. There is a requirement for improved reliability of risk assessments based on these databases, and the current limitations need to be fully understood [182].

6.3.4 Multivariate Techniques

Although cumulative experience with DPA methods described in section 6.3.3 has shown to be a promising adjunct in safety analysis, the reduction of drug-event combinations in two dimensions may result in the loss of crucial clinical information. Two-dimensional DPA approaches do not support the discovery and/or analysis of more complex or higher-dimensional drug safety phenomena that involve more than just one drug and one event. The importance and difficulty associated with the detection of these more complex drug safety phenomena have been noted in several prominent PV reports [174, 177], suggesting that more elaborate methods, henceforth collectively referred to as "multivariate methods", are required.

The multivariate logistic regression modeling based SDAs recently introduced in 2013, now adjust for confounding factors by co-medication (given the lack of other confounding information in the database). Confounding by co-medication can theoretically be addressed by using all drugs in a database as regression predictors for an event. Further efforts have been made in an attempt to address the concealed effects caused by confounding. However one study suggests that significant concealment is rare in large spontaneous databases, and that it mostly affects rare events [183]. These methods can also be difficult to implement, and their running process can be time consuming.

6.3.5 Bayesian Hierarchical Modeling in Clinical Trials

In the past detecting signals from clinical trials data has primarily been performed using traditional frequentist tests (e.g., fishers exact test, chi-squared

tests etc), which do not account for multiple testing. As an alternative Berry & Berry (2004) [184] proposed the Bayesian thee-level hierarchical mixture model (BHMM) for the analysis of AEs as a way of coping with multiple testing. This approach allows for explicitly modeling AEs with the existing MedDRA coding structure, so that strength can be borrowed within and across system organ classes (SOCs).

The idea is that there is a distribution of AEs inside each SOC group, then if we regard the AEs as being randomly picked from that distribution (exchangeability assumption), then we could use the distribution in each SOC group as a prior for each AE in that group. For example, the three-stage model assumes there are B body systems. Within body system b there are k_b types of AEs labeled A_{bj}, where b = 1,…., B and j = 1,…., k_b. Stage 1 priors have a normal prior distribution, Stage 2 we assign a prior distribution to a set of hyperparameters. In this stage the distribution varies from one body system to the next. Finally in the third level of the model the parameters of these distributions are assigned prior distributions to the hyperparameters of a beta distribution. The calculations of this model are carried out using Markov Chain Monte Carlo (MCMC) methods to simulate from the posterior distributions.

However, there have been questions in fitting the BHMM using ordinary logistic regression, suggesting that it may not be possible due to the sparsely reported nature of many AEs which will likely cause estimation to fail [185]. Therefore further research is needed into these Bayesian hierarchical methods and there is still a need for software development.

6.4 Signal Prioritization

Signal prioritization is a first critical step after the signal detection stage. Evaluating all signals generated (i.e. single or aggregated reports) in detail has major resource implications as many will turn out not to be real ("false alarm") or alternatively may require action. This is not to say that the signal can be dismissed without some kind of evaluation. The prioritization process implies that all signals will be reviewed but some more expeditiously than others. In this respect, there is general agreement that unexpected serious signals occurring during the first years post-marketing should be looked at as a priority in order to establish as rapidly as possible the safety of the drug under evaluation.

Given the number of signals produced, smaller companies may not need to prioritize signals for a particular product, choosing instead to assess all detected signals. However, for most companies a process for prioritization of these signals is required. Prioritizing allows action to be taken more expeditiously for higher priority signals than for other signals. For small to medium sized companies, assessing all signals in detail is resource intensive because of the high number of false positive signals. Larger companies may consider adopting an approach similar to the MHRA *'Impact Analysis'* for signal prioritization, where the impact of a signal is summarized through two scores [186]:

1. Quality of evidence (strength of evidence for causality, e.g. Bradford-Hill Criteria [187]).
2. Public health impact of the signal.

The MHRA impact analysis produces a four-level categorisation dependent on the strength of evidence for casual effect, potential public health implications, public perceptions and agency obligations. This then leads to a proposal for further action from high priority signals which need further assessment to the lowest priority signals which require no immediate action (e.g. may either be closed or require further monitoring only). For small and medium companies a more informal approach using the factors above can be used [150] as long as this is justified and documented. The company may consider prioritising using one or more of the "always serious" lists below:

- "Always Serious" ADRs and designated medical events.
- Other Examples:
    - The Council for international Organizations of Medical Sciences (CIOMS) working group V [150].
    - EMA Important Medical Events List [188].

Additionally, expectedness can often be used as part of the prioritisation process.

6.5 Signal Evaluation

After a signal is prioritised, other sources of data should be systematically assessed to determine whether sufficient evidence of "causality" exists, and what further action, if any, may be required. The sources of evidence can include [112]:

- The ICSR(s) that triggered the signal.

- Other ICSRs with similar event terms identified (e.g. by using Standardized MedDRA Queries (SMQs)).
- Scientific literature and/or systematic reviews
- Clinical trial and pre-clinical data (i.e., SmPCs and IBs)
- Epidemiological data.

The use of SMQs is recommended in order to retrieve and review similar cases of interest when potential signals are identified within a database. In practice many signals can be accessed on the strength of the ICSRs that triggered the signal in the first place. Depending on the case load (number/volume of cases), the data may be stratified according to age, gender, ethnicity, concomitant medication or disease. This may identify populations at highest risk for the event and also reduces confounding. A judgment about whether a signal is validated depends on the number and quality of case reports, the nature of the reaction, type of drug and the population exposure.

The evaluation stage of a signal is often a resource intensive and time consuming process. For example, in one study [189] investigating the use of the high-strength pancreatin supplement Nutrizyme for patients with cystic fibrosis, there were reported causes of sub-acute intestinal obstruction due to a fibrotic stricture of the ascending colon in a child with cystic fibrosis. Though, more recent similar cases suggest that this new pathology is linked to the use of enteric-coated high strength pancreatin microspheres, which resulted in a drug safety update in 1998 from the UK's committee on safety of medicines advising on the dosage of the treatment.

Once a signal has been evaluated there are three possible options following the decision making stage:

- *Close signal:* The signal was refuted based on the available evidence and no further action is required. The decision and rationale for closing a signal should be documented. However, if further evidence becomes available the signal can be re-assessed.
- *Continue monitoring:* In some circumstances a decision cannot be made until the evidence supporting the signal is strengthened. Except for situations of extreme risk, these signals are monitored until sufficient evidence becomes available to either confirm or refute the signal. The decision and rationale to justify monitoring a signal should be documented.
- *Take further action:* After a signal is validated further action is required. The decision and rationale to take further action for a signal should be documented. The actions may include the following; notify the Qualified Person for PV (QPPV), enhance monitoring or follow-up techniques, consult internal or external experts, targeted clinical investigations, comparative observational studies, active surveillance schemes and clinical trials.

6.6 Discussion

The development, testing and deployment of SDAs represent a quantum jump in PV. Although there is currently no scientific or regulatory basis to claim that

SDAs are a required element of good PV practice, they are an intuitively appealing solution to the operational challenges of screening steadily enlarging safety databases [109]. Higher-order phenomena, such as complex drug-drug interactions or drug-induced syndromes, may be especially difficult to identify through manual review of AE line listings, and it is this type of phenomena which might be most amenable to detection through the use of SDAs.

Retrospective applications indicate that SDAs can highlight some medically significant associations in a timely manner, often in advance of the published literature and traditional methods. As a result SDAs have been incorporated into routine signal management frameworks for most major national and transnational drug safety monitoring centers, including the MHRA (PRR), the WHO (BCPNN) and the FDA (GPS) [52]. However, SDAs and DPA methods may fail to highlight legitimate associations for various reasons; they often have an unclear opportunity cost associated with false alarms (false discoveries); and have yet to prospectively detect new drug hazards.

There are formidable challenges to validating SDAs beyond those already mentioned, such as the choice of appropriate reference AEs (true positive and false negative signals) for assessing SDA performances in the absence of perfect gold standards for adjudicating causality [174]. However findings of a disproportionality ratio for a drug should lead to a new reinvestigation of data from experimental pharmacology and RCTs. It should also stimulate specific case-control or cohort analysis to strengthen the generated hypothesis.

Accordingly, signal detection should be considered as one of many potentially performance-enhanced options in the toolkit for detecting safety signals that need to be assessed by each institution on an individual basis. They should only be considered potential supplements to, and not substitutes for, a comprehensive signal detection programme based on multiple approaches and data sets. In this chapter we have clearly underlined some of performance related issues with the SDAs when analyzing harms data and suggestions for improvements have been made. In chapter 7 we will explore the use of SDAs further to investigate their ability to detect signals in clinical trial databases of a smaller scale.

Chapter 7: Signal Detection Algorithms for Analyzing Harms data - Simulation Study

Part of the objective in chapter 5 was to explore current practice and future potential for use of SDAs to mine harms data. However the results have shown that there appears to be uncertainty of their application in CTU databases. Chapter 6 provided an extensive overview of SDAs, discussing in detail their characteristics and potential for refinement in the future.

In this chapter a literature review of recent studies that have assessed the use of SDAs is presented (Section 7.2). The performance of the three SDAs introduced in chapter 6 is then explored in detail in a simulation study (Section 7.3) to explore their properties under different conditions. The aim of the concluding part of this chapter is to explore whether these methods might be suitable for detecting signals in harms databases which are likely to be on a smaller scale than post-marketing surveillance systems, such as those which CTUs may have access to.

7.1 Introduction

For identifying safety signals of AEs from reported reactions, SDAs are increasingly being used to supplement the traditional expert review of the reports and to analyze the large volume of accumulated data more rapidly. Disproportionality analysis represents the main type of SDAs, where their methodologies use frequency analysis of 2 x 2 contingency tables (Table 19) to

quantify the degree to which a drug-event combination co-occurs disproportionately, as compared with what would be expected if there were no association [177].

In general SDAs are designed to compute surrogate measures of statistical association between drug-event pairs reported in a database [52]. These measures are often interpreted as signal scores, with large values representing true adverse drug reactions (ADRs). A signal score threshold is often used to highlight signals worthy of further review [173]. These threshold values can be adjusted to reduce false signals but at the expense of reduced power; in other words, the risk of missing a true signal will be potentially increased. Therefore, it is essential to identify statistical methods that can control false findings at an acceptable level without compromising on the power [167].

7.2 Literature Review

Although the value of SDAs has been widely recognized [109], their performance characteristics are not well understood [162]. This is due to the lack of evaluation guidelines and absence of established gold standards [158], and to a certain extent, acknowledged shortcomings in the studies that have been conducted so far.

The EMA have recently published their guideline on good pharmacovigilance practices [188], which states that the *"size of the data set should be taken into account when considering the use of SDAs"*. However from this it is unclear as to when they should and should not be used in relation to the database size. The guideline also states that the *"selection of the threshold criteria for the detection*

of signals should also be taken into account", although there is no explicit gold standard regarding the use of different thresholds for different scenarios.

To explore performance in more detail a literature review was firstly undertaken to summarise characteristics of other studies that have assessed SDAs. Studies with their primary objective(s) to explore the performance of alternative SDAs were included, and other studies were excluded. For example most studies have simply used one of the SDAs to generate a list of signals for further evaluation, and have not drawn any conclusions about the performance characteristics of the SDA. These studies were excluded. For the included studies information was collected on the journal of publication, purpose of research, methods used, data source and size of dataset (number of drugs and events reported), performance metrics, limitations and conclusions of the study.

The following strategy was used in MEDLINE which was searched from 2000 to 10th March 2014:

1. Signal detection.ti.
2. Data mining.ti.
3. Disproportionality analysis.ti.
4. 1 or 2 or 3
5. Limit 4 to yr = "2000 - 2014"

Sixty nine studies were identified in MEDLINE. Full articles were screened, and six studies met the inclusion criteria and assessed the performance of SDAs as their primary objective. These six studies are now described in Table 21.

Table 21: Characteristics of six studies assessing signal detection methods.

	Study Author (Journal publication date)					
Characteristic	Roux (2005)	Alvarez (2010)	Harpaz (2012)	Almenoff (2007)	Ahmed (2009)	Lehman (2007)
Journal of publication	Journal of Biomedical and Health Informatics	Drug Safety	Nature: Clinical Pharmacology & Therapeutics	Nature: Clinical Pharmacology & Therapeutics	Statistics in Medicine	Nature: Clinical Pharmacology & Therapeutics
Purpose of research	Evaluate the performance of signal detection methods on simulated data	Evaluate early signal detection in the Eudravigilance	A review of the recent methodological innovations and data sources used to support discovery and analysis of ADEs	A review of the statistical concepts behind DPA methods, and their application in pharmacovigilance	Explore two DPAs methods in a multiple hypothesis testing framework, for comparing multiple drug-event comparisons	Understand the value of the GPS to detect safety signals in relation to the current PV signal detection methods
Method(s) explored	PRR, ROR, Yule's Q, the sequential probability ratio test (SPRT), Poisson and chi-squared, IC, empirical bayes arithmetic mean (EBGM), and the alternative method empirical bayes probability (EbP). Standard thresholds used for all methods	PRR ≥ 1	All DPAs, multivariate methods and others reviewed in detail	All DPAs reviewed	GPS and BCPNN standard thresholds were used for both	GPS ≥ 2
Data source and size	Simulated data sets (150 drugs and 100 AEs)	The EMA EudraVigilance including 267 medicinal products between September 2003 - March 2007	FDA Adverse event reporting system and the WHO-UMC	Diverse	Simulated data including 500 data sets, 634 drugs and 756 AEs, and data from the French national PV database collected between 1984 - 2002, including 672 drugs	Merck's post-marketing safety database including four products with 4389 product-events pairs reported between 1993 and 2004

					and 820 AEs	
Performance metrics used	Receiver operator characteristic (ROC) curves	Time taken to detect safety signal and sensitivity performance recorded	-	-	False discovery rate (FDR)	Sensitivity, specificity, PPV and NPV calculated, and time evaluated until detection
Limitations	• Assessed on simulated data therefore real world value difficult to determine. • The methods have now been updated. Particularly the IC and GPS methods. • Only the standard thresholds were used.	• Investigation was carried out in large SRS. The characteristics of smaller based companies will differ. • Only the sensitivity was used as performance indicator, but the false positive signals are also important to consider.	• Analysis restricted to large SRSs (FDA adverse event reporting system and WHO-UMC). • The performance characteristics may differ in smaller databases. • Heterogeneous way in which data is collected in large spontaneous databases differs to reporting in CTUs	• The use of different threshold values was not investigated. • Implications on database size not invested. • No discussion for performance related to the diverse data sources used.	• Methods tested on national database and simulated data of limited size. • Only two SDAs were assessed in the study, other methods should be compared. • Only the standard thresholds were used will no investigation of the sensitivity – FDR trade-off.	• Pharmaceutical company database of limited size used to carry out this research. • Analysis restricted to only including four products doesn't represent a diverse range of products. • Only the standard thresholds were explored.
Conclusions	The EBAM and IC provide the better results, however their theoretical background and implementation are less obvious than other methods like chi squared and SPRT	Statistical signal detection can provide significant early warning in large proportion of safety problems, however not all safety issues can be detected more quickly than other PV processes	There is a diverse portfolio of signal detection approaches aligned to different strategies and objectives for the analysis and detection of post-approval ADEs. Although they lack proper gold standards	Statistical signal detection methods can be used to identify new safety issues. However additional tools are need for identifying and characterizing rare, serious events.	The GPS provide the lowest FDR. However further research is needed to explore the prospects of using FDR in signal detection analysis	GPS method using Merck's safety database demonstrates sufficient sensitivity and specificity to be considered for use as an adjunct to conventional signal detection methods

Roux [172] assessed the performance of ten signal detection methods on simulated data including 150 drugs and 100 AEs. These methods were investigated using only the standard thresholds and their performances were evaluated by constructing the receiver operator characteristic (ROC) curves. The empirical bayes arithmetic mean (EBAM) and information component (IC) methods provided the best results, as was determined from the ROC curves. However, these methods were more difficult to implement than the chi-squared and sequential probability ratio test (SPRT). Since this study was conducted over 9 years ago some of the methods have now been updated and are no longer in use.

Alvarez [165] evaluated whether statistical signal detection in the Eudravigilance database can lead to earlier detection of drug safety problems when using the proportional reporting ratio (PRR) method. 267 medicinal products were included in the study as reported between September 2003 and March 2007. The focus was mainly on sensitivity rather than on the trade-off between sensitivity and specificity for the PRR method. The study concluded that statistical signal detection can provide early detection and warning of safety problems, although not all safety issues are always detected.

Harpaz [173] reviewed all current SDAs, both DPAs methods and multivariate modeling methods. However this study is restricted by only investigating the use of these methods in the FDA adverse event reporting system and the WHO Uppsala monitoring centre. They discuss a range of different approaches that can be used in signal detection, but also highlight that further work is needed to

develop gold standards when using these different methods. Similarly, Almenoff [174] has reviewed the statistical concepts behind all DPA methods, and their application in PV across a diverse range of data sources. In this study there was no discussion regarding the use of different threshold values when evaluating the performance of the methods. The study also suggests that additional tools are need for identifying and characterizing rare and serious events.

Ahmed [191] explored two DPA methods (GPS and BCPNN) in a multiple hypothesis testing framework for comparing multiple drug-event comparisons. These methods now make it possible to derive, with a non-mixture modeling approach, Bayesian estimators of the false discovery rate (FDR). The FDR constraint determines how many false signals are generated, and can be useful when analyzing signals, as will be discussed later in this chapter. These methods were assessed on simulated data based on 634 drugs and 756 AEs, and data collected from the French national PV database including 672 drugs and 820 AEs. The methods produced identical performances according to the operating characteristics sensitivity and specificity, however the GPS method performed better by providing the lowest FDR. These methods based in a multiple hypotheses testing framework require further research to explore their full potential, additionally they need to be compared against other SDAs.

Finally, Lehman [169] evaluated the GPS method performance when detecting safety signals in relation to traditional PV methods. The sensitivity, specificity, positive predictive and negative predictive values were used as the metrics of performance. The study has assessed the performance of the GPS using only the

standard threshold, and the analysis was restricted to a pharmaceutical company database with data collected from 1993 to 2004 for four products only. There were a total of 4389 product-event pairs reported for these four products over the time period. The study concludes that the GPS method demonstrates sufficient sensitivity and specificity to be considered for use in addition to conventional detection methods.

7.2.1 Improving Signal Detection in the Future

Disproportionality analysis is based solely on aggregate numbers of reports and naively disregards report quality and content. However, these latter features are the very fundament of the ensuing clinical assessment. The following variables may provide strong predictors of emerging drug safety issues: the number of informative reports, recent reports, and reports with free-text descriptions; disproportional reporting; and geographic spread. Simultaneously accounting for these aspects of strength of evidence can significantly improve the accuracy of automated screening of individual case reports with disproportionality analysis alone [192].

Combinatorial signal detection has been pursued in few studies up until recently, employing a rather limited number of methods and data sources but illustrating well-promising outcomes. However, the large-scale realization of this approach requires systematic frameworks to address the challenges of the concurrent analysis setting. In a recent study [193] a semantically-enriched framework was designed to address some of these issues, and particularly highlight contribution in:

1. Annotating data sources and analysis methods with quality attributes to facilitate their selection given the analysis scope
2. Consistently defining study parameters such as health outcomes and drugs of interest, and providing guidance for study setup
3. Expressing analysis outcomes in a common format enabling data sharing and systematic comparisons
4. Assessing/supporting the novelty of the aggregated outcomes through access to reference knowledge sources related to drug safety.

This framework brings forth a new perspective on large-scale, knowledge-intensive signal detection, and aspires to increase the efficiency, automation, support and collaboration for PV stakeholders.

7.3 Simulation study

The evidence from the literature review has shown that there is a current lack of gold standard available when verifying the threshold criteria for SDAs. There also appears to be no guidance available when using the methods in databases of limited size, and their ability to characterize and identify rare events has not been fully explored. More recently the methods were extended in a multiple hypothesis testing framework which now allows the performance of the methods to be assessed in relation to the FDR. However these methods require further testing to understand their full potential.

To explore the use of SDAs to investigate each of these key component areas in more detail, a simulation study is required.

7.3.1 Simulation study objectives

The objectives of the simulation study are described below:

1. To investigate the use of SDAs in an AE reporting system considering different threshold values.
2. To investigate the use of SDAs for identifying and characterizing rare events, by considering different scenarios affecting the incidence and risk of signals.
3. To investigate the use of SDAs in smaller scale systems, by simulating scenarios to mirror the type of harms data that might be collected in CTU databases.

7.4 Methods

7.4.1 Signal Detection Algorithms (SDAs) under investigation

The SDAs examined in this simulation study were the PRR, IC and GPS. These methods were chosen as they are currently under use by national and international regulatory agencies (MHRA, EMA, WHO and FDA) and are the most commonly used methods. The standard threshold criteria for these three SDAs used were the PRR ($PRR_{02.5}$) > 1 [158], IC ($IC_{02.5}$) > 0 [163] and the GPS (GPS_{05}) > 2 [160], and are explained in chapter 6, Table 20. These thresholds are not a gold standard but are commonly used by the regulatory agencies due to their reasonable sensitivity-specificity trade-off performances on their AE databases. However, they may not be suitable in smaller scale databases and for detecting rare events since they are regarded as being too specific, and therefore have low sensitivity performances.

7.4.2 Simulation model – Data Generation

The model for simulating the data was proposed by Roux [172], which introduces a procedure for simulating an AE reporting system, where the reporting process is viewed as a Poisson-distribution. In this model for any given ADR during a given period, the number of reports (ρ_{ij}) is assumed to follow a Poisson distribution defined as:

$$\rho_{ij} \sim Po\left(T_j \cdot RR_{ij} \cdot I_i \cdot pr_{ij}\right)$$

Where the parameter T_j is the drug exposure frequency (i.e. the number of patients exposed to drug (j) during a given period), RR_{ij} is the risk ratio related to the ADR, I_i is the background incidence of the AE (i), and pr_{ij} is the reporting probability of the ADR combination.

7.4.2.1 Model parameter selection

A number of different data sources were used to inform the choice of parameters within the model used to simulate the data.

The EMA's 'Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT)' project [194] ADR database [195] was used to obtain information on the frequency of AEs. The database accumulates suspected reports of AEs as reported in the European summary product characteristics (SmPCs) for all EU licensed products, and then compiles the data into a central ADR repository which can be accessed by the public. The data lock point for collection of these reports is 31st December 2013. However this database does not contain information on drug exposures, therefore

prescribing-level data from the health and social care information centre (HSCIC) [196] was used to approximate the exposure frequencies of the UK marketed drugs. Since the prescription data is split annually and can only be accessed in one database by each year individually, the data collection period was restricted from 1st January 2013 to 31st December 2013. This time frame was also used to collect data from the PROTECT ADR database for consistency.

7.4.3 Metrics for comparing the performance of different SDAs

The SDA threshold is often used to highlight safety signals of interest. The threshold can be adjusted to reduce false signals, or to improve the sensitivity performance (power) when detecting true signals. However since there is a lack of gold standard for determining which thresholds to use for these SDAs, in this study we aim to try and identify thresholds that provide a balanced trade-off between the FDR and sensitivity performances.

To investigate this trade-off we firstly explore the use of the commonly used thresholds (i.e., $PRR_{02.5} > 1$, $IC_{02.5} > 0$ and $GPS_{05} > 2$) and then explore the use of different threshold values to achieve higher sensitivity performances. This will be explained in the following sections, along with the FDR, sensitivity and specificity estimations.

7.4.3.1 False Discovery Rate (FDR)

In 1995, Benjamini and Hochberg (BH) [197] introduced the concept of FDR, as a statistical method used to correct for multiple comparisons. In a list of findings, FDR procedures are designed to control the expected proportion of incorrectly rejected null hypotheses (“false discoveries”).

The FDR has attracted growing interest over the years, mainly in the genomic data-analysis field, because it is particularly adapted to screening studies involving large numbers comparisons of genomic expressions. There are similarities between the genomic data analysis and PV signal detection analysis fields, where many drug-AE comparisons are analysed in order to determine true signals. As a result the SDAs (PRR, IC and GPS) were recently revised in a multiple-hypothesis testing framework, where they are now formulated as the statistical choice of a tested hypothesis. These methodological developments have resulted in new decision rules based on P values for the frequentist PRR method [198], and on the posterior probability of the null hypothesis for the Bayesian methods (IC and GPS) [191]. In particular the PRR methods consist of the popular approach in assuming a mixture model for the marginal distribution of the p-values and the Bayesian methods (IC and GPS) involve a mixture model describing the distribution of the testing statistic with one of the components corresponding to the null hypothesis.

For these new decision rules it is now possible to obtain, for any detection threshold, an estimation of the FDR. This criterion, which may be defined in the PV signal detection field as the expectation of the proportion of false discoveries (FDP) among a generate list of signals, can easily be estimated by obtaining the FDP from each simulated dataset then averaging the FDP over all simulated datasets to obtain the FDR [191, 198]:

$$FDR = Exp(FDP) = \frac{1}{S}\sum_{1}^{S} FDP$$

where *S* is the total number of simulated datasets.

The advantages of these new SDAs based in a multiple hypothesis setting enable us to determine the measure of error within any generate list of signals, which could save the time spent during the analysis and clinical evaluation stages.

7.4.3.2 Sensitivity-Specificity Trade-Off

To measure and compare performance of the different SDAs the sensitivity and specificity for each simulated dataset are calculated using the notation in Table 22. The advantage of using simulated data is that we know the true status of the signal, which means that the sensitivity and specificity are exact in each dataset. This is opposed to using a real dataset, where the true status of a signal is normally unknown.

Table 22: Description of sensitivity and specificity calculations, for each simulated dataset.

		Truth		
		Signal (1)	No signal (0)	Totals
Signal detected	Yes	A	B	A + B
	No	C	D	C + D
	Totals	A + C	B + D	

For each dataset the sensitivity is calculated as;

$$Sensitivity = \frac{A}{A + C}$$

with mean sensitivity across all datasets calculated as

$$Mean\ Sensitivity = \frac{1}{s}\sum_{1}^{s}\left(\frac{A}{A+C}\right)$$

And the specificity for each dataset and mean specificity across all datasets is calculated as;

$$Specificity = \frac{D}{B + D}$$

$$Mean\ Specificity = \frac{1}{s}\sum_{1}^{s}\left(\frac{D}{B + D}\right)$$

There is also a trade-off relationship between sensitivity and specificity. Changing the SDA threshold value causes the sensitivity and specificity to change in tandem. Therefore the threshold value at a specific sensitivity was determined by changing the probability threshold in small increments (0.025, 0.05, 0.1, 0.2,…, 0.9, 0.95, 0.975), and the receiver operator characteristic (ROC) curves constructed, for all events combined, by plotting sensitivity along the vertical axis and '1-specificity' along the horizontal axis, as implemented by Roux [172]. The area under the ROC curve (AUC) and 95% CI was also calculated as a performance metric, and marked on the ROC curves.

The performance when maximising the sensitivity was also analysed, by determining the thresholds required to achieve mean sensitivity levels of 0.50, 0.60, 0.70, 0.80 and 0.90 with each of the SDAs. It was decided that sensitivity levels of above 50% are more acceptable, this has also been recommended by the observational medical outcomes partnership (OMOP) [199]. Therefore the specificity in this case can be compromised to improve the sensitivity when

detecting true signals. The average number of generated signals (the average number of false signals can be determined by multiplying this by the FDR), specificity, FDR and positive predictive value (PPV) were also evaluated at these levels of sensitivity.

7.4.4 Software Package for Signal Detection Analysis

Simulated datasets were generated in SAS version 9.3 [200] (Appendix D provides the SAS code for the simulation model), and the PhViD [201] package for PV in R (Version 3.1.1) was used to perform the signal detection analysis.

To use the PhViD package, simulated data must firstly be organised into a data frame consisting of the following three columns; 1st label of drugs, 2nd label of AEs and 3rd number of spontaneous reports (n_{ij}) of the corresponding couple ADR (Figure 15). Then reports generated are transformed to the elements of a 2x2 contingency table, where it is then possible to calculate the marginal counts ($n_{i\bar{j}}$ and $n_{\bar{i}j}$) which are required for the calculations.

The next stage involves calling the SDA with the appropriate syntax, this is done using the statistic argument set to the decision criterion (e.g., lower 5th percentile or 2.5% quantile), and then choosing a threshold upon the decision criterion requested for analysing the ADR couples, as explained in Table 20. Finally a list of generated signals is produced where the metrics of performance (average number of signals, FDR, sensitivity, specificity and PPV) can be computed.

Figure 15: Screenshot of simulated data with the corresponding 2x2 contingency table for Drug 1 and ADR 1.

1	Drug	ADR	Number of reports	n	nADR	nDrug
2	1	1	41	485398	13059	2848
3	1	2	43	485398	10686	2848
4	1	3	39	485398	10482	2848
5	1	4	39	485398	8674	2848
6	1	5	42	485398	8773	2848
7	1	6	20	485398	3617	2848
8	1	7	20	485398	3251	2848

	Drug 1	**Other Drugs**	
ADR 1	41 (n_{ij})	13018 $(n_{i\bar{j}})$	13059 $(n_{i.})$
Other ADRs	2807 $(n_{\bar{i}j})$	469532 $(n_{\bar{i}\bar{j}})$	472339 $(n_{\bar{i}.})$
	2848 $(n_{.j})$	482550 $(n_{.\bar{j}})$	485398 (n)

- n : Total number of reports in database.
- nADR : Marginal count involving ADR_i
- nDrug : Marginal count involving $Drug_j$

7.5 Simulation study 1 - To investigate the use of SDAs in a AE reporting system considering different threshold values

The objective in simulation study 1 is to firstly explore the use of the standard thresholds ($PRR_{02.5}$> 1, $IC_{02.5}$> 0 and GPS_{05}> 2) and then the use of different threshold values to achieve sensitivity performances above 50%, when also considering a trade-off with the FDR. The thresholds will be displayed in the results section, and recommended thresholds will be detailed in the conclusion.

7.5.1 Simulation procedure

A total of 1000 datasets were simulated. The datasets were representative of similar AE reporting databases including 60 UK marketed drugs and 150 AEs. This

was also based on the databases being of a manageable size to simulate. The 60 drugs were randomly chosen from the full set of 416 drugs held within the PROTECT database (Appendix D, Table 33) to provide information about the type of ADRs reported to the database.

For these 60 drugs, there were on average 150 (Range: 10, 1742) MedDRA preferred term coded ADRs reported per drug in the PROTECT database. An assumption was made that each drug had the possibility to report any of these 150 ADRs (i.e., a maximum 9,000 drug-ADR combinations were possible in each simulated dataset). The drug exposure frequencies (T_j) of the 60 drugs were approximated using data from the HSCIC, where for each of the UK marketed drugs the annual prescriptions were obtained. Then each drug was assigned to one of the four exposure levels in the simulations; 300,000 prescriptions for 5 drugs, 150,000 prescriptions for 10 drugs, 75,000 prescriptions for 15 drugs and 20,000 prescriptions for 30 drugs. However, 13 of the 60 randomly chosen drugs were not centrally marketed in the UK and therefore prescription data could not be obtained. These drug exposure frequencies were randomly assigned an exposure rate between the current ranges then placed into one of the four exposure levels as described above.

The data were generated under the condition that 15% of the drug-ADR combinations were *'true signals'*, albeit with varying signal strength levels by imposing a range of RR_{ij} of 2, 3, 5 or 10 each with equal probability across the 15% of drug-ADR combinations. Since no particular constraint is imposed for the definition of the background incidence I_i [172] and the actual number of

reported cases were not listed in the PROTECT database; half of the ADRs were assigned background incidence (I_i) 1/250, and the other half 1/500 to provide a split distribution of common and less commonly reported events. The reporting probability (pr_{ij}) is assumed to be at most equal to 0.1, as was described in one study [202] which determined the probability of reporting AEs in a national spontaneous reporting databases. The reporting probabilities (pr_{ij}) for the 150 ADRs, were evenly distributed and fixed at 0.1, 0.08, 0.06, 0.04, and 0.02.

7.5.2 Simulation study 1 results

One thousand datasets each with 60 drugs and 150 ADRs were generated. The average number of spontaneous reports (n) over 1000 datasets was 16,733 (standard deviation (SD) = 106). The average number of drug-ADR combinations per dataset was 8,893 (SD = 68) with an average 1261 (SD = 2.9) true signals per dataset.

7.5.2.1 At the Standard thresholds

Table 23 shows that the standard thresholds currently used by the SDAs under investigation (i.e. $PRR_{02.5} > 1$; $IC_{02.5} > 0$; $GPS_{05} > 2$) can lead to large differences in the numbers of signals generated (including the number that correspond to true signals), FDRs, sensitivities and PPVs.

Table 23: Comparison of the three signal detection algorithms for all ADRs, using the standard thresholds that are currently used in practice

Average number of ADRs (SD)	8,893 (68)					
Average number of true associations for ADRs (SD)	1261 (2.9)					
Signal detection algorithm and detection threshold	Average number of signals generated (mean (SD))[ŧ]	Corresponding average number of the true signals detected (mean (SD))[Δ]	FDR (mean (SD))[ח]	Sensitivity (mean (SD))[¥]	Specificity (mean (SD))[¥]	PPV (mean (SD))[ר]
$PRR_{02.5} > 1$	752 (6.8)	646 (6.7)	0.1409 (0.017)	0.512 (0.014)	0.969 (0.006)	0.859 (0.017)
$IC_{02.5} > 0$	602 (2.5)	573 (2.3)	0.0479 (0.014)	0.454 (0.012)	0.995 (0.002)	0.952 (0.014)
$GPS_{05} > 2$	405 (0.2)	405 (0.1)	0.0003 (0.012)	0.321 (0.011)	1.000 (0.0001)	1.000 (0.012)

[ŧ] The average number of signals generated is calculated by the sum of the number of generated signals (true/false) in each dataset divided by the total number of datasets.
[Δ] The average number of true signals detected is calculated by multiplying the PPV by the average number of generated signals.
[ח] False discovery rates (FDRs) are calculated as described in section 7.4.3.1, where the proportion of false discoveries (FDPs) is obtained from each dataset then Exp (FDPs) is the FDR.
[¥] The mean sensitivity and mean specificity are calculated as described in section 7.4.3.2.
[ר] The mean positive predictive value (PPV) is simply the complement of the FDR, which is different to the usual calculation of PPV as presented in clinical diagnostic studies.

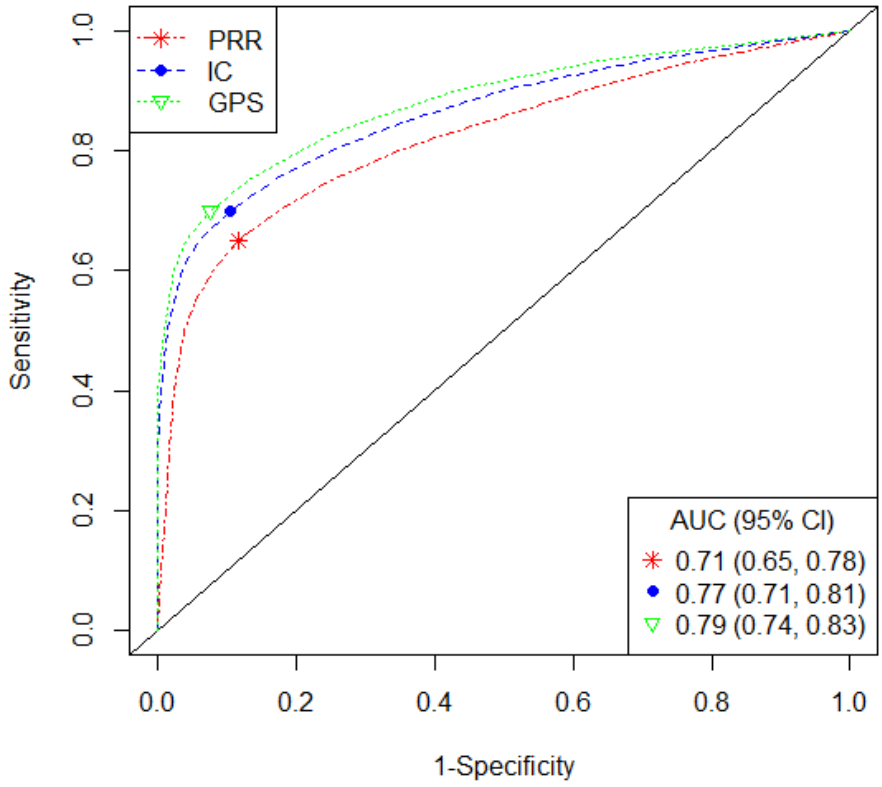
The $PRR_{02.5}$ method generated the most signals, with a mean of 752 (SD=6.8) of which 646 (51%) of these relate to true signals. The Bayesian methods ($IC_{02.5}$ and GPS_{05}) generated fewer signals with 602 (SD=2.5) for the $IC_{02.5}$, of which 573 (45%) of these relate to true signals and 405 (SD=0.2) for the GPS_{05} with 405 (100%) relating to true signals. In particular for the GPS_{05}, the standard threshold

on the criteria currently in use appears to be highly conservative in comparison with the other methods; producing a very high mean specificity (approximately 100%), favouring a very low proportion of false discoveries 0.0003. However this comes at the expense of a poor mean sensitivity of 32%, which would not be acceptable in smaller company databases. Therefore researchers are encouraged to lower the value of the threshold to improve the sensitivity performance, which will be explored in the next part of this study. Since the results also show low SD values, this indicates that the simulation model is consistent.

7.5.2.2 Exploring the effect at different thresholds

The threshold value at a specific sensitivity was determined by changing the probability threshold in small increments (0.025, 0.05, 0.1, 0.2,…, 0.9, 0.95, 0.975), then the ROC curves were constructed. The ROC curves (Figure 16) displayed a pattern of containment (no intersection), which emphasizes that there exist no levels of sensitivity, or specificity for which two methods interchangeably dominate each other. This is especially true for the relationship between the Bayesian approaches ($IC_{02.5}$ and GPS_{05}) and the $PRR_{02.5}$ method, and implies that the Bayesian approaches are better across all levels of sensitivity and specificity in this simulation study. This improved performance by the Bayesian approaches ($IC_{02.5}$ and GPS_{05}) was also indicated by the higher AUC estimates than the $PRR_{02.5}$, though the GPS_{05} achieved the best performance with the AUC = 0.79 (95% CI: 0.74, 0.83) (Figure 16).

Figure 16: Receiver operating characteristic (ROC) curves for each method in simulation study 1.



Finally the performance metrics at the desired sensitivity levels 0.50, 0.60, 0.70, 0.80 and 0.90 are presented in Table 24.

The thresholds and mean specificity, FDR and PPV were also recorded at this sensitivity value. The following include some examples:

- At the sensitivity level of 0.5, the $PRR_{02.5}$, $IC_{02.5}$ and GPS_{05} will result in 743, 602, 405 generated signals on average, specificities of 0.97, 0.99 and 1 and FDRs 0.13, 0.05 and 0. The thresholds required to obtain these performance characteristics, are 1.15, -0.60 and 1.50 respectively, and the PPVs were above 86%.

Table 24: Performance metrics for the three SDAs to achieve the sensitivities with the corresponding threshold in simulation study 1.

$PRR_{02.5}$						$IC_{02.5}$						GPS_{05}					
Thre	Sen*	Sig	Spe	FDR	PPV	Thre	Sen*	Sig	Spe	FDR	PPV	Thre	Sen*	Sig	Spe	FDR	PPV
1	0.51	752	0.97	0.14	0.86	0	0.45	602	1.00	0.05	0.95	2	0.32	405	1.00	0.00	1.00
1.15	**0.50**	743	0.97	0.13	0.87	-0.60	**0.50**	602	0.99	0.05	0.95	1.50	**0.50**	405	1.00	0.00	1.00
0.95	**0.60**	1104	0.90	0.32	0.68	-0.81	**0.60**	750	0.97	0.14	0.86	1.30	**0.60**	641	0.98	0.10	0.90
0.75	**0.70**	1286	0.78	0.41	0.59	-1.03	**0.70**	843	0.88	0.23	0.77	1.10	**0.70**	809	0.90	0.19	0.81
0.55	**0.80**	1483	0.62	0.49	0.51	-1.18	**0.80**	1022	0.76	0.32	0.68	0.90	**0.80**	985	0.79	0.28	0.72
0.35	**0.90**	1688	0.31	0.58	0.42	-1.30	**0.90**	1206	0.50	0.39	0.61	0.70	**0.90**	1093	0.54	0.34	0.66

Thre - Threshold required to achieve the corresponding sensitivity; Sen - mean sensitivity; Sig - average number of signals generated (false signals can be obtained by multiplying this by the FDR); Spe - mean specificity; FDR - false discovery rate; PPV - mean positive predictive value.
Gray shaded area represents the performance metrics when using the standard threshold criteria i.e., $PRR_{02.5} > 1$, $IC_{02.5} > 0$ and $GPS_{05} > 2$.
***since there were 1000 datasets the thresholds required to achieve the sensitivity value are not exact, and therefore were based on achieving the sensitivity value to 2 decimal places.**

- By increasing the sensitivity level to 0.7, we observed an increase in the number of signals generated 1286, 843 and 809, a drop in the specificities to 0.78, 0.88 and 0.90 and increase in FDR to 0.41, 0.23 and 0.19 for the $PRR_{02.5}$, $IC_{02.5}$ and GPS_{05} respectively. There were 527, 194 and 154 false signals for the $PRR_{02.5}$, $IC_{02.5}$ and GPS_{05}. This sensitivity level of 0.7 was achieved by lowering the thresholds further; this also resulted in decreased PPVs by approximately 20% for each method.
- When observing the sensitivity at a level of 0.90; 1688, 1206 and 1093 signals were generated, and the specificities were approximately half-

fold compared to when the sensitivity level was 0.50. The FDRs were all above 0.33 resulting in approximately 979, 470 and 372 false signals for $PRR_{02.5}$, $IC_{02.5}$ and GPS_{05} respectively. The PPVs ranged from 0.42 for the $PRR_{02.5}$, to 0.66 for the GPS_{05} which were the lowest across all levels of sensitivity.

7.5.3 Conclusion

Our results from this simulation study, suggest that the standard thresholds in use for the three SDAs result in large differences in terms of the performance metrics when analyzing AEs within a reporting system consisting of 60 drugs and 150 AEs. The Bayesian methods ($IC_{02.5}$ and GPS_{05}) outperformed the $PRR_{02.5}$ by displaying a lower value of FDR; in particular the GPS_{05} was the lowest. However, the standard threshold used for the GPS_{05} is considered too conservative as was indicated by the poor sensitivity performance of 32%.

When exploring the use of different thresholds for the SDAs, the Bayesian methods ($IC_{02.5}$ and GPS_{05}) were found to be superior to the $PRR_{02.5}$, and generally provided greater specificity when sensitivity was varied at values greater than 50%. The GPS_{05} method provided the best performance with the highest degree of accuracy when signaling true ADRs, as measured by the AUC. However, there was essentially very little difference in the sensitivity-specificity trade-off performance between the two Bayesian methods $IC_{02.5}$ and GPS_{05}, though when considering the trade-off results with the FDR also, the GPS_{05} proved most optimal if sensitivity is required to be above 50%.

Overall the results in simulation study 1 suggest that the GPS_{05} method controls the FDR well and also provides the better trade-off between sensitivity-specificity, although it is recommended that the threshold is adjusted to improve the sensitivity performance. For example we recommended that the GPS_{05} > 1.30 is used, which produced a sensitivity of 60% and provides a relatively small FDR of 10%.

7.6 Simulation study 2 - Detection of Rare Events

The purpose of this simulation study was to investigate the performance of the SDAs when detecting rare signals which are associated with low numbers of AE reports. There are two parts to this investigation to be carried out as explained below in the simulation procedure.

7.6.1 Simulation procedure

Firstly, a total of 1000 datasets were simulated with fixed parameters. The design was similar to simulation study 1 representing the type of data collected in an AE reporting database including 60 drugs and 150 AEs. Although the difference being, that all events were considered to have a background incidence rate $I_i = 1/500$ to consider the AEs as being less commonly reported. The RR_{ij} were imposed to take the values between: 1.2 to 5, again with equal probability. The result of changing these parameters meant that the true signals would have fewer reports on average, and were therefore potentially more difficult to detect. The performance characteristics of the standard thresholds and the use of different thresholds were explored similarly to simulation study 1.

Secondly, 24 individual scenarios, each with 1000 datasets, were simulated by fixing the I_i and RR_{ij} parameters. I_i was set to 1/250, 1/500 or 1/1000, the addition of the I_i= 1/1000 was to consider rare cases of events as is classified by the WHO [6], and the RR_{ij} was set to 1.2, 1.5, 2, 3, 4, 5, 7.5 or 10 respectively. The sensitivity and FDR were compared graphically. In this part of the investigation only the standard threshold criteria for the SDAs were explored (i.e. $PRR_{02.5} > 1$, $IC_{02.5} > 0$ and $GPS_{05} > 2$).

7.6.2 Simulation study 2 results

In simulation study 2, the average number of spontaneous reports (n) over the 1000 datasets was 12,767 (SD = 85). The average number of drug-ADR combinations per dataset was 3,686 (SD = 57) with 545 (SD = 3.6) true signals per dataset, which was less than half that displayed in simulation study 1.

7.6.2.1 At the Standard thresholds

Table 25 shows that the standard thresholds when detecting rare signals, also leads to large differences across the performance metrics. Again the $PRR_{02.5}$ method generated the most signals, with a total 282 on average. The Bayesian methods ($IC_{02.5}$ and GPS_{05}) generated fewer signals 213 and 150, and favoured a lower proportion of false discoveries 0.0012 and 0.0002, this was lower than displayed in simulation study 1. However, the mean sensitivity performances across all methods were below 47%, with the GPS_{05} performing worst with mean sensitivity of only 28%. These estimates of the mean sensitivity were also worse than displayed in simulation study 1 across all the methods, and represents the impediment when detecting rare signals. Therefore in the next section we will

investigate adjustments on the threshold value to try to improve the sensitivity performance.

Table 25: Comparison of the three signal detection algorithms for detecting Rare ADRs, using the standard thresholds that are currently used in practice.

Average number of ADRs (SD)	3,686 (57)	In these simulations $I_i = 1/500$ and RR_{ij} was imposed to take values between 1.2 to 5*				
Average number of true associations for ADRs (SD)	545 (3.6)					
Signal detection algorithm and detection threshold	Average number of signals generated (mean (SD))[ł]	Corresponding average number of the true signals detected (mean (SD))[Δ]	FDR (mean (SD))[Π]	Sensitivity (mean (SD))[¥]	Specificity (mean (SD))[¥]	PPV (mean (SD))[Γ]
$PRR_{02.5} > 1$	282 (6.4)	255 (6.5)	0.0940 (0.016)	0.468 (0.023)	0.946 (0.008)	0.906 (0.016)
$IC_{02.5} > 0$	213 (3.6)	212 (3.6)	0.0012 (0.014)	0.389 (0.021)	0.991 (0.003)	1.000 (0.014)
$GPS_{05} > 2$	150 (0.6)	150 (0.5)	0.0002 (0.006)	0.275 (0.012)	0.998 (0.0002)	1.000 (0.006)

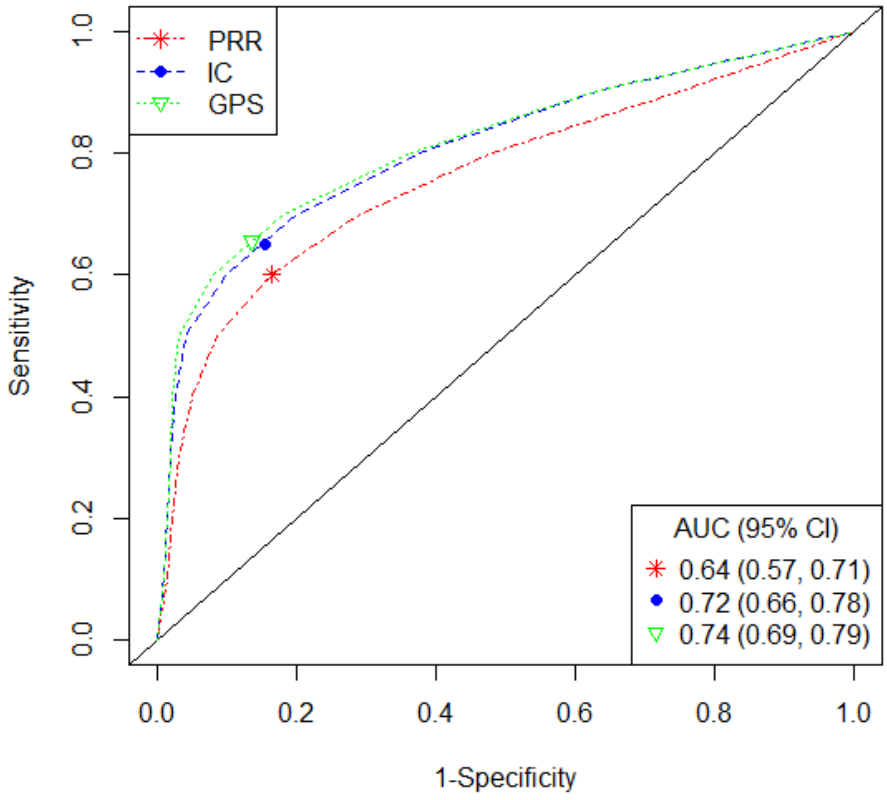
***Please see footnote from Table 23 for description of symbols.**

7.6.2.2 Exploring the effect at different thresholds

Similarly to simulation study 1, the ROC curves for the SDAs imply that the Bayesian approaches are better across all levels of sensitivity and specificity when detecting rare signals. The GPS_{05} achieved the best performance with the AUC = 0.74 (95% CI: 0.69, 0.79) (Figure 17). However, these AUCs estimates were noticeably lower than the AUC estimates obtained in simulation study 1 (Figure

16), which was expected since detecting signals with fewer numbers of reports (rare signals) is potentially more difficult.

Figure 17: Receiver operating characteristic (ROC) curves for each method when detecting rare signals in simulation study 2.



Furthermore, as was investigated in simulation study 1, the performance metrics at sensitivity levels of 0.50, 0.60, 0.70, 0.80 and 0.90 were assessed, and are presented in Table 26.

The mean specificity, FDR and PPV were also recorded at this sensitivity. Below are some examples:

- At the sensitivity level of approximately 0.5, the $PRR_{02.5}$, $IC_{02.5}$ and GPS_{05} will result in a higher number of signals generated with 341, 329 and 291 and lower specificity values with 0.92, 0.97 and 0.98. The FDRs increased

from those displayed when using the standard threshold to 0.11, 0.02 and 0, which was also less than the FDRs displayed in simulation study 1. The number of false signals included was 38 and 7 for the $PRR_{02.5}$ and $IC_{02.5}$, and 0 for the GPS_{05}. The thresholds required at this sensitivity level were 0.96, -0.63 and 1.24 respectively.

Table 26: Performance metrics for the three SDAs to achieve the sensitivities with the corresponding threshold when detecting rare signals in simulation study 2.

$PRR_{02.5}$						$IC_{02.5}$						GPS_{05}					
Thre	Sen*	Sig	Spe	FDR	PPV	Thr	Sen*	Sig	Spe	FDR	PPV	Thre	Sen*	Sig	Spe	FDR	PPV
1	0.47	282	0.95	0.09	0.91	0	0.39	213	0.99	0.00	1.00	2	0.28	150	1.00	0.00	1.00
0.96	**0.50**	341	0.92	0.11	0.89	-0.63	**0.50**	329	0.97	0.02	0.98	1.24	**0.50**	291	0.98	0.00	1.00
0.89	**0.60**	514	0.82	0.24	0.76	-0.86	**0.60**	396	0.90	0.09	0.91	1.15	**0.60**	338	0.92	0.07	0.93
0.68	**0.70**	683	0.69	0.38	0.62	-1.09	**0.70**	422	0.79	0.18	0.82	1.00	**0.70**	419	0.81	0.14	0.86
0.51	**0.80**	804	0.48	0.46	0.54	-1.24	**0.80**	471	0.61	0.29	0.71	0.84	**0.80**	486	0.61	0.23	0.77
0.36	**0.90**	1114	0.21	0.54	0.46	-1.32	**0.90**	547	0.32	0.35	0.65	0.62	**0.90**	532	0.32	0.27	0.33

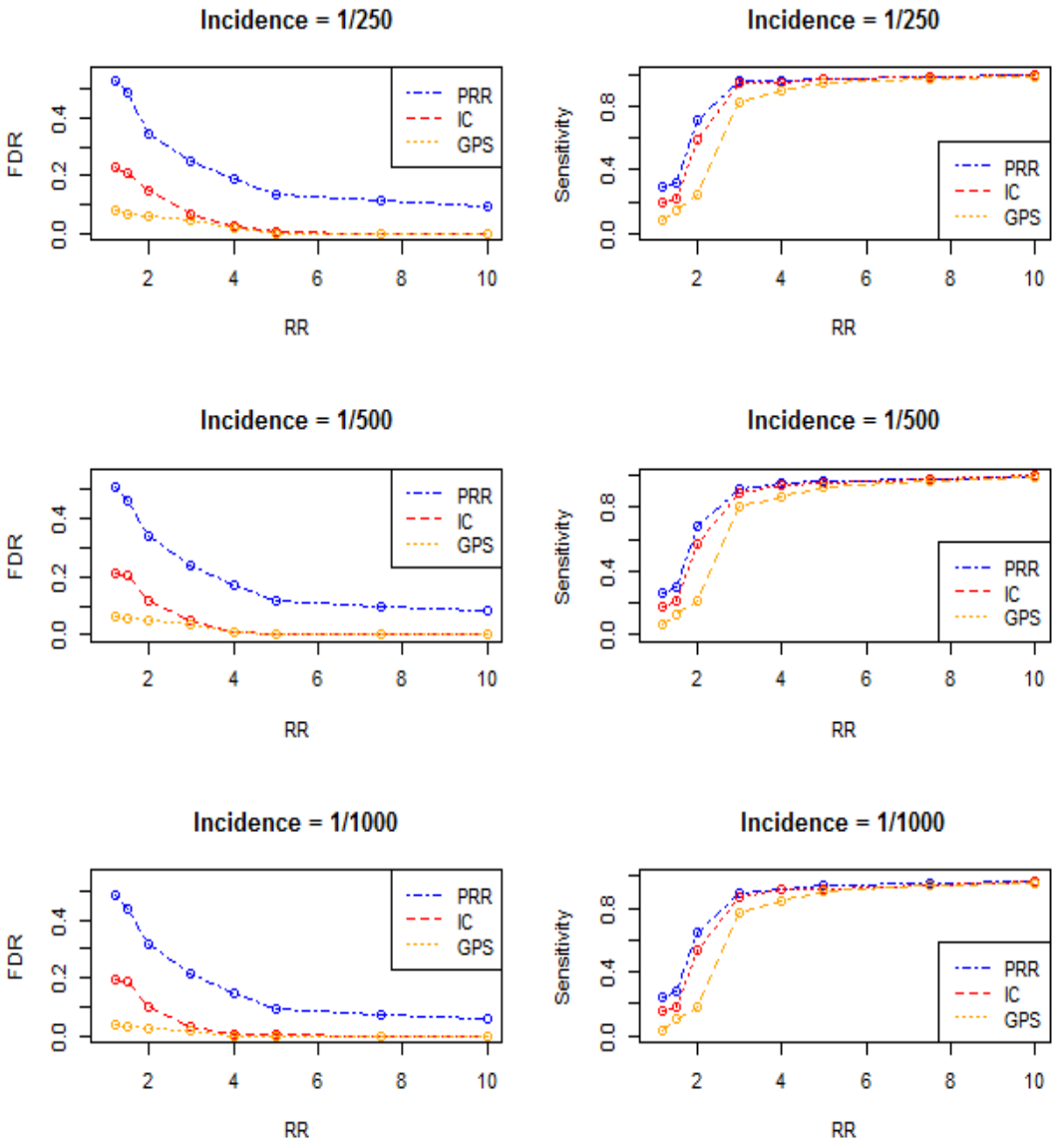
***Please see footnote from Table 24 for description of column names.**

- The specificities for achieving a sensitivity level of approximately 0.7 were decreased to 0.69, 0.79 and 0.81 respectively, and the FDRs increased further to 0.38, 0.18 and 0.14, resulting in considerably higher numbers of false signals 260, 76 and 59. The thresholds were adjusted to a lower value, and as a result the PPVs decreased.
- Overall for the different levels of sensitivity, the specificity and PPV decreased when detecting rare signals compared to simulation study 1. However, there was a minor improvement in the FDR performance across the levels of sensitivity.

7.6.2.3 Simulated scenarios to explore performance of SDAs for detecting signals of rare events

The results from the 24 simulated scenarios are presented in Figure 18 and see appendix D, Table 34 for full numerical results.

Figure 18: Simulation scenario results when detecting rare events.



***The points of the curve represent the RR_{ij} values investigated in the simulated scenarios. Values in-between are just extrapolations.**

When I_i = 1/250 and the RR_{ij} increases, the sensitivity also increases from 0.289 when RR_{ij} is 1.2, to 0.998 when RR_{ij} is 10 for the $PRR_{02.5}$ method; 0.197 to

0.998 for the $IC_{02.5}$, and 0.080 to 0.985 for the GPS_{05}. The FDR is highest when the RR_{ij} is 1.2, and was lowest when RR_{ij} set at 10. This was the case across all methods. The FDR is particularly high for the $PRR_{02.5}$ when compared with the Bayesian methods; the GPS_{05} however outperforms the $IC_{02.5}$ by small margins across all RR_{ij} values. For example when RR_{ij} is set to 1.2, the FDR for the $PRR_{02.5}$ is 0.528, 0.229 for the $IC_{02.5}$ and 0.081 for the GPS_{05}. At a RR_{ij} of 10, the FDR for the $PRR_{02.5}$ is 0.092, 0.0009 for the $IC_{02.5}$ and 0.0003 for the GPS_{05}.

For I_i = 1/500, a similar pattern is observed, although the estimated values for sensitivity were lower. Moreover there was an observed improvement in the FDR as compared to when I_i = 1/250. Again, the $PRR_{02.5}$ provided the best performance of sensitivity between 0.27 and 1.0 but at the expense of increased FDRs between 0.08 and 0.51. The GPS_{05} method produced the lowest FDR which was less than 0.07 across all scenarios.

Finally, with I_i = 1/1000 the sensitivity decreased slightly across all three methods, though the pattern of increasing sensitivity as the RR_{ij} increased was similar. The FDRs were lowest across all scenarios for this incidence rate; again the $PRR_{02.5}$ method produced the highest FDR, and the GPS_{05} achieved the lowest FDR.

7.6.3 Conclusion

The first part of this simulation study shows the performance of methods for detecting rare signals with $RR_{ij} \leq 5$ and a I_i = 1/500. As was the case in simulation study 1, the $IC_{02.5}$ and GPS_{05} were superior to the $PRR_{02.5}$, providing greater specificity at levels of sensitivity greater than 50%. Again the GPS_{05}

method provided the best performance with the highest degree of accuracy. However, in general rare signals were detected with less accuracy as indicated with the lower AUC than shown in simulation study 1. The Bayesian methods ($IC_{02.5}$ and GPS_{05}) also outperformed the $PRR_{02.5}$ by displaying the lowest FDR at all levels of sensitivity above 50%, with the GPS_{05} producing the lowest FDR.

The methods were evaluated more extensively by assessing their performance on 24 simulated scenarios using the standard threshold criteria. For signals with high numbers of reports with RR_{ij} above 5, the $IC_{02.5}$ and GPS_{05} provide the best FDR performance, and the sensitivity was similar across all methods, above 80%. For signals with a low number of reports with RR_{ij} below 4, the GPS_{05} had the lowest FDR, although the GPS_{05} also produced the lowest sensitivity. The sensitivity performance was similar with the $PRR_{02.5}$ and $IC_{02.5}$ methods. Therefore considering the trade-off between the FDR and sensitivity performance, the $IC_{02.5}$ proved to be the method of best choice. However, the GPS_{05} with its standard threshold criteria (i.e., $GPS_{05} > 2$) is regarded as conservative, and hence changing the threshold would improve the sensitivity performance. For example using the $GPS_{05} > 1$ will result in an improved sensitivity of 70% and FDR of 14% when detecting rare signals, as shown in section 7.6.2.2.

7.7 Simulation study 3 - Exploring performance within small databases reflective of Clinical Trial Unit Databases

The aim in this simulation study is to investigate the performance of SDAs in smaller scale systems, by simulating different scenarios to mirror the type of harms data that might be collected in CTU databases.

7.7.1 Simulation procedure

As part of the survey in chapter 5 information was collected on the number of drugs trialed, and events reported in CTUs with central databases (section 5.4.1.2). The results from six CTUs with a central database are provided in Table 27. The majority of these CTUs involved cancer trials where signal detection methods would more likely have been useful. Therefore, the simulated scenarios are not necessarily reflective of the wider network of CTUs.

Table 27: The specific sizes of harms databases from five clinical trial units.

Number of Drugs $_j$	Number of Events $_i$ (AEs or SAEs)
12	100 SAEs
20	200 AEs
20	Not reported
40	200 AEs
40	33 SAEs
34	140 AEs

Using the results from Table 27, five scenarios of different sized clinical trial database were chosen. The range of sizes explored were: (1) 60 drugs and 150 events, to reflect a large database as explored previously (2) 40 drugs and 120 events, (3) 30 drugs and 100 events, (4) 20 drugs and 80 events, and (5) 10 drugs with 50 events. For simulating the clinical trial databases the same model was

used as in simulation study 1 and 2; although with each of these five scenarios I_i = 1/250, 1/500 or 1/1000 was explored separately, resulting in the total of 15 simulated scenarios, with 1000 simulated datasets in each scenario. The SDAs were assessed again by comparing sensitivity and FDR, and all SDAs were assessed using the standard threshold criteria only (i.e. $PRR_{02.5} > 1$, $IC_{02.5} > 0$ and $GPS_{05} > 2$).

7.7.2 Simulation study 3 results

The simulation results including sensitivity and FDR for the different scenarios when considering the database size are presented in Figure 19 and see appendix D, Table 35 for full numerical results.

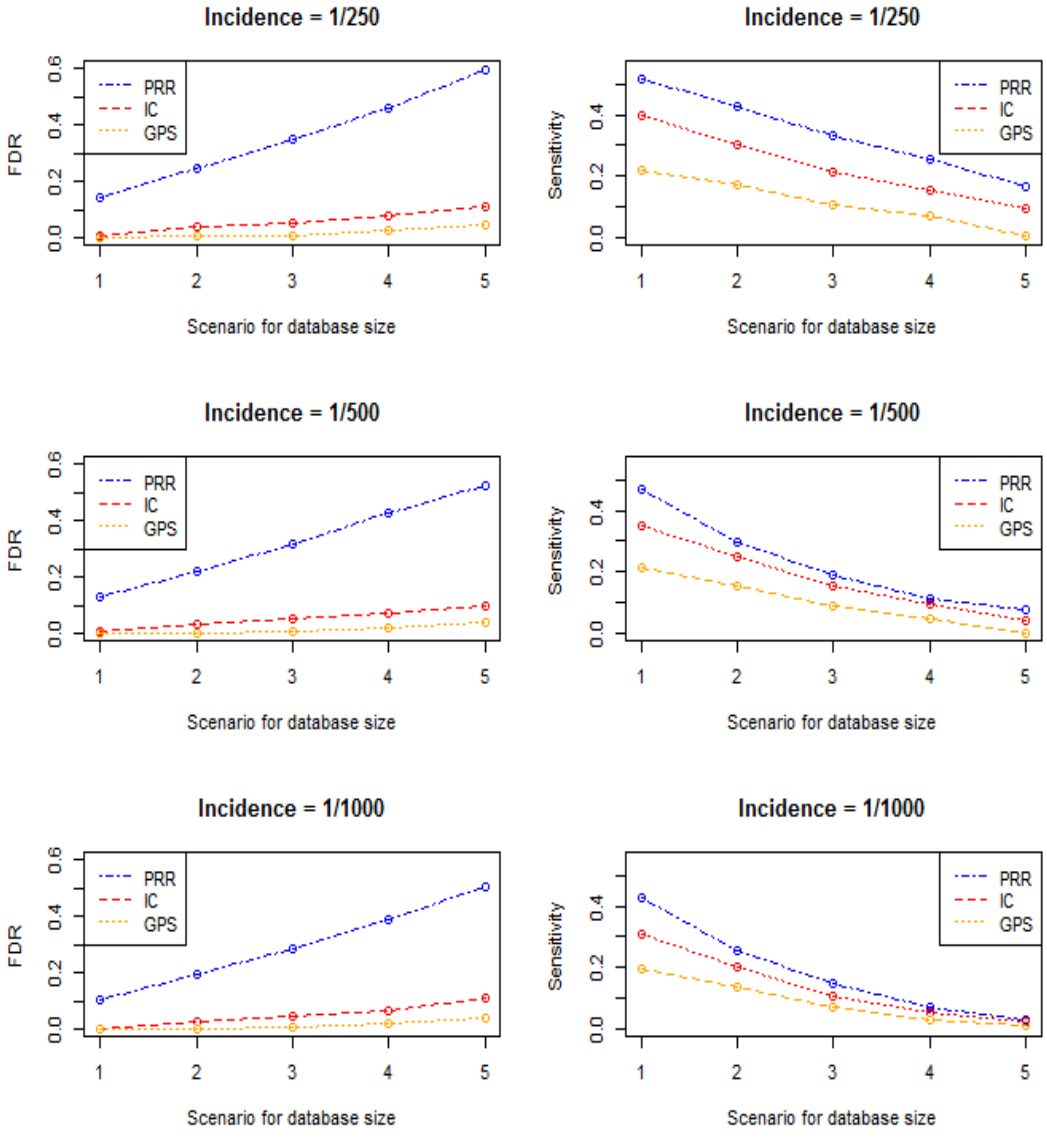
When I_i = 1/250 the sensitivity decreased as the database reduced in size (See appendix D, Table 35). The $PRR_{02.5}$ method proved the best method with the sensitivity ranging from 0.51 to 0.17 for scenarios (1) - (5) respectively. The $PRR_{02.5}$ displayed the highest FDRs of above 0.14 across all scenarios. The GPS_{05} method produced the lowest FDR across all scenarios, with FDRs below 0.05. When I_i = 1/500, the trends were similar, although the sensitivity was further decreased for each method, and the FDRs are generally better for all methods. When $I_i = 1/1000$, the same pattern is displayed with decreased sensitivity and improved FDRs.

7.7.3 Conclusion

When reducing the database size of that similar to CTU databases, it was shown that the sensitivity of all methods reduced considerably. For example, for databases containing only 20 drugs (scenario (4)), the sensitivity was

approximately below 20% for all methods and across all levels of I_i. The FDR across all the scenarios was increased for the $PRR_{02.5}$ compared with the Bayesian methods ($IC_{02.5}$ and GPS_{05}) which both produced similar FDRs. As was suggested in simulation study 2 the $IC_{02.5}$ would be the best method to consider, though the GPS_{05} with a different threshold criteria would outperform the $IC_{02.5}$.

Figure 19: Simulated scenario results to assess the FDR and sensitivity performances at different incidences.



***The points of the curve represent each scenario for the database size i.e., with specific number of drugs and events as explained in section 7.1.1. Values in-between are just extrapolations.**

Considering these poor performances with regard to sensitivity for databases of reduced size, a different threshold should be used. The results from the sensitivity-specificity trade-offs in simulation studies 1 and 2 with sensitivity values above 50%, should be considered in practice. However it is still important to balance the sensitivity with the FDR, to optimize the performance when detecting true signals as much as possible.

7.8 Discussion

These simulation studies have provided a systematic assessment of the performance of commonly used SDAs.

In simulation study 1 for each of the SDAs, different thresholds were explored to assess the balance between FDR and sensitivity when detecting signals in an AE reporting system containing 60 drugs and 150 AEs. The results from this study have shown that the $PRR_{02.5}$ was not able to control the FDR, and for achieving values of sensitivity above 50% lower threshold values were required than the standard thresholds currently in use. The GPS_{05} performed better than the $IC_{02.5}$ and $PRR_{02.5}$ methods, as displayed in the ROC curves and by the AUC values. For AE reporting systems of similar size, it is recommended that a lower threshold of the GPS_{05} is used, e.g., $GPS_{05} > 1.30$ to improve the sensitivity performance to approximately 60%, but also control the FDR to 10%.

Similarly in simulation study 2, the use of different thresholds on the SDAs was explored to assess the balance between FDR and sensitivity when detecting rare signals. The results again suggested that the $PRR_{02.5}$ produce the highest FDRs, and that the GPS_{05} was the better SDA for achieving high values of sensitivity

above 50% whilst also controlling the FDR, which was achieved by lowering the threshold. These performances however were worse than displayed in simulation study 1. The methods were also assessed at the standard threshold criteria in 24 different simulated scenarios considering variation on the I_i and RR_{ij} parameters. The results showed that the $PRR_{02.5}$ provides the highest sensitivity as the RR_{ij} increases; however the $PRR_{02.5}$ also provided the highest FDRs of the three methods when detecting rare signals. On the other hand the $IC_{02.5}$ and GPS_{05} control the FDR much better, although their sensitivity performance was relatively poor, particularly the $GPS_{05.}$ It is therefore recommended that a threshold of $GPS_{05} > 1$ could be used to improve the sensitivity performance to approximately 70% and control the FDR to a level of 14%.

Finally in simulation study 3, the SDAs were assessed in five scenarios considering different database sizes, to mirror current CTU systems. The results from this study have shown that the sensitivity decreases as the size of the database decreases, and the SDAs with their standard threshold criteria are only able to detect few signals for small or sparse data similar to harms data contained in CTUs. Specifically it has been shown that when the database contains fewer than 20 trials and 80 different AEs, that SDAs become unreliable signal generating tools, with poor sensitivity below 10% at times. Therefore traditional signal detection methods (i.e., cases and cases series reviews) should be used on databases of a smaller scale, as discussed in chapter 6.

Most of the safety signal detection methods were developed in the last two decades, and there have been some attempts to compare the performance of some of these methods as highlighted in the literature review. For example, a recent study [165] assessed the performance of the PRR using real spontaneous reported data from the Eudravigilance database; another study [169] assessed the performance of the GPS on a pharmaceutical company AE database. However, it is very challenging to assess the performance of the methods in terms of sensitivity and FDR using real databases when the status of true signals is unknown. In an earlier study [172], 10 methods published before 2000 were compared by simulating the incidence reporting process based on Poisson distribution. However the FDR was not estimated in this study as the methods had not been developed in a multiple hypothesis testing framework, and they did not investigate the use of different threshold values.

The simulation model used in this study has considered real prescription data from the HSCIC, and ADR reports were obtained from the EMA PROTECT database to formulate accurate and reliable parameters during the data generation process. However, the simulated data only represents fictional drug classes and outcome types, and therefore no clinical interpretations should be drawn from the data.

It was not possible to examine the onset of signals relative to the time point at which an ADR is confirmed. Therefore a comparative assessment between SDAs compared with more traditional methods could not be made. Due to the constraint on time for simulating the data, the use of different thresholds could

not be investigated in the simulated scenarios in simulation study 2 and simulation study 3. This could have enhanced the performance, and improved the sensitivity when detecting rare events and signals in CTUs; and needs to be carried out in future work. Furthermore the assessments of the SDAs on the five different scenarios representing different database sizes in CTUs were restricted to the number of drugs and events in each scenario. However, as suggested in Table 27 there may be a high number of drugs with very few AEs (or the opposite) which might produce different results, this would need to be explored in future work for consistency also.

More recent signal detection methods have been developed including multivariate modeling techniques [192] and the likelihood ratio test [203] which now enable adjustments for potential confounding factors. Confounding has been investigated primarily in the context of poly-pharmacy, wherein a true association of an AE with one drug may bias its estimated association with another drug when the two drugs tend to be prescribed and reported together [183, 204, 205]. These methods need to be researched further to understand their full potential in the context of signal detection analysis.

For clinical trials data, more traditional statistical tests such as Pearson's chi-square test, Fisher's exact test, and the chi-squared test for rates comparison are often used for flagging safety signals. These methods, however, do not control for multiplicity. Multiple testing is highly important when making assessments about AEs, as stated in the ICH E9 good clinical practice guideline [206]: *"when hypothesis tests are used to evaluate safety data, statistical*

adjustments for multiplicity to quantify the type I error are appropriate". In 2004, Mehrotra and Heyse [207] developed a procedure to control FDR based on Benjamini and Hochbergs (BH) procedure [197] considering the hierarchical structure of MedDRA coding, namely, AE preferred terms (PTs) are grouped into body systems, referred to as system organ classes (SOCs). This procedure adjusts FDR at both SOC level and PT level, and hence, it is often referred to as the double FDR. Following Mehrotra and Heyse, Berry and Berry [184] introduced the Bayesian hierarchical mixture model (BHMM) to detect safety signals (Section 6.3.4), which has the same assumption as the double FDR (DFDR) method. It is assumed that the probability that a drug has caused a type of AE is greater if its rate is elevated for multiple AE PTs within the same SOC, than if the AE PTs with elevated rates belonged to different SOCs. Most recently, Mehrotra and Adewale [208] developed a newer DFDR adjustment approach and demonstrated that it has better performance in terms of FDR and sensitivity. This method needs to be researched further in clinical trials, and there is a demand for software developments to encourage its use.

This study has shown that the two Bayesian ($IC_{02.5}$ and GPS_{05}) methods, particularly the GPS_{05} when using a lower threshold than the standard threshold criteria performs well when considering the sensitivity and the FDR. As shown in chapter 5, SDAs do not appear to be used currently in CTUs, and this simulation study suggests that the SDA methods that have been explored could be particularly unreliable on small datasets. However, the EMAs guideline on good pharmacovigilance practices (risk management) states that *"signal detection is*

an important element in identifying new risks for all products, and should be used as part of a pharmacovigilance tool-kit".

It is also important to remember that spontaneously reported data comes with a number of inherent limitations, and therefore the danger of over-interpreting SDA outcomes has been well highlighted in the past [209]. Further efforts are therefore needed to improve access to other sources of data from clinical trials and observational data so that adverse effects can be evaluated in a more comprehensive and unbiased manner. The Eudravigilance clinical trials module (EVCTM) from 2004 is designed to receive reports on SUSARs that occur in clinical trials, and data can sometimes be accessed by sponsors of clinical trials to inform on the DSURs or for use of traditional signal detection methods like aggregate analysis. Other ongoing initiatives like the exploring and understanding-adverse drug reactions (EU-ADR) [210] project, the innovation in medical evidence development and surveillance [211] program and the pilot project Mini-Sentinel [212] sponsored by the FDA have developed electronic systems which have been setup with the aim to promote the use of observational data to complement existing methods of safety surveillance. However, public sector access to some of these systems is not possible, and data requests can often be very costly.

Finally it is important to remember that SDAs serve as screening tools to identify possible safety signals for further investigation. Safety scientists need to further evaluate the identified possible signals using medical rationale and additional

information such as biological plausibility, outcome of the event, severity and seriousness of the event, and other concomitant medications used.

Chapter 8: Conclusions and further work

This thesis details research into some of the challenges that stem from the reporting, conduct, analysis and interpretation of harms in clinical research.

8.1 Overview

The systematic review in chapter 2 has shown that the current standards of reporting harms in RCTs, after the release of the CONSORT-harms still remains poor and inadequate [55]. Readers of RCT publications should be able to balance the trade-offs between the benefits and harms of interventions [213], however this review highlighted inconsistencies and at times inadequate reporting for all 10 CONSORT-harms recommendations across seven systematic reviews, which included RCTs of a diverse range of clinical areas and conditions. The review highlights the need for wider adoption of the CONSORT-harms extension by journals. This research was published as a review in the British medical journal (BMJ) open, and was added to the list of important publications by the Cochrane adverse effect methods group (AEMG).

The debate around open access to clinical trials data continues, with ongoing developments for better data transparency of clinical trial results. The value of unpublished data and results held within CSRs has proven highly influential in the past, when evaluating both the safety and efficacy of marketed drugs [17, 83]. The case study in chapter 3 which includes a representative sample of five published RCTs for the obesity drug olistat, has shown that the CSRs provide

more harms data, and were generally more transparent in their findings than the journal publications. They also detailed more about the design, conduct, and analysis of the trial which help facilitate the assessment of risk of bias in an evidence synthesis. This study is currently under review for publication.

The unpredictable and diverse nature of harms substantially increases the complexity of the study designs and data sources used in a systematic review. When searching and identifying relevant data sources it is important to consider a structured approach, so that harms can be evaluated in a comprehensive, unbiased manner (Chapter 4). Due to past disasters in drug safety like the thalidomide tragedy, PV has resulted in the development of systems that collect individual case histories of ADRs to improve the safety profile of medicines [109]. These systems can support a more comprehensive resource for harms data held within health databases, which are now frequently being used in hypothesis-strengthening observational studies to assess the risks of harms [140].

At present there is a lack of empirical evidence discussing the methods and procedures used in the trial safety monitoring within UK clinical trial units (CTUs) [142]. Therefore it was important to investigate this further by conducting a national survey to communicate with the CTUs. The survey (chapter 5) has shown that very few CTUs database harms centrally, and it was identified that a diverse range of data sources external to the trial are being used during the trial monitoring. This included not only published literature but also observational data sources like health databases and spontaneous reports. These data were

used when monitoring ongoing trials, preparing safety documents like the develop safety update report and to support expedited reporting to the trial sponsor, regulatory authorities and research ethics committee.

Over the past forty years SRSs have often been at the forefront for detecting delayed, uncommon and rare harms [52, 140]. Since 2004 it is a mandatory requirement now for SUSARs from clinical trials to be submitted to the Eudravigilance clinical trial module, which can also be used for drug safety surveillance purposes. Due to the accumulating number of spontaneous and SUSAR reports, data standards now make it possible to use signal detection algorithms (SDAs) to systematically explore safety data and generate hypotheses (chapter 6). Unlike traditional signal detection, SDAs can detect drug-drug interactions or drug-related syndromes which otherwise may not be detected. However, the use of SDAs in clinical trial settings has not been investigated in detail; therefore a simulation study was needed to explore this and other performance characteristics.

The simulation study in chapter 7 has explored the use of SDAs, and suggests that some are more suitable to use than others. The study investigated the performance of three SDAs across different simulation studies with aims to assess the performance in an AE reporting database of fixed size (60 drugs and 150 AEs), an AE reporting database including rare signals and harms databases similar to those in CTUs. Of the three SDAs the Bayesian gamma Poisson shrinker (GPS_{05}) method produced the lowest number of false signals across all scenarios, as measured with the false discovery rate (FDR). The GPS_{05} was also found to be

very conservative in its approach, which is why an investigation was carried out to explore the use of different thresholds. This was mainly to try and maximize the sensitivity performance of detecting true signals, but also control the FDR. When maximizing the sensitivity above 50%, the GPS_{05} outperformed the other SDAs with lower FDR values. However it was suggested that these SDAs are unsuitable and potentially unstable when used on CTU databases of smaller size. Mainly due to their poor performance on the sensitivity, this was shown to be below 20% at times.

8.2 Limitations

The systematic review (chapter 2) did not assess changes in reporting over time, to observe for any improvements since the release of the CONSORT-harms extension in 2004. Since some of the included studies contained trials reported prior to the publication of the CONSORT-harms guideline (Pre-CONSORT), it may have been beneficial to provide a Pre vs. Post-CONSORT in a meta-analysis comparison to observe for levels of improvement after the release of the guideline. In fact this has been assessed for the standard CONSORT guideline [59], and they generally found vast improvements for each item, but some were still found to be lagging. Due to the limited number of studies published that have systematically reviewed the standards for reporting harms in RCTs using the CONSORT-harms as a benchmark, we were only able to obtained seven studies which were of varying clinical areas and conditions. This made it difficult to assess the heterogeneity across studies, and to determine which recommendations preformed the worst.

In the case study (chapter 3) we were able to obtain access to CSRs from Roche; however we requested access to 31 CSRs and only received 5. So the analysis is based only on a subset of representative trials, and therefore should not be considered as clinical evidence. Since orlistat is also centrally licensed by the EMA requests were made for access to the CSRs, particularly for those trials that pre-dated Roche's policy act. Though the ongoing legal proceedings from 2013 onwards, meant that the EMA were unable to provide any CSRs. However their policy has now been re-instated, and therefore it may now be possible to obtain more CSRs from them [89]. Furthermore, no clinical assessments of causality or relatedness for missing AEs and SAEs in the journal publications were made, though the protocol did mention that only related events were to be reported in the publication, but this was not accessed in detail.

The survey provided valuable insight into some of the current practices involved in UKCRC registered CTUs (chapter 5). However 51% of the CTUs did not respond to the survey, this may have been affected by the limited collection time period which was restricted to approximately three months. In addition, many of the open-ended questions in the survey where the participants were asked to elaborate and provide further comments often lacked quality and quantity. For example when using an external harms data source like CPRD data, very few CTUs provided extensive detail on how and why they used the data. We were unable to follow-up on any outstanding queries, to try and determine more detailed responses from the participants. Therefore the results only represent a subset of the responses.

The design and parameters used in the simulation model in chapter 7 was dependent upon summary data obtained from the SmPCs, prescriptions data from the HSCIC and data obtained from the survey in chapter 5. A real data set would have improved this simulation study, and enabled a more expansive detailed assessment when detecting real life safety signals. Moreover, we were unable to research other signal detection methods like the multivariate logistic-regression modeling technique [192], which allow adjustments for potential confounding factors during the analysis of drug-event relationships; and the Bayesian hierarchical mixture modeling method for detecting signals from clinical trials data.

8.3 Integration with current research

Over the past 15 years there has been an accumulation of research demonstrating the existence of poor and inadequate reporting for harms in RCTs [9, 10, 24, 214]. As a result of these findings in 2004 the CONSORT group developed their harms extension, to help improve upon the standards of reporting harms in RCTs. Our review was the first to empirically assess the standards of harms reporting using the CONSORT-harms as a benchmark. The review supports findings from previous studies that the reporting of harms is still inconsistent and inadequate, and that greater emphasis should be in place for wider adoption and full adherence of reporting guidelines to help improve these standards.

More recently researchers have discovered potentially new and more comprehensive sources for information on clinical trials results, including CSRs.

The information contained within CSRs has proved vital for evaluating both the efficacy [84] and safety [85] of clinical interventions, with some of the evidence from journal publications questioned, and even overturned by findings from unpublished information reported in the CSR [86]. The case study has carried out an extensive assessment of the harms reporting in CSRs against the journal publication for a sample of orlistat trials. This study supports previous findings about CSRs, that they should be considered in any evidence synthesis of clinical trial results, and that researchers should not just rely on the findings from journal publications and systematic reviews of RCTs when assessing harms.

The value of signal detection methodologies has been widely recognized over the past decade. Although past studies [165, 167, 173, 174] have reported that the performance characteristics of SDAs are not well understood, and that there exists a lack of guidelines and gold standards when using them. The aim of the simulation study was to investigate the performance of SDAs in three different simulation studies that have not been researched previously in detail. In particular, the performances of the SDAs when applied to simulated data designed to mirror CTU harms databases, was investigated. The parameters and design of the simulation model were informed from data collected from the survey.

8.4 Recommendations for Researchers

Full adoption of the CONSORT-harms by journal editors is imperative to improve the standards for reporting harms [55]. Peer reviewers should also be properly instructed on how to assess RCTs with adherence to the reporting guidelines

accordingly. The Equator network [107] is an international initiative that seeks to improve the reliability and value of published health research literature, by promoting transparent and accurate reporting and wider use of robust reporting guidelines like the CONSORT-harms extension. The Cochrane AEMG [46] have also developed systematic review methods to address the issues of imbalanced reporting between harms and benefits in RCTs [47], which should be addressed when conducting reviews of harms. Also the PRISMA harms statement is currently under development. This statement aims to develop a checklist of items to guide researchers when conducting systematic reviews and performing meta-analysis on harms.

Open access to clinical trial results and data will undoubtedly continue to improve, with the various stakeholders including funders, academics, industry, publishers and regulators all supporting the move towards greater transparency. It is also important for the continued registration of clinical trials even if the outcome of the trial is unpublished. In the past it has been suggested that approximately 50% of trials results are unpublished and therefore hidden, access to the data from abandoned trials is equally as important as published trials [215].

Harms data is archived and collected individually by trials within CTUs across the UK, although some have implemented central systems. The survey suggests that CTUs with a central reporting system experience certain benefits including; a better coverage of trials and tracking of AEs as they are stored in the same way and easier to compare workloads for future PV and useful for reporting in

development safety update reports and periodic safety update reports [142]. There are also obvious needs to improve access to existing harms data from CTUs in a more coherent and systematic approach, to allow for larger-scale drug safety monitoring. As discussed, it is now a requirement for any SUSAR to be reported to the MHRA and the EMA EudraVigilance clinical trials module, and these reports can be accessed through the EudraViglance gateway, but access by research organizations in public sector is still very limited. Therefore it is important to improve access at affordable costs to systems like the EudraVigilance so that CTUs can learn from each other to move forward.

For identifying safety signals of AEs from reported reactions, SDAs are increasingly being used to supplement the traditional expert review of reports and to analyze the large volume of accumulated data more rapidly. Though there is a lack of gold standards for applying SDAs in practice, and they should only be used as hypothesis generating tools and not hypothesis testing purposes [216]. The current SDAs (PRR, IC and GPS) have in the past reported many performance related issues, from their failure to control the number of false discoveries, the uncertainty of appropriate thresholds that should be used in practice and their performance on small and/or sparse datasets [174]. The recent development of the methods based in a multiple hypothesis testing framework now enable SDAs to control the number of false discoveries by providing an estimate of the FDR at any threshold. In the simulation study we provided an extensive evaluation when using different thresholds to compare the sensitivity performance and FDR. It was recommended that the GPS_{05}

method should be used with a lower threshold to maximize the sensitivity performance above 50%, and also provide an optimal FDR. However these SDAs did not appear to be suitable when applied to the simulated scenarios designed to mirror CTU databases, as indicated by the poor sensitivity performance. Therefore it is recommended that traditional PV methods like case and case series reviews are used in smaller databases for a higher demand on sensitivity for improved detection of safety signals.

8.5 Further work

The systematic review found that the reporting of harms was poor and inadequate even after the release of the CONSORT harms extension. Although this was only assessed using a small cohort of seven published reviews at the time, it is recommended that the review is updated in the future. There also needs to be some guidance provided to reviewers conducting similar studies using the CONSORT-harms checklist. The risk of bias amendment in this study provides some important recommendations when conducting similar reviews, but other considerations may include guidance on using appropriate search criteria for locating the trial reports and how reporting over time could be assessed.

Regression modeling with time of publication included in the model could be used to determine any improvements of reporting over time in reviews. This could be encouraged by asking reviewers to separate RCT reports by year of publication, which would then allow for regression analysis to be carried out. Alternatively, reporting over time could be assessed by taking the median time

points for collecting the trial reports from each study, then in ascending time order show the proportions of reporting in the forest plots.

We recommend that this review is updated in the future, where a more detailed assessment can be undertaken of the impact of the CONSORT harms statement over time. A similar approach was undertaken to assess the uptake of the standard CONSORT statement [60, 217], and this study found that certain items were still lagging post-CONSORT. This is likely to be similar with the CONSORT-harms items.

In the past, CSRs have provided more accurate harms information on the design, conduct and analysis in a clinical trial. For example a recent study [218] to investigate and describe the potential benefits and harms of Tamiflu by reviewing all CSRs of RCTs, the study found significant evidence of increased risks of nausea, vomiting, headaches and renal and psychiatric syndromes. It is anticipated that this data will be compared with the journal publication in a separate study in the future by the same authors. There is also the potential for more information on harms being unveiled by exploring the use other formats of clinical trial results [105]. The information from case report forms (CRFs) could also be useful in an evidence synthesis of harms along with the information obtained within the CSR. The CSRs for the orlistat case study in this thesis removed all CRFs, due to the patient confidential information contained within. A sample of the CRF was provided, and it is easy to see from this the potential value of the additional harms data that could be obtained on each patient individually. A sensitivity analysis considering each of the key points:

relatedness, causality, severity grading and attribution should also be performed in the meta-analysis. Also multiple testing between events and other statistical methods should be considered when handling rare events.

Accessing CSRs can be difficult as found in our case study and has also been exemplified in past studies. However it has been shown, the extent of missing information (whether efficacy or harms) from journal publications, does support the use of CSRs in evidence synthesis. Though, reviewing CSRs can be difficult, as they are extremely lengthy documents and therefore represent a considerable challenge to researchers. Alternative to CSRs, registry reports can often be accessed instantly through a clinical trial results database with the trial ID, and they have occasionally found additional information on harms. However, recent studies suggest that registry reports have also been found to be unreliable [81] with missing information [80]. There is a need to develop tools and methodological approaches that will reduce the workload and still allow researchers to use CSRs in an accurate and efficient manner.

Many unknowns still remain about the current safety monitoring practices involved in CTUs, and possibly how improvements could be made. There are still a number of outstanding questions left unanswered from the survey that may help to determine some valuable opinions towards making future progress. For example, it is important to understand the choices made for collecting and storing harms data, and to determine the potential advantages for developing a central database which was not fully understood from the survey. However this appears to be more of a complex issue, as was also highlighted in a UKCRC CTU

which recently developed a central PV system [142]. This CTU encountered a number of issues, particularly with the costs involved for training staff to manage the system, and the time spent transcribing the PV processes involved into SOPs. It also appears evident that a central system may only be beneficial for CTUs investigating certain diseases (e.g., cancer or surgical), where there is a greater volume of harms data.

It was clear from the survey that CTUs do use existing harms data from external sources during the trial safety monitoring, although to determine more about the exact methods and processes used, further in-depth discussions would be needed. Nevertheless, published trials and systematic reviews were amongst the most common external harms data source being used; mainly to support data monitoring committees, the preparation of development safety update reports, and to improve expedited reporting to sponsors and research ethic committees. But these responses still lacked detail. Further work would be to conduct interviews with the responding CTU members to understand their reasons for using the data in first place and what implications the data may have had in the long run. For example, was the data used to improve the design of a trial (e.g., recruitment, sample size, etc), or was it used to improve the trial safety monitoring and conduct.

The SDAs in the simulation study were assessed on simulated data sets; these assessments should also be carried out on a real dataset, preferably in similar real-world environments like in CTUs. This kind of assessment would also allow evaluations into the impact on resources used to evaluate detected signals and

the numbers required to screen the databases. In the simulation study we compared estimates for the false discovery rate (FDR) across these methods, although currently there is a lack of guidance for determining an optimal FDR. However as noted this may be affected by the time and resources needed when evaluating a signal, but also may vary in CTUs due to the limited resources available. This would need to be explored in future work.

Multivariate logistic regression modeling methods are now being explored by the FDA [173]. Now with the introduction of confounding with these methods, this potentially has improved the sensitivity-specificity performance in signal detection [192]. Although another study reports that these methods can also be restrictive in the detection of rare signals [183]. Therefore the potential of these methods is still unclear. Bayesian hierarchical mixture modeling techniques have also been researched for use in clinical trials [184], though the proposed idea of grouping and lumping AEs into one group then allocating it a prior distribution has been questioned in the past [185]. However this method introduces the potential of drug safety analysis using a Bayesian approach which is less tied to type I errors unlike the disproportionality analysis methods, and shows the potential promise these approaches may have in this area in the future and may even replace the use of standard meta-analysis techniques currently under use.

Finally there is an overwhelming requirement to determine accurate guidelines and gold standards when using SDAs, which currently is lacking in many of the good pharmacovigilance clinical practice guidelines, including the EMA and ICH.

Bibliography

1. Meyboom, R.H.B., *The case for good pharmacovigilance practice.* Pharmacoepidemiology and Drug Safety, 2000. 9(4): p. 335-336.
2. Edwards, I.R. and J.K. Aronson, *Adverse drug reactions: definitions, diagnosis, and management.* Lancet, 2000. 356(9237): p. 1255-9.
3. Ray, W.A., M.R. Griffin, and J. Avorn, *Evaluating drugs after their approval for clinical use.* N Engl J Med, 1993. 329(27): p. 2029-32.
4. Strom, B.L., *How the US drug safety system should be changed.* JAMA, 2006. 295(17): p. 2072-5.
5. Stricker, B.H. and B.M. Psaty, *Detection, verification, and quantification of adverse drug reactions.* Bmj, 2004. 329(7456): p. 44-7.
6. *Good Clinical Research Practice (GCP). Guidance for implementation.World Health Organization (WHO). Available at: http://apps.who.int/prequal/info_general/documents/GCP/GCP_handbook.pdf*
7. Cuervo, L.G. and M. Clarke, *Balancing benefits and harms in health care.* Bmj, 2003. 327(7406): p. 65-6.
8. Ioannidis, J.P., et al., *Better reporting of harms in randomized trials: an extension of the CONSORT statement.* Annals of Internal Medicine, 2004. 141(10): p. 781-8.
9. Ioannidis, J.P. and J. Lau, *Completeness of safety reporting in randomized trials: an evaluation of 7 medical areas.* JAMA. 285(4): p. 437-43.
10. Loke, Y.K. and S. Derry, *Reporting of adverse drug reactions in randomised controlled trials - a systematic survey.* BMC clinical pharmacology, 2001. 1: p. 3-3.
11. Tsang, R., L. Colley, and L.D. Lynd, *Inadequate statistical power to detect clinically significant differences in adverse event rates in randomized controlled trials.* Journal of Clinical Epidemiology, 2009. 62(6): p. 609-16.
12. Rothwell, P.M., *External validity of randomised controlled trials: "to whom do the results of this trial apply?".* Lancet, 2005. 365(9453): p. 82-93.
13. Chan A, H.A., Haahr MT, Gotzsche PC, Altman DG., *Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles.* JAMA, 2004. 291(20): p. 2457-2465.
14. Saini, P., et al., *Selective reporting bias of harm outcomes within studies: findings from a cohort of systematic reviews*. Vol. 349. 2014: BMJ.
15. Vedula, S.S., L. Tianjing, and K. Dickersin, *Differences in Reporting of Analyses in Internal Company Documents Versus Published Trial Reports: Comparisons in Industry-Sponsored Trials in Off-Label Uses of Gabapentin.* PLoS Medicine, 2013. 10(1): p. 1-13.
16. Rodgers, M.A., et al., *Reporting of industry funded study outcome data: comparison of confidential and published data on the safety and effectiveness of rhBMP-2 for spinal fusion.* BMJ (Clinical research ed.), 2013. 346: p. f3981-f3981.

17. Eyding, D., et al., *Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials*. BMJ, 2010.
18. *ClinicalTrials.gov. A service of the U.S. National Institutes of Health. Available at: https://clinicaltrials.gov/. Last accessed 25th March 2015.*
19. Klepper, M. and B. Cobert, *Drug Safety Data: How to Analyze, Summarize and Interpret to Determine Risk*. 2010: Jones & Bartlett Learning.
20. *Directive 2001/20/EC of the European Parliament and of the Council of 4 April 2001 on the approximation of the laws, regulations and administrative provisions of the member states relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use.* Med Etika Bioet, 2002. 9(1-2): p. 12-9.
21. *The Medicines for Human Use (Clinical Trials) Regulations 2004 S.I. 2004 No. 1031. Norwich: The Stationery Office; 2004 [http://www.legislation.gov.uk/uksi/2004/1031/contents/made].*
22. *International Conference on Harmonisation (1996) Harmonised Tripartite Guideline for Good Clinical Practice. Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E6/E6_R1_Guideline.pdf. .*
23. Moher, D., et al., *Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation.* JAMA, 2001. 285(15): p. 1992-5.
24. Altman, D.G., K.F. Schulz, and D. Moher, *CONSORT statement requires closer examination.* Bmj, 2002. 325(7376): p. 1364.
25. *John Talbot, Jeffrey K. Aronson. Stephens' Detection and Evaluation of Adverse Drug Reactions: Principles and Practice, 6th Edition. December 2011, Wiley-Blackwell. .*
26. *The Uppsala Monitoring Centre.The WHO Adverse Reaction Terminology (WHO-ART) Terminology for coding clinical information in relation to drug therapy. Available at: https://www.umc-products.com/graphics/28010.pdf.*
27. *National Institute of Health (NIH) U.S. National Library of Medicine. Coding Symbols for Thesaurus of Adverse Reaction Terms (COSTART) Source Information. Available at: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CST/.*
28. *The World Health Organisation (WHO). Classifications - International Classification of Diseases (ICD). Available at: http://www.who.int/classifications/icd/en/.*
29. *Medical Dictionary for Regulatory Activities Maintenance and Support Services Organization*. 2nd January 2013; Available from: http://www.meddramsso.com/index.asp.
30. Schroll, J.B., E. Maund, and P.C. Gøtzsche, *Challenges in Coding Adverse Events in Clinical Trials: A Systematic Review.* PLoS ONE, 2012. 7(7): p. e41174.
31. Vandenbroucke, J.P., *When are observational studies as credible as randomised trials?* Lancet, 2004. 363(9422): p. 1728-31.

32. von Elm, E., et al., *The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: Guidelines for Reporting Observational Studies.*: Epidemiology November 2007;18(6):800-804.
33. Vandenbroucke, J.P., et al., *Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration.*: Epidemiology November 2007;18(6):805-835.
34. Golder, S., Y.K. Loke, and M. Bland, *Meta-analyses of Adverse Effects Data Derived from Randomised Controlled Trials as Compared to Observational Studies: Methodological Overview.* PLoS Med, 2011. 8(5): p. e1001026.
35. Papanikolaou, P.N., G.D. Christidi, and J.P.A. Ioannidis, *Comparison of evidence on harms of medical interventions in randomized and nonrandomized studies.* CMAJ Canadian Medical Association Journal, 2006. 174(5): p. 635-41.
36. Singh, S. and Y.K. Loke, *Drug safety assessment in clinical trials: methodological challenges and opportunities.* Trials, 2012. 13(138): p. 1745-6215.
37. de Gans, J. and D. van de Beek, *Dexamethasone in Adults with Bacterial Meningitis.* New England Journal of Medicine, 2002. 347(20): p. 1549-1556.
38. Song, F., et al., *Dissemination and publication of research findings : an updated review of related biases.* Health Technology Assessment, 2010. 14(8): p. 234.
39. Santaguida, P.L. and P. Raina. *The Development of the McHarm Quality Assessment Scale for adverse events: Delphi Consensus on important criteria for evaluating harms*. 2008; Available from: http://hiru.mcmaster.ca/epc/mcharm.pdf.
40. Loke, Y.K. and K. Mattishent, *If nothing happens, is everything all right? Distinguishing genuine reassurance from a false sense of security.* Cmaj, 2015. 187(1): p. 15-6.
41. Higgins, J.P.T. and S. Green. *The Cochrane Collaboration. Cochrane Handbook for Systematic Reviews of Interventions.* March 2011; Version 5.1.0:[Available from: handbook.cochrane.org.
42. Aronson, J.K., *Monitoring for harms of therapy.* British Journal of Clinical Pharmacology, 2006. 61(4): p. 365-6.
43. Golder, S. and Y. Loke, *Search strategies to identify information on adverse effects: a systematic review.* Journal of the Medical Library Association. 97(2): p. 84-92.
44. Derry, S., Y.K. Loke, and J.K. Aronson, *Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials.* BMC Medical Research Methodology. 1: p. 7.
45. *The Cochrane Collaboration. Available at http://www.cochrane.org/. Last accessed 24th March 2015.*
46. *The Methods Group of the Cochrane Collaboration. Cochrane Adverse Effects Methods Group (AEMG). Last updated: Thursday 4th Jul 2013. Available at: http://aemg.cochrane.org/*

47. Loke, Y., D. Price, and A. Herxheimer, *Systematic reviews of adverse effects: framework for a structured approach.* BMC Medical Research Methodology, 2007. 7(1): p. 1-9.
48. Liberati, A., et al., *The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration.* PLoS Med, 2009. 6(7): p. e1000100.
49. Moher, D., et al., *Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses.* Lancet, 1999. 354(9193): p. 1896-900.
50. Zorzela, L., et al., *Quality of reporting in systematic reviews of adverse events: systematic review.* Bmj, 2014. 348.
51. Gould, A.L., *Practical pharmacovigilance analysis strategies.* Pharmacoepidemiol Drug Saf, 2003. 12(7): p. 559-74.
52. Stephenson, W.P. and M. Hauben, *Data mining for signals in spontaneous reporting databases: proceed with caution.* Pharmacoepidemiology and Drug Safety, 2007. 16(4): p. 359-365.
53. Moore, T.J. and C.D. Furberg, *Electronic Health Data for Postmarket Surveillance: A Vision Not Realized.* Drug Saf, 2015. 38(7): p. 601-10.
54. Robertson, A., et al., *Implementation and adoption of nationwide electronic health records in secondary care in England: qualitative analysis of interim results from a prospective national evaluation*. Vol. 341. 2010.
55. Hodkinson, A., et al., *Reporting of harms data in RCTs: a systematic review of empirical assessments against the CONSORT harms extension.* BMJ Open, 2013. 3(9).
56. Stevens, A., et al., *Relation of completeness of reporting of health research to journals' endorsement of reporting guidelines: systematic review.* Bmj, 2014. 25(348).
57. Altman, D.G., *Better reporting of randomised controlled trials: the CONSORT statement.* Bmj, 1996. 313(7057): p. 570-1.
58. Schulz, K.F., et al., *CONSORT 2010 changes and testing blindness in RCTs.* Lancet, 2010. 375(9721): p. 1144-6.
59. Hopewell, S., et al., *Endorsement of the CONSORT Statement by high impact factor medical journals: a survey of journal editors and journal 'Instructions to Authors'.* Trials, 2008. 9(20).
60. Plint, A.C., et al., *Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review.* Medical Journal of Australia, 2006. 185(5): p. 263-7.
61. *Consolidated Standards Of Reporting Trials (CONSORT) 2001 statement. Available at http://www.consort-statement.org/about-consort/history/consort-statement-2001/. Last accessed December 29, 2011.*
62. *Consolidated Standards Of Reporting Trials (CONSORT) statement 2010. Available at http://www.consort-statement.org/consort-statement/. Last accessed December 2011.*

63. Cobo, E., et al., *Effect of using reporting guidelines during peer review on quality of final manuscripts submitted to a biomedical journal: masked randomised trial.* Bmj. 343: p. d6783.
64. *Zetoc: Informing Research. Available at www.zetoc.mimas.ac.uk/. .*
65. Moher D, P.A., Altman DG, Schulz KF, Kober T, Galloway EK, Weeks L, Dias S, *Consolidated standards of reporting trials (CONSORT) and the quality of reporting of randomized controlled trials (Protocol).* The Cochrane Library 2010(3).
66. Bagul, N.B. and J.J. Kirkham, *The Reporting of Harms in Randomized Controlled Trials of Hypertension Using the CONSORT Criteria for Harm Reporting.* Clinical and Experimental Hypertension, 2012. 34(8): p. 548-554.
67. Breau, R.H., et al., *Reporting of harm in randomized controlled trials published in the urological literature.* Journal of Urology, 2010. 183(5): p. 1693-7.
68. Turner, L.A., et al., *An evaluation of the completeness of safety reporting in reports of complementary and alternative medicine trials.* BMC Complementary & Alternative Medicine, 2011. 11(67).
69. Shukralla, A.A., et al., *Reporting of adverse events in randomised controlled trials of antiepileptic drugs using the CONSORT criteria for reporting harms.* Epilepsy Research, 2011. 97(1-2): p. 20-9.
70. Capili, B., J.K. Anastasi, and J.N. Geiger, *Adverse event reporting in acupuncture clinical trials focusing on pain.* Clinical Journal of Pain, 2010. 26(1): p. 43-8.
71. Pitrou, I., et al., *Reporting of safety results in published reports of randomized controlled trials.* Archives of Internal Medicine. 169(19): p. 1756-61.
72. Haidich, A.B., et al., *The quality of safety reporting in trials is still suboptimal: survey of major general medical journals.* Journal of Clinical Epidemiology, 2011. 64(2): p. 124-35.
73. *Consolidated Standards Of Reporting Trials (CONSORT) Statement. CONSORT Endorsers - Journals. 4th August 2011. Accssed at: http://www.consort-statement.org/about-consort/consort-endorsement/consort-endorsers---journals/.*
74. Hirst, A. and D.G. Altman, *Are Peer Reviewers Encouraged to Use Reporting Guidelines? A Survey of 116 Health Research Journals.* PLoS ONE, 2012. 7(4).
75. Kirkham, J.J., et al., *The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews.* Bmj, 2010. 340.
76. Smyth, R.M.D., et al., *Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists.* Bmj, 2011. 342.
77. McGauran, N., et al., *Reporting bias in medical research - a narrative review.* Trials, 2010. 11(37): p. 1745-6215.
78. De Angelis, C., et al., *Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors.* New England Journal of Medicine, 2004. 351(12): p. 1250-1251.

79. *The World Health Organization (WHO). The International Clinical Trials Registry Platform (ICTRP). Available at: http://apps.who.int/trialsearch/Default.aspx. Last accessed December 2014.*
80. Wieseler, B., et al., *Completeness of Reporting of Patient-Relevant Clinical Trial Outcomes: Comparison of Unpublished Clinical Study Reports with Publicly Available Data.* PLoS Medicine, 2013. 10(10): p. 1-13.
81. Prayle, A.P., M.N. Hurley, and A.R. Smyth, *Compliance with mandatory reporting of clinical trial results on ClinicalTrials.gov: cross sectional study*. Vol. 344. BMJ, 2012.
82. *Structure and content of clinical study reports: E3*.
83. Doshi, P., T. Jefferson, and C. del Mar, *The imperative to share clinical study reports: Recommendations from the Tamiflu experience.* PLoS Medicine, 2012. 9(4).
84. Jefferson, T., et al., *Possible harms of oseltamivir--a call for urgent action.* Lancet, 2009. 374(9698): p. 1312-3.
85. Jureidini, J.N., L.B. McHenry, and P.R. Mansfield, *Clinical trials and drug promotion: Selective reporting of study 329.* The International Journal of Risk and Safety in Medicine, 2008. 20(1): p. 73-81.
86. Eyding, D., et al., *Reboxetine for acute treatment of major depression: systematic review and meta-analysis of published and unpublished placebo and selective serotonin reuptake inhibitor controlled trials.* Bmj, 2010. 341.
87. Gøtzsche, P.C. and A.W. Jørgensen, *Opening up data at the European Medicines Agency.* Bmj, 2011. 342.
88. Dyer, C., *European drug agency's attempts to improve transparency stalled by legal action from two US drug companies*. Vol. 346. BMJ, 2013.
89. *European Medicines Agency (EMA) policy on publication of clinical data for medicinal products for human use. 2nd October 2014 EMA/240810/2013. Policy/0070. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf.*
90. *GlaxoSmithKline (GSK) gives update on plans to share detailed clinical trial data as part of its commitment to transparency. Available at: http://www.gsk.com/media/press-releases/2013/gsk-gives-update-on-plans-to-share-detailed-clinical-trial-data-.html. Last accessed: 6 August 2013.*
91. Guerciolini, R., *Mode of action of orlistat.* International Journal of Obesity, 1997. 21(SUPPL. 3): p. S12-S23.
92. Johansson, K., et al., *Discontinuation due to adverse events in randomized trials of orlistat, sibutramine and rimonabant: a meta-analysis.* Obes Rev, 2009. 10(5): p. 564-75.
93. Li, Z., et al., *Meta-analysis: pharmacologic treatment of obesity.* Ann Intern Med, 2005. 142(7): p. 532-46.
94. Douglas, I.J., et al., *Orlistat and the risk of acute liver injury: self controlled case series study in UK Clinical Practice Research Datalink.* Bmj, 2013. 346: p. f1936.

95. *Any Count Software. Software for word counting for PDF file. Available at: http://www.anycount.com/. Last updated June 2014.*
96. Chanoine, J.P., et al., *Effect of orlistat on weight and body composition in obese adolescents: a randomized controlled trial.* JAMA, Journal of the American Medical Association, 2005. 293(23): p. 2873-2883.
97. Halpern, A., et al., *Latin-American trial of orlistat for weight loss and improvement in glycaemic profile in obese diabetic patients.* Diabetes, Obesity & Metabolism, 2003. 5(3): p. 180-188.
98. Sweeting, M.J., A.J. Sutton, and P.C. Lambert, *What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data.* Stat Med, 2004. 23(9): p. 1351-75.
99. Higgins, J.P.T., et al., *Measuring inconsistency in meta-analyses.* BMJ : British Medical Journal, 2003. 327(7414): p. 557-560.
100. Hanefeld, M. and G. Sachse, *The effects of orlistat on body weight and glycaemic control in overweight patients with type 2 diabetes: a randomized, placebo-controlled trial.* Diabetes, Obesity & Metabolism, 2002. 4(6): p. 415-423.
101. Kelley, D.E., et al., *Clinical efficacy of orlistat therapy in overweight and obese patients with insulin-treated type 2 diabetes: A 1-year randomized controlled trial.* Diabetes Care, 2002. 25(6): p. 1033-1041.
102. Torgerson, J.S., et al., *XENical in the Prevention of Diabetes in Obese Subjects (XENDOS) study: a randomized study of orlistat as an adjunct to lifestyle changes for the prevention of type 2 diabetes in obese patients.* Diabetes Care, 2004. 27(1): p. 155-161.
103. Maund, E., et al., *Benefits and harms in clinical trials of duloxetine for treatment of major depressive disorder: comparison of clinical study reports, trial registries, and publications.* Bmj, 2014. 4(348).
104. Bradburn, M.J., et al., *Much ado about nothing: a comparison of the performance of meta-analytical methods with rare events.* Statistics in Medicine, 2007. 26(1): p. 53-77.
105. Doshi, P., et al., *Restoring invisible and abandoned trials: a call for people to publish the findings.* Bmj, 2013. 346.
106. *REGULATION (EU) No 536/2014 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 16 April 2014. on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC. Official Journal of the European Union. Available at: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2014.158.01.0001.01.ENG. .*
107. *The Equator Network. Enhancing the QUAlity and Transparency Of health Research. Available at: http://www.equator-network.org/resource-centre/library-of-health-research-reporting/reporting-guidelines/other-reporting-guidelines/. .*
108. Loke, Y.K., S.P. Golder, and J.P. Vandenbroucke, *Comprehensive evaluations of the adverse effects of drugs: importance of appropriate study selection and data sources.* Therapeutic Advances in Drug Safety, 2011. 2(2): p. 59-68.
109. *Mann's Pharmacovigilance, 3rd Edition. Edited by Elizabeth B. Andrews and Nocholas Moore May 2014, Wiley-Blackwell.*

110. Singh, S., Y.K. Loke, and C.D. Furberg, *Thiazolidinediones and Heart Failure: A Teleo-Analysis.* Diabetes Care, 2007.
111. Scott, P.A., et al., *Non-steroidal anti-inflammatory drugs and myocardial infarctions: comparative systematic review of evidence from observational studies and randomised controlled trials.* Ann Rheum Dis, 2007. 66(10): p. 1296-304.
112. Brain L. Strom, S.E.K., Sean Hennessy, *Pharmacoepidemiology, February 2012, 5th Edition, Wiley-Blackwell. Last accessed May 2013.*
113. Hazell, L. and S.A. Shakir, *Under-reporting of adverse drug reactions : a systematic review.* Drug Saf, 2006. 29(5): p. 385-96.
114. Golder, S., et al., *Developing efficient search strategies to identify reports of adverse effects in MEDLINE and EMBASE.* Health Info Libr J, 2006. 23(1): p. 3-12.
115. Waller, P.C., *Making the most of spontaneous adverse drug reaction reporting.* Basic Clin Pharmacol Toxicol, 2006. 98(3): p. 320-3.
116. Edwards, I.R., *Spontaneous reporting—of what? Clinical concerns about drugs.* British Journal of Clinical Pharmacology, 1999. 48(2): p. 138-141.
117. Kazi, D., *Rosiglitazone and implications for pharmacovigilance.* BMJ : British Medical Journal, 2007. 334(7606): p. 1233-1234.
118. *Guidance for Industry: Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment: US Food and Drug Administration, 2005. Available at: http://www.fda.gov/downloads/regulatoryinformation/guidances/ucm126834.pdf.* .
119. *The Importance of Pharmacovigilance - Safety Monitoiring of medicinal products. The World Health Oranization (WHO) 2002. Last accessed 2013. Available at: http://apps.who.int/medicinedocs/en/d/Js4893e/.* .
120. AJ Avery, C.A., CM Bond, H Fortnum, A Gifford, PC Hannaford, L Hazell, J Krska, AJ Lee, DJ McLernon, E Murphy, S Shakir and MC Watson *Evaluation of patient reporting of adverse drug reactions to the UK 'Yellow Card Scheme': literature review, descriptive and qualitative analyses, and questionnaire surveys. May 2011. Published: NIHR HTA Programme www.hta.ac.uk*. Vol. 15.
121. *Medicines and Healthcare Products Regulatory Agency (MHRA) Yellow Card Scheme*. 12 August 2011]; Available from: http://www.mhra.gov.uk/Safetyinformation/Howwemonitorthesafetyofproducts/Medicines/TheYellowCardScheme/index.htm.
122. *British National Formulary (2005) BNF49. London: British Medical Association and the Royal Pharmaceutical Society of Great Britain. Available at: http://www.bnf.org/bnf/index.htm.*
123. Edwards, I.R., et al., *Global Drug Surveillance: The WHO Programme for International Drug Monitoring*, in *Pharmacoepidemiology*. 2007, John Wiley & Sons, Ltd. p. 161-183.
124. *VigiBase™. The Uppsala Monitoring Centre: Safeguarding patients.*; Available from: http://www.umc-products.com/DynPage.aspx?id=73590&mn1=1107&mn2=1132.

125. *The European Medicines Agenecy (EMA). EudraVigilance: Mandatory e-reporting essentials. Last accessed 2013.*; Available from: http://eudravigilance.ema.europa.eu/human/index.asp.
126. Davies, E.C., et al., *Adverse drug reactions in hospital in-patients: a pilot study.* J Clin Pharm Ther, 2006. 31(4): p. 335-41.
127. *Drug Safety Research Unit (DSRU). Modified-Prescription Event Monitoring (M-PEM). Available at: http://www.dsru.org/studies-for-risk-management/m-pem-modified-prescription-event-monitoring-studies.*
128. Layton, D., L. Hazell, and S.A. Shakir, *Modified prescription-event monitoring studies: a tool for pharmacovigilance and risk management.* Drug Saf, 2011. 34(12).
129. *Safety of Medicines - A guide to detecting and reporting adverse drug reactions. Why health professionals need to take action. World Health Organization, 2002. Available at: http://whqlibdoc.who.int/hq/2002/WHO_EDM_QSM_2002.2.pdf.*
130. *The General Practice Research Database (GPRD)*. Further Information for patients; Available from: http://www.erskinepractice.scot.nhs.uk/website/S11486/files/GPRD_PatientLeaflet.pdf.
131. *Clinical Practice Research Datalink (CPRD), Medicines and Healthcare products Regulatory Agency (MHRA). Available at: http://www.cprd.com/intro.asp. Last accessed 12th Febrary 2014.*
132. Garcia Rodriguez, L.A. and S. Perez Gutthann, *Use of the UK General Practice Research Database for pharmacoepidemiology.* British Journal of Clinical Pharmacology, 1998. 45(5): p. 419-25.
133. *The Health Improvement Network (THIN) Research Team*. 15 December 10]; Available from: http://www.ucl.ac.uk/pcph/research-groups-themes/thin-pub/database.
134. *The Tayside Medicines Monitoring Unit (MEMO): A Record-Linkage System for Pharmacovigilance. Available at: http://www.dundee.ac.uk/memo/memoonly/RL.HTM.*
135. Lazarou, J., B.H. Pomeranz, and P.N. Corey, *Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies.* JAMA, 1998. 279(15): p. 1200-5.
136. Beijer, H.J. and C.J. de Blaey, *Hospitalisations caused by adverse drug reactions (ADR): a meta-analysis of observational studies.* Pharm World Sci, 2002. 24(2): p. 46-54.
137. Xu, H., Y. Wang, and N. Liu, *A hospital-based survey of healthcare professionals in the awareness of pharmacovigilance.* Pharmacoepidemiology & Drug Safety, 2009. 18(7): p. 624-630.
138. Belton, K.J., *Attitude survey of adverse drug-reaction reporting by health care professionals across the European Union. The European Pharmacovigilance Research Group.* European Journal of Clinical Pharmacology, 1997. 52(6): p. 423-427.
139. Wood, L. and C. Martinez, *The general practice research database: role in pharmacovigilance.* Drug Safety, 2004. 27(12): p. 871-81.

140. McNaughton, R., G. Huet, and S. Shakir, *An investigation into drug products withdrawn from the EU market between 2002 and 2011 for safety reasons and the evidence used to support the decision-making.* BMJ Open, 2014. 4(1).
141. Clarke, A., J. Deeks, and S.W. Shakir, *An Assessment of the Publicly Disseminated Evidence of Safety Used in Decisions to Withdraw Medicinal Products from the UK and US Markets.* Drug Safety, 2006. 29(2): p. 175-181.
142. *Eleanor M Dinnett, Sharon Kean, Elizabeth P Tolmie, et al. Implementing a centralised pharmacovigilance service in a non-commercial setting in the United Kingdom. Trials 2013, 14:171. .*
143. Chalmers, I., et al., *Data sharing among data monitoring committees and responsibilities to patients and science.* Trials, 2013. 14(102): p. 1745-6215.
144. Hicks, L.K., A. Laupacis, and A.S. Slutsky, *A primer on data safety monitoring boards: mission, methods, and controversies.* Intensive Care Med, 2007. 33(10): p. 1815-8.
145. *DAMOCLES Study Group, NHS Health Technology Assessment Programme. A proposed charter for clinical trial data monitoring committees: helping them to do their job well.* Lancet, 2005. 365(9460): p. 711-22.
146. Pocock, S.J., *When to stop a clinical trial.* BMJ : British Medical Journal, 1992. 305(6847): p. 235-240.
147. *UK Clinical Research Collaboration (CRC) Registered Clinical Trial Unit Network. Last updated September 2014. Available at: http://www.ukcrc-ctu.org.uk/. .*
148. *EudraVigilance. Mandatory e-reporting essentials. Last update: Wednesday, 24 September 2014. Available at: https://eudravigilance.ema.europa.eu/human/. .*
149. *Management of Safety Information from Clinical Trials: Report of CIOMS Working Group VI. CIOMS 2005. Available at: http://www.cioms.ch/index.php/available-publications?task=view&id=25&catid=54.*
150. *Council for International Organizations of Medical Sciences (CIOMS): Practical aspects of signal detection in Pharmacovigilance: Report of CIOMS Working Group VIII. 2010. Available at: http://www.cioms.ch/index.php/publications/available-publications?task=view&id=27&catid=54. .* Available from: http://www.cioms.ch/index.php/publications/available-publications?task=view&id=27&catid=54.
151. *ICH Harmonised Tripartite Guideline*. Revision of the ICH guideline on clinical safety data management: Data elements for transmission of individual case safety reports E2B(R) 12th May 2005]; Available from: http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm129399.pdf.
152. Meyboom, R.H., et al., *Principles of signal detection in pharmacovigilance.* Drug Safety, 1997. 16(6): p. 355-65.

153. *Waller P. An Introduction to Pharmacovigilance, October 2009. Wiley-Blackwell.* .
154. *Periodic Safety Update Reports (PSURs)*. 11th July 2012]; Available from: http://www.mhra.gov.uk/Howweregulate/Medicines/Licensingofmedicines/Informationforlicenceapplicants/PeriodicSafetyUpdateReports/index.htm.
155. Bate, A., M. Lindquist, and I.R. Edwards, *The application of knowledge discovery in databases to post-marketing drug safety: example of the WHO database.* Fundamental & Clinical Pharmacology, 2008. 22(2): p. 127-140.
156. Bate, A. and S.J.W. Evans, *Quantitative signal detection using spontaneous ADR reporting.* Pharmacoepidemiology and Drug Safety, 2009. 18(6): p. 427-436.
157. Aronson, J.K. and R.E. Ferner, *Clarification of terminology in drug safety.* Drug Safety. 28(10): p. 851-70.
158. Evans, S.J.W., P.C. Waller, and S. Davis, *Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports.* Pharmacoepidemiology and Drug Safety, 2001. 10(6): p. 483-486.
159. Bate, A., et al., *A Bayesian neural network method for adverse drug reaction signal generation.* European Journal of Clinical Pharmacology, 1998. 54(4): p. 315-321.
160. DuMouchel, W. and D. Pregibon, *Empirical bayes screening for multi-item associations*, in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*2001, ACM: San Francisco, California. p. 67-76.
161. van Puijenbroek, E.P., et al., *A comparison of measures of disproportionality for signal detection in spontaneous reporting systems for adverse drug reactions.* Pharmacoepidemiology and Drug Safety, 2002. 11(1): p. 3-10.
162. Hauben, M. and A. Bate, *Decision support methods for the detection of adverse events in post-marketing data.* Drug Discov Today, 2009. 14(7-8): p. 343-57.
163. Norén, G.N., et al., *Extending the methods used to screen the WHO drug safety database towards analysis of complex associations and improved accuracy for rare events.* Statistics in Medicine, 2006. 25(21): p. 3740-3757.
164. Szarfman, A., S.G. Machado, and R.T. O'Neill, *Use of screening algorithms and computer systems to efficiently signal higher-than-expected combinations of drugs and events in the US FDA's spontaneous reports database.* Drug Saf, 2002. 25(6): p. 381-92.
165. Alvarez, Y., et al., *Validation of statistical signal detection procedures in eudravigilance post-authorization data: a retrospective evaluation of the potential for earlier signalling.* Drug Saf, 2010. 33(6): p. 475-87.
166. *European Medicines Agency (EMA) Guideline on good pharmacovigilance practices (GVP): Module IX - Signal management. 22 June 2012, EMA/827661/2011. Available at:*

http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2012/06/WC500129138.pdf.

167. Levine, J.G., J.M. Tonning, and A. Szarfman, *Reply: The evaluation of data mining methods for the simultaneous and systematic detection of safety signals in large databases: lessons to be learned.* British Journal of Clinical Pharmacology, 2006. 61(1): p. 105-113.
168. Hammond, I.W., et al., *Database size and power to detect safety signals in pharmacovigilance.* Expert Opin Drug Saf, 2007. 6(6): p. 713-21.
169. Lehman, H.P., et al., *An evaluation of computer-aided disproportionality analysis for post-marketing signal detection.* Clin Pharmacol Ther, 2007. 82(2): p. 173-80.
170. de Boer, A., *When to publish measures of disproportionality derived from spontaneous reporting databases?* British Journal of Clinical Pharmacology, 2011. 72(6): p. 909-911.
171. Montastruc, J.-L., et al., *Benefits and strengths of the disproportionality analysis for identification of adverse drug reactions in a pharmacovigilance database.* British Journal of Clinical Pharmacology, 2011. 72(6): p. 905-908.
172. Roux, E., et al., *Evaluation of statistical association measures for the automatic signal generation in pharmacovigilance.* IEEE Trans Inf Technol Biomed, 2005. 9(4): p. 518-27.
173. Harpaz, R., et al., *Novel Data-Mining Methodologies for Adverse Drug Event Discovery and Analysis.* Clin Pharmacol Ther, 2012. 91(6): p. 1010-1021.
174. Almenoff, J.S., et al., *Novel Statistical Tools for Monitoring the Safety of Marketed Drugs.* Clin Pharmacol Ther, 2007. 82(2): p. 157-166.
175. Bäckström, M., T. Mjörndal, and R. Dahlqvist, *Spontaneous reporting of adverse drug reactions by nurses.* Pharmacoepidemiology and Drug Safety, 2002. 11(8): p. 647-650.
176. Brown, E., *Effects of Coding Dictionary on Signal Generation.* Drug Safety, 2002. 25(6): p. 445-452.
177. Bate, A. and S. Evans, *Quantitative signal detection using spontaneous ADR reporting.* Pharmacoepidemiology and Drug Safety, 2009. 18(6): p. 427-436.
178. *PIPA Guidelines for Signal Management: PIPA Guidelines for Signal Management for Small and Medium Sized Pharmaceutical Companies. Accessed 25th August 2015. Available at: http://www.pipaonline.org/write/MediaManager/Members%20Area/Pharmacovigilance/Signal%20Detection/PIPA_UK_Guidelines_Signal_Management_May_2012.pdf*
179. Evans, S.J., P.C. Waller, and S. Davis, *Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports.* Pharmacoepidemiol Drug Saf, 2001. 10(6): p. 483-6.
180. Meyboom, R.H., et al., *Causal or casual? The role of causality assessment in pharmacovigilance.* Drug Saf, 1997. 17(6): p. 374-89.

181. Brown, J., et al., *Drug Adverse Event Detection in Health Plan Data Using the Gamma Poisson Shrinker and Comparison to the Tree-based Scan Statistic.* Pharmaceutics, 2013. 5(1): p. 179-200.
182. Trifiro, G., et al., *EU-ADR healthcare database network vs. spontaneous reporting system database: preliminary comparison of signal detection.* Stud Health Technol Inform, 2011. 166: p. 25-30.
183. Maignen, F., et al., *Assessing the extent and impact of the masking effect of disproportionality analyses on two spontaneous reporting systems databases.* Pharmacoepidemiol Drug Saf, 2014. 23(2): p. 195-207.
184. Berry, S.M. and D.A. Berry, *Accounting for Multiplicities in Assessing Drug Safety: A Three-Level Hierarchical Mixture Model.* Biometrics, 2004. 60(2): p. 418-426.
185. Berry, D., *Discussion of "Multivariate Bayesian Logistic Regression for Analysis of Clinical Trial Safety Issues" by W. DuMouchel.* 2012(3): p. 344-345.
186. Waller, P., E. Heeley, and J. Moseley, *Impact analysis of signals detected from spontaneous adverse drug reaction reporting data.* Drug Saf, 2005. 28(10): p. 843-50.
187. Hill, A.B., *The Environment and Disease: Association or Causation?* Proc R Soc Med, 1965. 58: p. 295-300.
188. *Guideline on the use of statistical signal detection methods in the Eudravigilance data analysis system*.
189. Ong, P.S., et al., *Colonic stricture in a boy with cystic fibrosis.* Postgraduate Medical Journal, 1995. 71(835): p. 309-312.
190. Goldman, S.A., *Limitations and strengths of spontaneous reports data.* Clinical Therapeutics, 1998. 20: p. C40-C44.
191. Ahmed, I., et al., *Bayesian pharmacovigilance signal detection methods revisited in a multiple comparison setting.* Statistics in Medicine, 2009. 28(13): p. 1774-1792.
192. Caster, O., et al., *Logistic regression in signal detection: another piece added to the puzzle*. Clin Pharmacol Ther. 2013 Sep;94(3):312.
193. Koutkias, V.G. and M.-C. Jaulent, *Computational Approaches for Pharmacovigilance Signal Detection: Toward Integrated and Semantically-Enriched Frameworks.* Drug Safety, 2015. 38(3): p. 219-232.
194. *The Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT). Last updated May 2013. Available at: http://www.imi-protect.eu/.*
195. *The Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium (PROTECT) Adverse Drug Reactions Database. Last update 30th June 2013. Available at: http://www.imi-protect.eu/adverseDrugReactions.shtml.* . Available from: http://www.imi-protect.eu/adverseDrugReactions.shtml.
196. *Health and social care information centre (HSCIC). Prescribing by GP practice. Available at: http://www.hscic.gov.uk/gpprescribingdata.*
197. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. 57(1): p. 289-300.

198. Ahmed, I., et al., *False Discovery Rate Estimation for Frequentist Pharmacovigilance Signal Detection Methods.* Biometrics, 2010. 66(1): p. 301-309.
199. *Observational Medical Outcomes Partnership (OMOP). Available at: http://omop.org/. Last accssed May 2013.*
200. *SAS 9.3 product documentation. Available at: http://www.sas.com/en_us/home.html.*
201. *PhViD: an R package for PharmacoVigilance signal Detection. Available at: http://cran.r-project.org/web/packages/PhViD/PhViD.pdf.*
202. Tubert, P., et al., *Power and weakness of spontaneous reporting: A probabilistic approach.* Journal of Clinical Epidemiology, 1992. 45(3): p. 283-286.
203. *Lance A. Waller, Carol A. Gotway. Applied Spatial Statistics for Public Health Data. Wiley, July 2004.*
204. Maignen, F., et al., *A conceptual approach to the masking effect of measures of disproportionality.* Pharmacoepidemiol Drug Saf, 2014. 23(2): p. 208-17.
205. Huang, L., et al., *Zero-inflated Poisson model based likelihood ratio test for drug safety signal detection.* Stat Methods Med Res, 2014. 3.
206. *ICH harmonised tripartit guideline. Statistcal principles for clinical trials E9. Current Step 4 version dated 5 February 1998 Available at: http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf. .*
207. Mehrotra, D.V. and J.F. Heyse, *Use of the false discovery rate for evaluating clinical safety data.* Stat Methods Med Res, 2004. 13(3): p. 227-38.
208. Mehrotra, D.V. and A.J. Adewale, *Flagging clinical adverse experiences: reducing false discoveries without materially compromising power for detecting true signals.* Stat Med, 2012. 31(18): p. 1918-30.
209. Hauben, M., et al., *The role of data mining in pharmacovigilance.* Expert Opinion on Drug Safety, 2005. 4(5): p. 929-948.
210. Trifiro, G., et al., *The EU-ADR project: preliminary results and perspective.* Stud Health Technol Inform, 2009. 148: p. 43-9.
211. *Innovation in medical Evidence development and surveillance (IMEDS). Advancing Regulatory Science for Public Health. Last updated 2013. Available at: http://imeds.reaganudall.org/.*
212. *Mini-Sentinel. Last updated 21st November 2013. Available at: http://www.mini-sentinel.org/.* . Available from: http://www.mini-sentinel.org/.
213. Ioannidis, J.P., *Adverse events in randomized trials: neglected, restricted, distorted, and silenced*. Arch Intern Med. 2009 Oct 26;169(19):1737-9. doi: 10.1001/archinternmed.2009.313.
214. Moher, D., et al., *Assessing the quality of reports of randomized trials in pediatric complementary and alternative medicine.* BMC Pediatrics, 2002. 2(2).
215. Doshi, P., M. Jones, and T. Jefferson, *Rethinking credible evidence synthesis.* Bmj, 2012. 344.

216. Szarfman, A., J.M. Tonning, and P.M. Doraiswamy, *Pharmacovigilance in the 21st century: new systematic tools for an old problem*. Pharmacotherapy. 2004 Sep;24(9):1099-104.
217. Hopewell, S., et al., *Effect of editors' implementation of CONSORT guidelines on the reporting of abstracts in high impact medical journals: interrupted time series analysis.* Bmj, 2012. 344: p. e4178.
218. Jefferson, T., et al., *Oseltamivir for influenza in adults and children: systematic review of clinical study reports and summary of regulatory comments.* Bmj, 2014. 9(348).

Appendix A – Search strategy and Forest plots from Chapter 2

Search strategy:

Ovid MEDLINE:

1. harm*
2. Safe*
3. CONSORT
4. Consolidation of standards reporting trials
5. (#3 OR #4)
6. (#1 OR #3 OR #4)
7. (#2 OR #5)
8. Toxic*
9. Adverse events
10. Adverse effects
11. Adverse
12. (#9 OR #10 OR #11)
13. (#5 OR #12)
14. Randomised
15. Randomized
16. RCTs
17. Randomised controlled trials
18. Randomized controlled trials
19. (#14 OR #15 OR #16 OR #17 OR #18)
20. Clinical trials
21. Side effect
22. Risk*
23. Complication*
24. Treatment next emergent
25. Post marketing next surveillance
26. drug next surveillance
27. (#5 OR #12 OR #19)
28. (#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 OR #18 OR #19 OR #20 OR #21 OR #22 OR #23 OR #24 OR #25 OR #26)
29. consort or consolidat$ standard$
30. *randomized controlled trials/
31. *clinical trials/

truncates the word e.g. harm (harm, harms or harmful).

ISI Web of Knowledge:

1. harm*
2. Safety
3. CONSORT
4. Consolidation of standards reporting trials
5. Adverse events
6. Adverse effects
7. (#1 OR #2 OR #3 OR #4 OR #5 OR #6)
8. ((consort OR 'consolidat*') AND (checklist* OR quality))

Scopus:

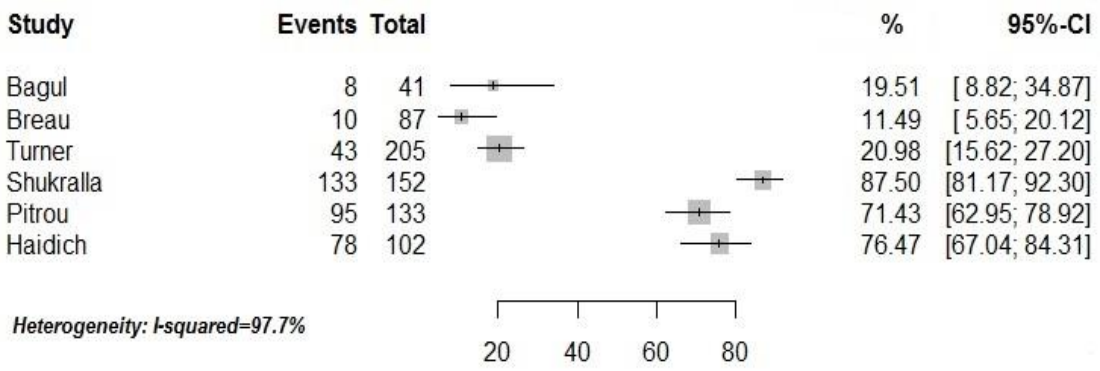
1. harm*
2. Safety
3. CONSORT
4. Consolidation of standards reporting trials
5. Adverse events
6. Adverse effects
7. (#1 OR #2 OR #3 OR #4 OR #5 OR #6)

Cochrane Library:

1. CONSORT
2. Consolidation of standards reporting trials
3. harms
4. Safety
5. Adverse events
6. RCTs
7. Randomised controlled trials
8. (#1 OR #2 OR #3 OR #4 OR #5 OR #6 OR #7)

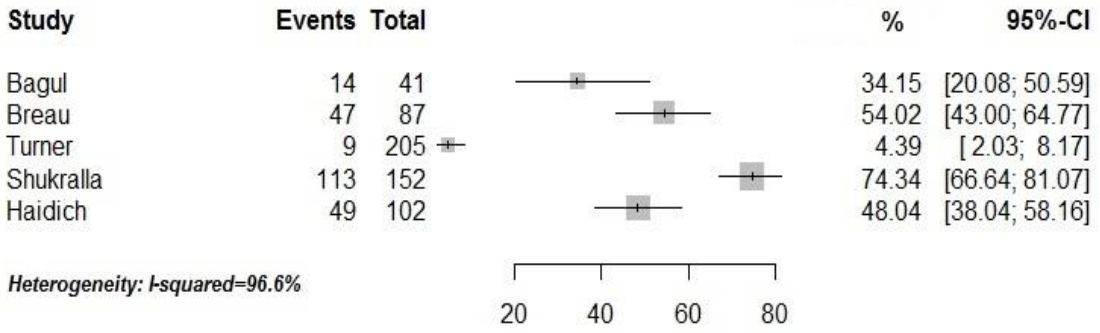
Figures 20: Forest plots for the CONSORT harms recommendations

Recommendation 1: Title & Abstract



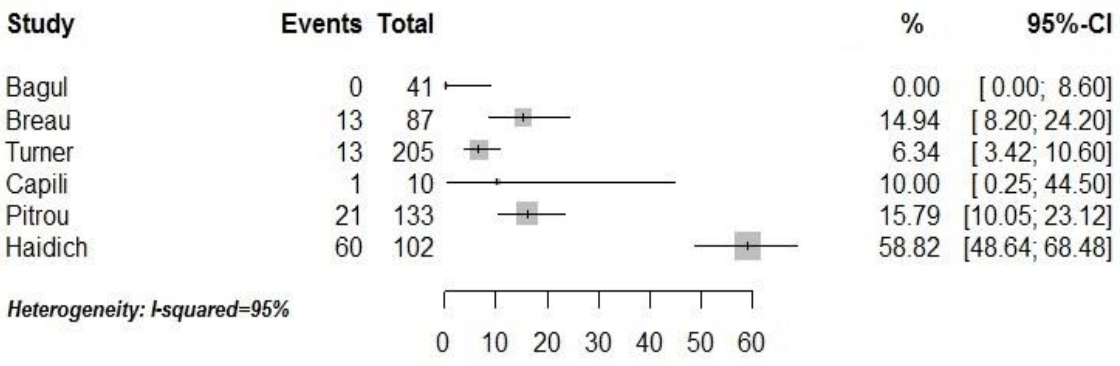
Recommendation not assessed in Capili study.

Recommendation 2: Introduction



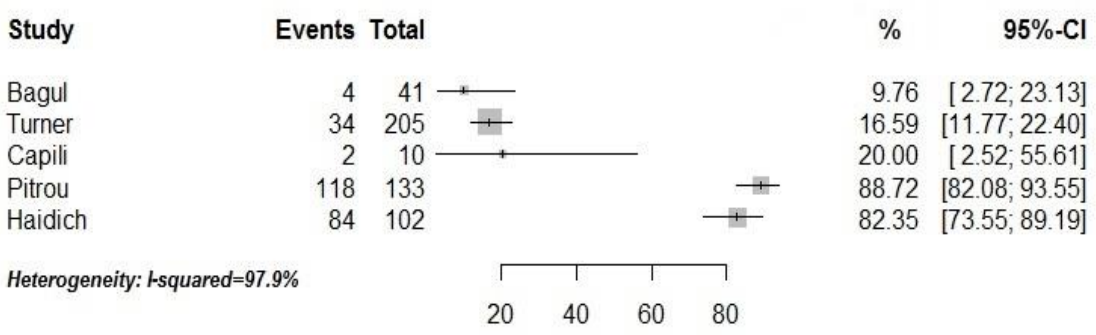
Studies Capili and Pitrou did not report recommendation

Recommendation 3: Definition of Adverse events



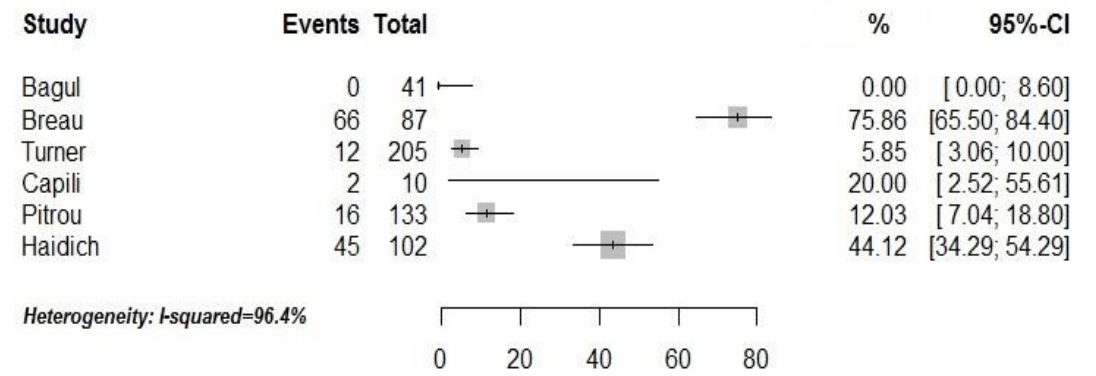
Shukralla reports recommendation as multiple items.

Recommendation 4: Collection of harms data



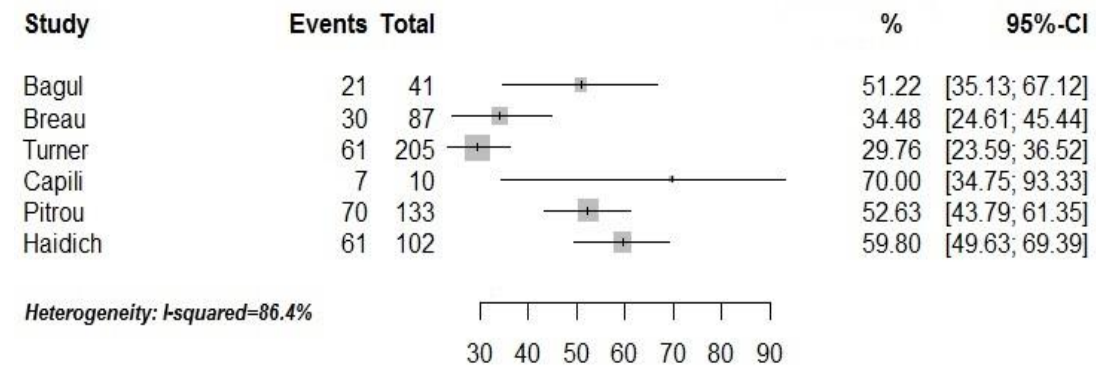
Breau & Shukralla report recommendation with multiple items

Recommendation 5: Analysis of harms



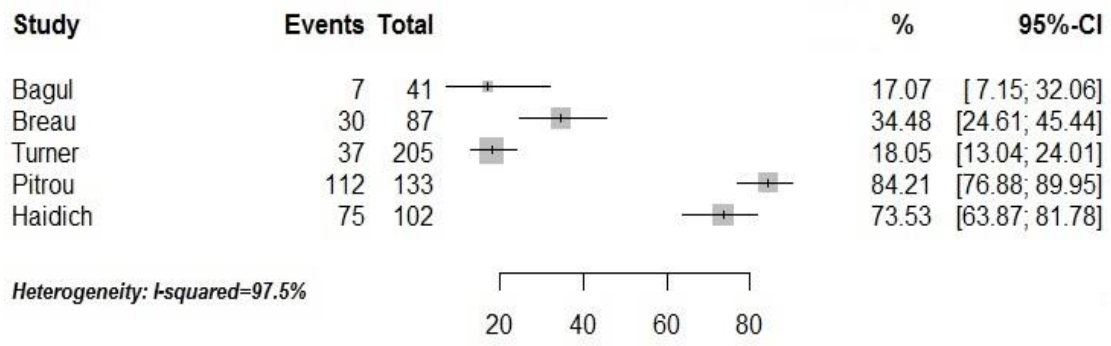
Shukralla reports the recommendation with multiple items

Recommendation 6: Withdrawals



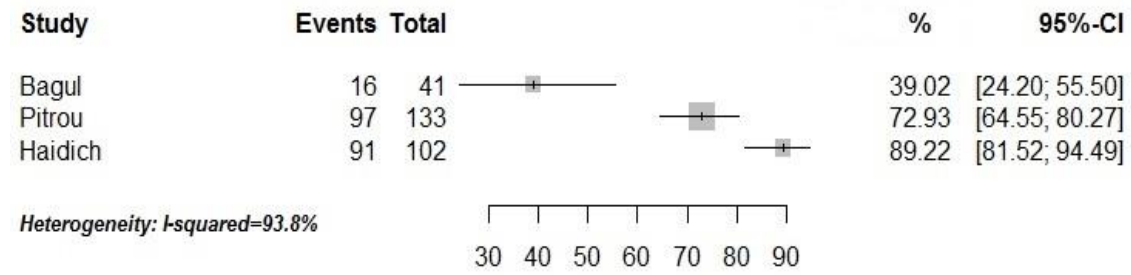
Shukralla reports the recommendation with multiple items

Recommendation 7: Number of patients analyzed



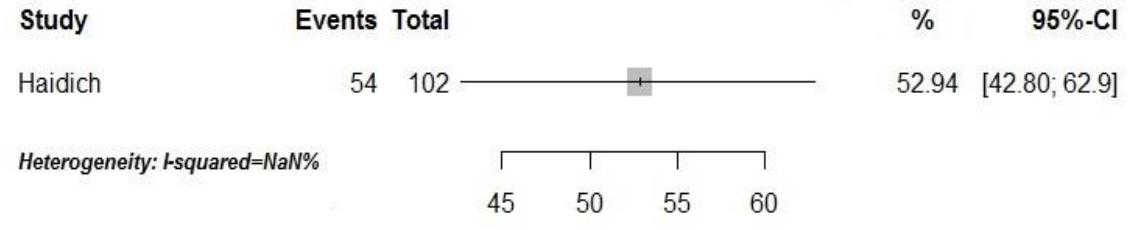
Shukralla reports the recommendation with multiple items, and Capili did not report the recommendation.

Recommendation 8: Results for each adverse event



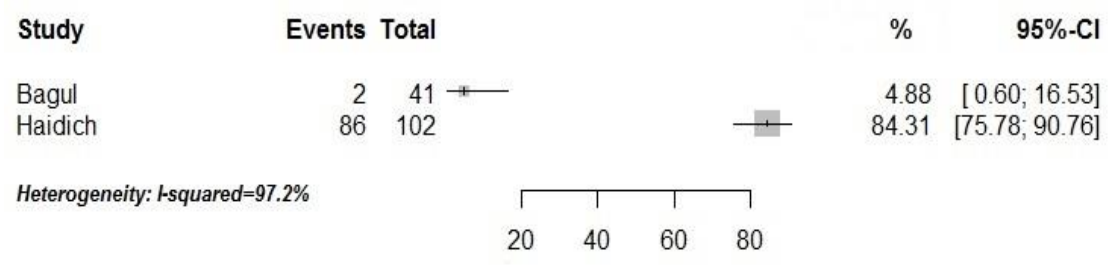
Breau and Shukralla report with multiple items. Turner chose not to assess this recommendation and Capili did not report.

Recommendation 9: Subgroup analysis



Bagul, Breau, Turner and Shukralla chose not to assess this recommendation. Capili and Pitrou did not report.

Recommendation 10: Balanced discussion



Breau and Shukralla report with multiple items. Turner chose not to assess. Capili and Pitrou did not report.

Appendix B – Search Strategy and Further Results from Chapter 3

Table 28: Search criteria used in both Cochrane central and MEDLINE.

Search set	CENTRAL	MEDLINE
1	(Orlistat or Xenical)	randomized controlled trial.pt.
2		controlled clinical trial.pt.
3		randomized.ab.
4		placebo.ab.
5		clinical trials as topic.sh.
6		randomly.ab.
7		trial.ti.
8		7 or 5 or 2 or 6 or 1 or 4 or 3
9		(orlistat or Xenical).tw.
10		8 and 9

Table 29: Reporting of serious adverse events (SAEs) in Clinical study reports (CSRs) and journal articles for Olistat trials.

Trial ID	NM16189	M37013	M37002	M37047	BM15421
Listed Serious adverse event- MedDRA preferred term.					
Respiratory, thoracic and medicinal disorders					
Adenoidal hypertrophy	BOTH+	NR	NR	NR	NR
Asthma aggravated	BOTH+	NR	NR	CSR	CSR
Asthma	BOTH+	NR	NR	NR	CSR
Bronchospsm aggravated	CSR	NR	NR	NR	NR
Nasal septum deviation	BOTH+	NR	NR	NR	NR
Nasal polyps	NR	NR	NR	CSR	NR
Pleural effusion	NR	NR	NR	CSR	NR
Dyspnoea	NR	NR	NR	NR	CSR
Epistaxis	NR	NR	NR	NR	CSR
Oropharyngeal swelling	NR	NR	NR	NR	CSR
Pleurisy	NR	NR	NR	NR	CSR
Rhinitis seasonal	NR	NR	NR	NR	CSR
Nervous system disorders					
Convulsions	CSR	NR	NR	NR	CSR
Demyelination	CSR	NR	NR	NR	NR
Facial palsy	BOTH+	NR	NR	NR	CSR
Meningitis aseptic	BOTH+	NR	NR	NR	NR
Trigeminal neuralgia	NR	NR	CSR	NR	NR
Migraine	NR	NR	NR	CSR	CSR
Radiculopathy	NR	NR	NR	CSR	NR
Spinal stenosis	NR	NR	NR	CSR	NR

Cerebrovascular accident	NR	NR	NR	NR	CSR
Headache	NR	NR	NR	NR	CSR
Syncope	NR	NR	NR	NR	CSR
Epilepsy	NR	NR	NR	NR	CSR
Multiple sclerosis	NR	NR	NR	NR	CSR
Sleep apnoea syndrome	NR	NR	NR	NR	CSR
Carpal tunnel syndrome	NR	NR	NR	NR	CSR
Dizziness (exc. vertigo)	NR	NR	NR	NR	CSR
Entrapment neuropathy	NR	NR	NR	NR	CSR
Haemorrhagic stroke	NR	NR	NR	NR	CSR
Loss of consciousness	NR	NR	NR	NR	CSR
Migraine aggravated	NR	NR	NR	NR	CSR
Multiple sclerosis aggravated	NR	NR	NR	NR	CSR
Obstructive sleep apnoea syndrome	NR	NR	NR	NR	CSR
Encephalomyelitis	Pub	NR	NR	NR	NR
Spinal cord compression	NR	NR	NR	NR	CSR
Hepato-Biliary Disorders					
Choleithiasis	BOTH+	NR	NR	CSR	CSR
Gall bladder disorder	BOTH+	NR	NR	NR	NR
Cholecystitis	NR	NR	NR	CSR	CSR
Bile duct stone	NR	NR	NR	NR	CSR
Biliary colic	NR	NR	NR	NR	CSR
Bile duct obstruction	NR	NR	NR	NR	CSR
Cholangitis	NR	NR	NR	NR	CSR
Gall bladder disease	NR	NR	NR	NR	CSR
Gall bladder pain	NR	NR	NR	NR	CSR
Hepatotoxicity	NR	NR	NR	NR	CSR

Infections and infestations					
Erysipelas	NR	NR	CSR	NR	CSR
Acute exacerbation of chronic	NR	NR	CSR	NR	NR
Bronchitis	NR	NR	CSR	NR	CSR
Cholecystitis acute	NR	NR	CSR	NR	CSR
Fungal infection	NR	NR	CSR	NR	NR
Sepsis	NR	NR	CSR	NR	NR
Pilonidal abscess	BOTH+	NR	NR	NR	NR
Pneumonia	BOTH+	NR	NR	CSR	CSR
Cellulitis	NR	NR	NR	CSR	CSR
Infected skin ulcer	NR	NR	NR	CSR	NR
Pyelonephritis	NR	NR	NR	CSR	NR
Staphylococcal infection	NR	NR	NR	CSR	NR
Upper respiratory tract	NR	NR	NR	CSR	NR
Infection	NR	NR	NR	NR	NR
Urosepsis	NR	NR	NR	CSR	NR
Pyelonephritis	NR	NR	NR	NR	CSR
Gastroenteritis	NR	NR	NR	NR	CSR
Bacterial infection	NR	NR	NR	NR	CSR
Localised infection	NR	NR	NR	NR	CSR
Salpingitis	NR	NR	NR	NR	CSR
Gastroenteritis helicobacter	NR	NR	NR	NR	CSR
Abscess	NR	NR	NR	NR	CSR
Ano-rectal infection	NR	NR	NR	NR	CSR
Bronchitis acute	NR	NR	NR	NR	CSR
Cervicitis	NR	NR	NR	NR	CSR
Empyema	NR	NR	NR	NR	CSR

Eye abscess	NR	NR	NR	NR	CSR
Gastroenteritis salmonella	NR	NR	NR	NR	CSR
Haemorrhagic fever	NR	NR	NR	NR	CSR
Hepatitis b	NR	NR	NR	NR	CSR
Herpes zoster	NR	NR	NR	NR	CSR
Meningitis bacterial	NR	NR	NR	NR	CSR
Meningitis	NR	NR	NR	NR	CSR
Ovarian abscess	NR	NR	NR	NR	CSR
Pyelonephritis acute	NR	NR	NR	NR	CSR
Sinusitis	NR	NR	NR	NR	CSR
Tuberculosis	NR	NR	NR	NR	CSR
Urinary tract infection	NR	NR	NR	NR	CSR
Psychiatric disorders					
Depression	BOTH-*	NR	NR	NR	CSR
Suicidal ideation	CSR	NR	NR	NR	NR
Alcoholism	NR	NR	NR	NR	CSR
Alcoholic withdrawal symptoms	NR	NR	NR	NR	CSR
Anxiety disorder	NR	NR	NR	NR	CSR
Completed suicide	NR	NR	NR	NR	CSR
Suicide attempt	NR	NR	NR	NR	CSR
Gastrointestinal disorders					
Diarrhea & dehydration	NR	BOTH+	NR	NR	NR
Cholecistectomy due to chronic Cholelithiasis	NR	CSR	NR	NR	NR
Appendicitis	BOTH+	NR	NR	NR	CSR
Diverticulitis	NR	NR	CSR	NR	CSR
Ileus	NR	NR	CSR	NR	CSR

Inguinal hernia	NR	NR	CSR	NR	CSR
Umbilical hernia	NR	NR	CSR	NR	CSR
Gastrointestinal haemorrhage	NR	NR	NR	CSR	CSR
Impaired gastric emptying	NR	NR	NR	CSR	NR
liquid stools	CSR	NR	CSR	CSR	CSR
Abdominal pain	NR	NR	NR	NR	CSR
Abdominal pain upper	NR	NR	NR	NR	CSR
Pancreatitis	NR	NR	NR	NR	CSR
Gastritis	NR	NR	NR	NR	CSR
Incisional hernia	NR	NR	NR	NR	CSR
Diverticulum nos	NR	NR	NR	NR	CSR
Duodenal ulcer	NR	NR	NR	NR	CSR
Gastric ulcer	NR	NR	NR	NR	CSR
Oesophagitis	NR	NR	NR	NR	CSR
Pancreatitis acute	NR	NR	NR	NR	CSR
Anal fissure	NR	NR	NR	NR	CSR
Appendicitis perforated	NR	NR	NR	NR	CSR
Colitis	NR	NR	NR	NR	CSR
Decreased defecation	NR	NR	NR	NR	CSR
Diverticulum intestinal	NR	NR	NR	NR	CSR
Duodenitis	NR	NR	NR	NR	CSR
Enteritis	NR	NR	NR	NR	CSR
Flatulence	NR	NR	NR	NR	CSR
Haemorrhoids	NR	NR	NR	NR	CSR
General disorders and administration site conditions					
Death	NR	CSR	NR	NR	NR

Pain	BOTH+	NR	NR	NR	NR
Fall	NR	NR	CSR	CSR	CSR
Chest pain	NR	NR	NR	CSR	CSR
Weakness	NR	NR	NR	CSR	NR
Pyrexia	NR	NR	NR	NR	CSR
Groin pain	NR	NR	NR	NR	CSR
Hernia	NR	NR	NR	NR	CSR
Shivering	NR	NR	NR	NR	CSR
Renal and urinary system disorders					
Nephrectomy due to previous renal carcinoma	NR	CSR	NR	NR	NR
Nephrotomy & lithotripsy due to previous nephrolithiasis	NR	CSR	NR	NR	NR
Calculus renal	NR	NR	CSR	CSR	CSR
Bladder neck obstruction	NR	NR	NR	NR	CSR
Urinary incontinence	NR	NR	NR	NR	CSR
Metabolic and nutritional disorder					
Diabetes mellitus	NR	CSR	NR	NR	NR
Metabolic disorder	NR	NR	CSR	NR	NR
Hyperglycaemia	NR	NR	NR	CSR	NR
Hypokalaemia	NR	NR	NR	CSR	NR
Calcinosis	NR	NR	NR	NR	CSR
Diabetes mellitus aggravated	NR	NR	NR	NR	CSR
Diabetes mellitus non-insulin-dependent	NR	NR	NR	NR	CSR
Hypoglycaemia	NR	NR	NR	NR	CSR
Reproductive disorders, female					

Ovary carcinoma & ascites	NR	CSR	NR	NR	NR
Hysterectomy & perineoplasty	NR	CSR	NR	NR	NR
Benign prostatic hyperplasia	NR	NR	NR	CSR	NR
Endometriosis	NR	NR	NR	CSR	CSR
Prostatitis	NR	NR	NR	CSR	CSR
Dysmenorrhoea	NR	NR	NR	NR	CSR
Uterine prolapse	NR	NR	NR	NR	CSR
Endometrial hyperplasia	NR	NR	NR	NR	CSR
Ovarian cyst	NR	NR	NR	NR	CSR
Cervical dysplasia	NR	NR	NR	NR	CSR
Cervical stricture	NR	NR	NR	NR	CSR
Cystocele	NR	NR	NR	NR	CSR
Haemorrhage into ovarian cyst	NR	NR	NR	NR	CSR
Menometrorrhagia	NR	NR	NR	NR	CSR
Vaginal haemorrhage	NR	NR	NR	NR	CSR
Vaginal prolapse	NR	NR	NR	NR	CSR
Cardiovascular disorders					
Mitral lesion	NR	CSR	NR	NR	NR
Atrial fibrillation	NR	NR	CSR	NR	CSR
Myocardial infarction	NR	NR	CSR	CSR	CSR
Cardiac failure congestive	NR	NR	NR	CSR	NR
Angina pectoris	NR	NR	NR	CSR	CSR
Angina pectoris aggravated	NR	NR	NR	CSR	CSR
Angina unstable	NR	NR	NR	CSR	CSR
Coronary artery disease	NR	NR	NR	CSR	NR
Chest pressure sensation	NR	NR	NR	CSR	NR
Congestive cardiac failure	NR	NR	NR	CSR	NR

Cardiac failure	NR	NR	NR	NR	CSR
Arrhythmia	NR	NR	NR	NR	CSR
Atrial flutter	NR	NR	NR	NR	CSR
Myocarditis	NR	NR	NR	NR	CSR
Tachycardia	NR	NR	NR	NR	CSR
Atrial tachycardia	NR	NR	NR	NR	CSR
Bradycardia	NR	NR	NR	NR	CSR
Endocarditis	NR	NR	NR	NR	CSR
Myocardial ischaemia	NR	NR	NR	NR	CSR
Oedema lower limb	NR	NR	NR	NR	CSR
Pericarditis	NR	NR	NR	NR	CSR
Supraventricular tachycardia	NR	NR	NR	NR	CSR
Ventricular extrasystoles	NR	NR	NR	NR	CSR
Ventricular fibrillation	NR	NR	NR	NR	CSR
Musculoskeletal, connective tissue and bone disorders					
Intervertebral disc prolapse	NR	NR	CSR	NR	CSR
Osteoarthritis aggravated	NR	NR	CSR	CSR	CSR
Baker's cyst	NR	NR	CSR	NR	NR
Ligament laxity	NR	NR	CSR	NR	NR
Rheumatoid arthritis	NR	NR	CSR	NR	CSR
Aggravated	NR	NR	NR	NR	NR
Spinal osteoarthritis	NR	NR	CSR	NR	NR
Spondylolisthesis acquired	NR	NR	CSR	NR	NR
Arthritis nos aggravated	NR	NR	NR	CSR	NR
Back pain aggravated	NR	NR	NR	NR	CSR
Back pain	NR	NR	NR	NR	CSR
Bursitis	NR	NR	NR	NR	CSR

Chest wall pain	NR	NR	NR	NR	CSR
Localised osteoarthritis	NR	NR	NR	NR	CSR
Arthralgia	NR	NR	NR	NR	CSR
Rotator cuff syndrome	NR	NR	NR	NR	CSR
Sciatica	NR	NR	NR	NR	CSR
Spondylosis	NR	NR	NR	NR	CSR
Tendonitis	NR	NR	NR	NR	CSR
Arthritis	NR	NR	NR	NR	CSR
Cartilage disorder	NR	NR	NR	NR	CSR
Chondritis	NR	NR	NR	NR	CSR
Juvenile rheumatoid arthritis	NR	NR	NR	NR	CSR
Myositis	NR	NR	NR	NR	CSR
Neck pain	NR	NR	NR	NR	CSR
Psoriatic arthropathy	NR	NR	NR	NR	CSR
Tenosynovitis	NR	NR	NR	NR	CSR
Neoplasms benign and malignant (including cysts and polyps)					
Benign ovarian tumour	NR	NR	CSR	NR	NR
Renal cell carcinoma stage 1	NR	NR	CSR	NR	CSR
Unspecified	NR	NR	NR	NR	NR
Gall bladder cancer	NR	NR	CSR	NR	CSR
Acoustic neuroma	NR	NR	NR	CSR	NR
Hepatic neoplasm malignant	NR	NR	NR	CSR	NR
Uterine fibroids	NR	NR	NR	NR	CSR
Malignant melanoma site/stage unspecified	NR	NR	NR	NR	CSR
Basal cell carcinoma	NR	NR	NR	NR	CSR
Prostate cancer	NR	NR	NR	NR	CSR

Uterine cancer	NR	NR	NR	NR	CSR
Benign neoplasm	NR	NR	NR	NR	CSR
Breast cancer female	NR	NR	NR	NR	CSR
Benign anorectal neoplasm	NR	NR	NR	NR	CSR
Brain neoplasm benign	NR	NR	NR	NR	CSR
Cervical cancer stage 0	NR	NR	NR	NR	CSR
Chronic lymphocytic leukaemia	NR	NR	NR	NR	CSR
Gastric cancer	NR	NR	NR	NR	CSR
Haemangioma of liver	NR	NR	NR	NR	CSR
Hodgkin's disease nos 1	NR	NR	NR	NR	CSR
Laryngeal neoplasm benign	NR	NR	NR	NR	CSR
Leiomyoma	NR	NR	NR	NR	CSR
Lymphoma	NR	NR	NR	NR	CSR
Ovarian cancer	NR	NR	NR	NR	CSR
Rectal cancer	NR	NR	NR	NR	CSR
Urethral cancer nos	NR	NR	NR	NR	CSR
Injury and poisoning					
Cartilage injury	NR	NR	CSR	NR	NR
Fibula fracture	NR	NR	CSR	NR	CSR
Foot fracture	NR	NR	CSR	NR	CSR
Leg fracture	NR	NR	CSR	NR	CSR
Ankle fracture	NR	NR	NR	CSR	CSR
Ligament sprain	NR	NR	NR	CSR	CSR
Road traffic accident	NR	NR	NR	CSR	CSR
Spinal compression fracture	NR	NR	NR	CSR	NR
Wrist fracture	NR	NR	NR	CSR	NR
Concussion	NR	NR	NR	NR	CSR

Limb injury	NR	NR	NR	NR	CSR
Upper limb fracture	NR	NR	NR	NR	CSR
Accident	NR	NR	NR	NR	CSR
Muscle rupture	NR	NR	NR	NR	CSR
Radius fracture	NR	NR	NR	NR	CSR
Rib fracture	NR	NR	NR	NR	CSR
Alcohol poisoning	NR	NR	NR	NR	CSR
Back injury	NR	NR	NR	NR	CSR
Forearm fracture	NR	NR	NR	NR	CSR
Hand fracture	NR	NR	NR	NR	CSR
Humerus fracture	NR	NR	NR	NR	CSR
Injury	NR	NR	NR	NR	CSR
Joint dislocation	NR	NR	NR	NR	CSR
Laceration	NR	NR	NR	NR	CSR
Multiple fractures	NR	NR	NR	NR	CSR
Multiple injuries	NR	NR	NR	NR	CSR
Muscle sprain	NR	NR	NR	NR	CSR
Non-accidental injury	NR	NR	NR	NR	CSR
Spinal fracture	NR	NR	NR	NR	CSR
Tibia fracture	NR	NR	NR	NR	CSR
Whiplash injury	NR	NR	NR	NR	CSR
Vascular disorders					
Brain stem infarction	NR	NR	CSR	NR	NR
Carotid artery stenosis	NR	NR	CSR	CSR	NR
Cerebral ischaemia	NR	NR	CSR	NR	NR
Hypertensive crisis	NR	NR	CSR	NR	NR
Varicose veins	NR	NR	CSR	NR	CSR

Hypertension aggravated	NR	NR	NR	CSR	NR
Hypotension	NR	NR	NR	CSR	NR
Intracranial haemorrhage	NR	NR	NR	CSR	NR
Pulmonary hypertension	NR	NR	NR	CSR	NR
Venous thrombosis deep limb	NR	NR	NR	NR	CSR
Pulmonary embolism	NR	NR	NR	NR	CSR
Aortic aneurysm	NR	NR	NR	NR	CSR
Aortic aneurysm rupture	NR	NR	NR	NR	CSR
Aorto-iliac arterial stenosis	NR	NR	NR	NR	CSR
Cerebral infarction	NR	NR	NR	NR	CSR
Cranial arteritis	NR	NR	NR	NR	CSR
Peripheral vascular disease	NR	NR	NR	NR	CSR
Subarachnoid haemorrhage	NR	NR	NR	NR	CSR
Transient ischaemic attack	NR	NR	NR	NR	CSR
Vasculitis	NR	NR	NR	NR	CSR
Venous thrombosis nos limb	NR	NR	NR	NR	CSR
Investigations					
Blood glucose abnormal	NR	NR	CSR	NR	NR
Blood in stool	NR	NR	CSR	NR	CSR
Endocrine disorders					
Diabetes mellitus inadequate	NR	NR	CSR	NR	NR
Control	NR	NR	NR	NR	NR
Surgical and medical procedures					
Post-operative complications	NR	NR	CSR	NR	NR
Post-operative haemorrhage	NR	NR	NR	CSR	CSR
Abortion induced	NR	NR	NR	NR	CSR
Post-operative pain	NR	NR	NR	NR	CSR

Post-operative wound infection	NR	NR	NR	NR	CSR
Eye disorders					
Blindness nec	NR	NR	NR	CSR	NR
Retinopathy diabetic	NR	NR	NR	CSR	NR
Dacryocystitis	NR	NR	NR	NR	CSR
Iridocyclitis	NR	NR	NR	NR	CSR
Papilloedema	NR	NR	NR	NR	CSR
Retinal detachment	NR	NR	NR	NR	CSR
Blood and lymphatic system disorders					
Secondary anaemia	NR	NR	NR	CSR	NR
Lymphadenitis acute	NR	NR	NR	NR	CSR
Neutropenia	NR	NR	NR	NR	CSR
Ear and labyrinth disorders					
Vertigo	NR	NR	NR	NR	CSR
Vestibular neuronitis	NR	NR	NR	NR	CSR
Meniere's disease	NR	NR	NR	NR	CSR
Deafness	NR	NR	NR	NR	CSR
Hearing impaired	NR	NR	NR	NR	CSR
Vertigo aggravated	NR	NR	NR	NR	CSR
Immune system disorders					
Sarcoidosis	NR	NR	NR	NR	CSR
Hypersensitivity	NR	NR	NR	NR	CSR
Anaphylactic reaction	NR	NR	NR	NR	CSR
Drug hypersensitivity	NR	NR	NR	NR	CSR
Amyloidosis	NR	NR	NR	NR	CSR
Endocrine disorders					
Goitre	NR	NR	NR	NR	CSR

Hyperparathyroidism	NR	NR	NR	NR	CSR
Hyperthyroidism	NR	NR	NR	NR	CSR
Hypothyroidism	NR	NR	NR	NR	CSR
Thyrotoxicosis	NR	NR	NR	NR	CSR
Skin & subcutaneous tissue disorders					
Dermatitis	NR	NR	NR	NR	CSR
Urticaria nos	NR	NR	NR	NR	CSR
Angioneurotic oedema	NR	NR	NR	NR	CSR
Pregnancy, puerperium and perinatal conditions					
Abortion spontaneous	NR	NR	NR	NR	CSR
Abortion	NR	NR	NR	NR	CSR
Total number of serious adverse event terms reported across CSR and publication	19	9	40	53	255
Total number of serious adverse event terms reported in CSR (% of total)	18 (95)	9 (100)	40 (100)	53 (100)	255 (100)
Total number of serious adverse event terms reported in journal publication (% of total)	14 (74)	1 (11)	0 (0)	0 (0)	0 (0)

BOTH+ = 'reported in CSR and the corresponding journal publication and agreed in data'; BOTH- = 'reported in CSR and the corresponding full academic journal publication but disagreed in results'; CSR = 'only reported within the CSR'; Pub = 'only reported in the journal publication (specific events may have been subsumed under other categories in summary reports)'; NR = 'neither reported in the CSR or journals publication'. *Depression: CSR reports (1) event on orlistat; Publication reports (2) events on orlistat arm.

Table 30: Meta-analysis for serious adverse events (SAEs) reported at least once in CSR and corresponding journal publication.

These meta-analysis results are based on a subset of the eligible trials of orlistat and are presented for the purpose of methodological comparison rather than definitive clinical results.

						Pooled estimates		Heterogeneity	
Event	**Document**	**Orlistat events**	**Total randomised to orlistat**	**Placebo events**	**Total randomised to placebo**	**Risk Difference (RD) (accurate to 2 dp's)**	**95% CI (accurate to 2 dp's)**	**I^2**	**P-Value**
Respiratory, Thoracic And Medicinal Disorders									
Adenoidal Hypertrophy	Publication	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA
	CSR	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA
Asthma	Publication	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA
	CSR	2	2001	1	1836	0.00	(0.00, 0.00)	0	0.5254
Asthma aggravated	Publication	0	352	1	181	-0.01	(-0.02, 0.01)	NA	NA
	CSR	1	2268	2	2105	0.00	(0.00, 0.00)	0	0.5541
Nasal Septum Deviation	Publication	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA
	CSR	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA
Nervous System Disorders									
Facial Palsy	Publication	0	352	1	181	-0.01	(-0.02, 0.01)	NA	NA
	CSR	1	2001	2	1836	0.00	(0.00, 0.00)	12.5	0.258
Meningitis Aseptic	Publication	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA
	CSR	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA
Hepato-Biliary Disorders									
Choleithiasis	Publication	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA

	CSR	18	2268	10	2105	0.00	(0.00, 0.01)	6.5	0.3431
Gall Bladder disorder	Publication	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA
	CSR	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA
Infections And Infestations									
Pilonidal Abscess	Publication	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA
	CSR	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA
Pneumonia	Publication	0	352	1	181	-0.01	(-0.02, 0.01)	NA	NA
	CSR	3	2268	2	2105	0.00	(0.00, 0.00)	3.9	0.3634
Psychiatric Disorders									
Depression	Publication	2	352	0	181	0.01	(-0.01, 0.02)	NA	NA
	CSR	2	2001	1	1836	0.00	(0.00, 0.00)	0	0.5254
Gastrointestinal Disorders									
Appendicitis	Publication	1	352	0	181	0.00	(-0.01, 0.01)	NA	NA
	CSR	5	2001	4	1836	0.00	(0.00, 0.00)	0	0.6032
Diarrhoea & Dehydration	Publication	1	174	0	169	0.01	(-0.01, 0.02)	NA	NA
	CSR	1	174	0	169	0.01	(-0.01, 0.02)	NA	NA
General Disorders And Administration Site Conditions									
Pain	Publication	0	352	1	181	-0.01	(-0.02, 0.01)	NA	NA
	CSR	0	352	1	181	-0.01	(-0.02, 0.01)	NA	NA

P-value represents significance of the heterogeneity between studies reporting the event within CSR or journal publication.

NA – Only one study reports the event and therefore can't calculate heterogeneity or P-value.

Table 31: Serious adverse events (SAEs) reported only in CSR.

					Pooled estimates		Heterogeneity	
Event	**Orlistat events**	**Total randomised to orlistat**	**Placebo events**	**Total randomised to placebo**	**Risk Difference (RD)**	**95% CI**	**I^2 (%)**	**P - Value**
Respiratory, Thoracic And Medicinal Disorders								
Bronchospsm aggravated	0	352	1	181	-0.01	(-0.02, 0.01)	NA	NA
Dyspnoea	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Epistaxis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Nasal Polyps	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Oropharyngeal swelling	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Pleural Effusion	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Pleurisy	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Rhinitis Seasonal	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Nervous System Disorders								
Carpal Tunnel Syndrome	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Cerebrovascular Accident	3	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Convulsions	3	2001	0	1836	0.00	(0.00, 0.00)	NA	NA
Demyelination	0	352	1	181	-0.01	(-0.02, 0.01)	NA	NA
Dizziness (Exc vertigo)	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Entrapment Neuropathy	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA

Epilepsy	1	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Haemorrhagic stroke	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Headache	4	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Loss of Consciousness	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Migraine	1	1916	2	1924	0.00	(0.00, 0.00)	0	0.3891
Migraine aggravated	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Multiple Sclerosis	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Multiple Sclerosis aggravated	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Obstructive sleep apnoea syndrome	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Radiculopathy	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Sleep Apnoea Syndrome	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Spinal cord compression	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Spinal Stenosis	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Syncope	2	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Trigeminal Neuralgia	0	190	1	182	-0.01	(-0.02, 0.01)	NA	NA
Hepato-Biliary Disorders								
Bile Duct Obstruction	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Bile Duct Stone	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Biliary Colic	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Cholangitis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Cholecystitis	8	1916	7	1924	0.00	(0.00, 0.00)	0	0.5139
Gall Bladder disease	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA

Gall Bladder pain	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Hepatotoxicity	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Infections And Infestations								
Abscess	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Acute Exacerbation of Chronic	0	190	1	182	-0.01	(-0.02, 0.01)	NA	NA
Ano-Rectal Infection	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Bacterial Infection	2	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Bronchitis	1	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Bronchitis acute	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Cellulites	1	1916	1	1924	0.00	(0.00, 0.00)	0.2	0.3168
Cervicitis	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Cholecystitis Acute	1	1839	1	1837	0.00	(0.00, 0.00)	0	0.3217
Empyema	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Erysipelas	4	1839	7	1837	0.00	(-0.01, 0.00)	57.8	0.1238
Eye Abscess	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Fungal Infection	1	190	0	182	0.01	(-0.01, 0.02)	NA	NA
Gastroenteritis	3	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Gastroenteritis Helicobacter	2	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Gastroenteritis Salmonella	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Haemorrhagic fever	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Hepatitis B	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Herpes Zoster	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Infected skin Ulcer	0	267	1	269	0.00	(-0.01, 0.01)	NA	NA
Localised infection	1	1649	2	1655	0.00	(0.00, 0.00)	NA	NA

Meningitis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Meningitis bacterial	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Ovarian Abscess	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Pyelonephritis	4	1916	3	1924	0.00	(-0.01, 0.01)	NA	NA
Pyelonephritis Acute	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Salpingitis	3	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Sepsis	0	190	1	182	-0.01	(-0.02, 0.01)	NA	NA
Sinusitis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Staphylococcal infection	0	267	1	269	0.00	(-0.01, 0.01)	NA	NA
Tuberculosis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Upper Respiratory Tract	0	267	1	269	0.00	(-0.01, 0.01)	NA	NA
Urinary Tract Infection	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Urosepsis	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Psychiatric Disorders								
Alcoholic Withdrawal symptoms	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Alcoholism	4	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Anxiety disorder	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Completed suicide	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Suicide ideation	1	362	0	181	0.00	(-0.01, 0.01)	NA	NA
Suicide attempt	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Gastrointestinal Disorders								
Abdominal pain	4	1649	3	1655	0.00	(0.00, 0.00)	NA	NA

Abdominal Pain Upper	2	1649	3	1655	0.00	(0.00, 0.00)	NA	NA
Anal Fissure	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Appendicitis Perforated	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Cholecistectomy due to chronic Cholelithiasis	1	174	0	169	0.01	(-0.01, 0.02)	NA	NA
Colitis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Decreased Defecation	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Diverticulitis	6	1839	5	1837	0.00	(0.00, 0.00)	0	0.3853
Diverticulum Intestinal	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Diverticulum Nos	2	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Duodenal ulcer	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Duodenitis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Enteritis	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Flatulence	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Gastric Ulcer	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Gastritis	1	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Gastrointestinal haemorrhage	2	1916	0	1924	0.00	(0.00, 0.00)	0	0.4682
Haemorrhoids	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Ileus	3	1839	1	1837	0.00	(0.00, 0.00)	0	0.4726
Impaired Gastric Emptying	0	267	1	269	0.00	(-0.01, 0.01)	NA	NA
Incisional hernia	2	1649	1	1655	0.00	(0.00, 0.00)	NA	NA

Inguinal hernia	2	1839	2	1837	0.00	(0.00, 0.00)	0	0.3524
Liquid Stools	5	2458	2	2287	0.00	(0.00, 0.00)	0	0.8666
Oesophagitis	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Pancreatitis	2	1649	3	1655	0.00	(0.00, 0.00)	NA	NA
Pancreatitis acute	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Umbilical Hernia	2	1839	0	1837	0.00	(0.00, 0.00)	0	0.432
General Disorders And Administration Site Conditions								
Chest Pain	9	1916	15	1924	0.00	(-0.01, 0.00)	0	0.4705
Death	0	174	1	169	-0.01	(-0.02, 0.01)	NA	NA
Fall	2	2106	4	2106	0.00	(0.00, 0.00)	0	0.6017
Groin pain	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Hernia	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Pyrexia	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Shivering	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Weakness	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Renal And Urinary System Disorders								
Bladder Neck Obstruction	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Calculus Renal	2	2106	4	2106	0.00	(0.00, 0.00)	44.5	0.1652
Nephrectomy due to previous renal carcinoma	1	174	0	169	0.01	(-0.01, 0.02)	NA	NA
Nephrotomy & lithotripsy due to previous nephrolithiasis	1	174	0	169	0.01	(-0.01, 0.02)	NA	NA
Urinary Incontinence	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Metabolic And Nutritional Disorder								

Calcinosis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Diabetes Mellitus	0	174	1	169	-0.01	(-0.02, 0.01)	NA	NA
Diabetes Mellitus Aggravated	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Diabetes Mellitus Non-Insulin-Dependent	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Hyperglycaemia	0	267	1	269	0.00	(-0.01, 0.01)	NA	NA
Hypoglycaemia	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Hypokalaemia	0	267	1	269	0.00	(-0.01, 0.01)	NA	NA
Metabolic Disorder	0	190	1	182	-0.01	(-0.02, 0.01)	NA	NA
Reproductive Disorders, Female								
Benign Prostatic Hyperplasia	0	267	1	269	0.00	(-0.01, 0.01)	NA	NA
Cervical Dysplasia	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Cervical Structure	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Cystocele	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Dysmenorrhoea	2	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Endometrial Hyperplasia	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Endometriosis	3	1916	2	1924	0.01	(-0.01, 0.01)	0	0.5741
Haemorrhage into ovarian Cyst	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Hysterectomy & perineoplasty	0	174	1	169	-0.01	(-0.02, 0.01)	NA	NA
Menometorrhagia	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Ovarian Cyst	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Ovary carcinoma &	1	174	0	169	0.01	(-0.01, 0.02)	NA	NA

Ascites								
Prostatitis	0	1916	2	1924	0.00	(0.00, 0.00)	0	0.4707
Uterine Prolapse	2	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Vaginal Haemorrhage	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Vaginal Prolapse	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Cardiovascular Disorders								
Angina Pectoris	10	1916	6	1924	0.00	(0.00, 0.01)	0	0.674
Angina Pectoris Aggravated	2	1916	3	1924	0.00	(0.00, 0.00)	0	0.8944
Angina Unstable	1	1916	2	1924	0.00	(0.00, 0.00)	65.4	0.0892
Arrhythmia	2	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Atrial Fibrillation	7	1839	3	1837	0.00	(0.00, 0.01)	0	0.6482
Atrial Flutter	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Atrial Tachycardia	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Bradycardia	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Cardiac Failure	1	1649	5	1655	0.00	(-0.01, 0.00)	NA	NA
Cardiac Failure Congestive	3	267	0	269	0.00	(0.00, 0.03)	NA	NA
Chest pressure sensation	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Congestive Cardiac Failure	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Coronary Artery Disease	1	267	1	269	0.00	(-0.01, 0.01)	NA	NA
Endocarditis	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Mitral lesion	0	174	1	169	-0.01	(-0.02, 0.01)	NA	NA
Myocardial	14	2106	9	2106	0.00	(0.00, 0.01)	11.5	0.323

Infraction								
Myocardial Ischaemia	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Myocarditis	2	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Oedema lower limb	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Pericarditis	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Supraventricular Tachycardia	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Tachycardia	2	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Ventricular Extrasystoles	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Ventricular Fibrillation	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Musculoskeletal, Connective Tissue And Bone Disorders								
Arthralgia	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Arthritis	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Arthiritis NOS Aggravated	0	267	1	269	0.00	(-0.01, 0.01)	NA	NA
Back pain	3	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Back Pain Aggravated	2	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Bakers Cyst	0	190	1	182	-0.01	(-0.02, 0.01)	NA	NA
Bursitis	0	1649	3	1655	0.00	(0.00, 0.00)	NA	NA
Cartilage Disorder	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Chest wall pain	2	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Chondritis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Intervertebral Disc prolapse	2	1839	3	1837	0.00	(0.00, 0.00)	59.9	0.1141

Juvenile Rheumatoid arthritis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Localised Osteoarthritis	2	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Myositis	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Neck Pain	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Osteoarthritis aggravated	5	2106	3	2106	0.00	(0.00, 0.00)	0	0.3749
Psoriatic Arthropathy	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Rheumatoid arthritis	1	1839	1	1837	0.00	(0.00, 0.00)	3.6	0.3084
Rotator Cuff syndrome	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Sciatica	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Spinal Osteoarthritis	1	190	0	182	0.01	(-0.01, 0.02)	NA	NA
Spondylolisthesis Aquired	1	190	0	182	0.01	(-0.01, 0.02)	NA	NA
Spondylosis	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Tendonitis	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Tenosynovitis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Neoplasm's Benign And Malignant								
Acoustic Neuroma	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Basal Cell Carcinoma	2	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Benign Anorectal Neoplasm	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Benign Ovarian Tumour	2	190	0	182	0.01	(-0.01, 0.03)	NA	NA
Brain Neoplasm	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA

Benign								
Breast Cancer Female	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Cervical Cancer stage 0	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Chronic Lymphocytic Leukaemia	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Gall Bladder Cancer	0	1839	2	1837	0.00	(0.00, 0.00)	0	0.4138
Gastric Cancer	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Haemangioma of liver	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Hepatic Neoplasm malignant	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Hodgkin's Disease Nos 1	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Laryngeal neoplasm benign	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Leiomyoma	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Lymphoma	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Malignant Melanoma site/stage unspecified	1	1649	3	1655	0.00	(0.00, 0.00)	NA	NA
Ovarian Cancer	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Prostate cancer	2	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Rectal cancer	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Renal Cell carcinoma stage 1	2	1839	2	1837	0.00	(-0.02, 0.01)	NA	NA
Urethral Cancer Nos	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA

Uterine Cancer	2	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Uterine Fibroids	4	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Injury And Poisoning								
Accident	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Alcohol poisoning	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Ankle Fracture	0	1916	2	1924	0.00	(0.00, 0.00)	0	0.4707
Back injury	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Cartilage Injury	1	190	0	182	0.01	(-0.01, 0.02)	NA	NA
Concussion	1	1649	5	1655	0.00	(-0.01, 0.00)	NA	NA
Fibula Fracture	1	1839	1	1837	0.00	(0.00, 0.00)	3.6	0.3084
Foot fracture	0	1839	3	1837	0.00	(0.00, 0.00)	0	0.5133
Forearm fracture	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Hand fracture	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Humerus fracture	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Injury	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Joint Dislocation	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Laceration	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Leg fracture	3	1839	2	1837	0.00	(0.00, 0.00)	0	0.4407
Ligament Sprain	2	1916	1	1924	0.00	(0.00, 0.00)	8	0.297
Limb injury	2	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Multiple fractures	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Multiple injuries	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Muscle rupture	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Muscle Sprain	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Non-accidental injury	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Radius fracture	2	1649	0	1655	0.00	(0.00, 0.00)	NA	NA

Rib fracture	2	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Road traffic accident	3	1916	1	1924	0.00	(0.00, 0.00)	0	0.508
Spinal compression fracture	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Spinal fracture	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Tibia Fracture	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Upper Limb Fracture	1	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Whiplash Injury	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Wrist Fracture	1	267	0	269	0.00	(-0.01, 0.01)	NA	1
Vascular Disorders								
Aortic Aneurysm	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Aortic Aneurysm Rupture	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Aorto-Iliac Arterial Stenosis	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Brain Stem Infarction	1	190	0	182	0.01	(-0.01, 0.02)	NA	NA
Carotid Artery Stenosis	5	457	0	451	0.01	(0.00, 0.02)	0	0.3644
Cerebral Infarction	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Cerebral Ischaemia	1	190	0	182	0.01	(-0.01, 0.02)	NA	NA
Cranial Arthritis	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Hypertension Aggravated	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Hypertensive Crisis	0	190	1	182	-0.01	(-0.02, 0.01)	NA	NA
Hypotension	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Intracranial haemorrhage	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Peripheral Vascular	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA

disease								
Pulmonary Embolism	1	1649	4	1655	0.00	(0.00, 0.00)	NA	NA
Pulmonary hypertension	0	267	1	269	0.00	(-0.01, 0.01)	NA	NA
Subarachnoid haemorrhage	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Transient Ischaemic attack	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Varicose Veins	2	1839	0	1837	0.01	(0.01, 0.02)	NA	NA
Vasculitis	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Venous Thrombosis deep limb	2	1649	4	1655	0.00	(0.00, 0.00)	NA	NA
Venous Thrombosis Nos Limb	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Investigations								
Blood Glucose Abnormal	0	190	1	182	-0.01	(-0.02, 0.01)	NA	NA
Blood in stool	0	1839	2	1837	0.00	(0.00, 0.00)	0	0.4138
Endocrine Disorders								
Diabetes Mellitus Inadequate	1	190	0	182	0.01	(-0.01, 0.02)	NA	NA
Goitre	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Hyperparathyroidism	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Hyperthyroidism	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Hypothyroidism	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Thyrotoxicosis	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Surgical And Medical Procedures								
Abortion induced	2	1649	0	1655	0.00	(0.00, 0.00)	NA	NA

Post-operative complications	0	190	1	182	-0.01	(-0.02, 0.01)	NA	NA
Post-operative Haemorrhage	1	1916	2	1924	0.00	(0.00, 0.00)	0	0.3891
Post-operative pain	2	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Post-operative wound infection	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Eye Disorders								
Blindness Nec	0	267	1	269	0.00	(-0.01, 0.01)	NA	NA
Dacryocystitis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Iridocyclitis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Papilloedema	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Retinal detachment	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Retinopathy diabetic	0	267	1	269	0.00	(-0.01, 0.01)	NA	NA
Blood And Lymphatic System Disorders								
Lymphadenitis Acute	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Neutropenia	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Secondary Anaemia	1	267	0	269	0.00	(-0.01, 0.01)	NA	NA
Ear And Labyrinth Disorders								
Deafness	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Hearing impaired	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Menieres disease	2	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Vertigo	6	1649	1	1655	0.00	(0.00, 0.01)	NA	NA
Vertigo aggravation	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Vestibular neuronitis	3	1649	3	1655	0.00	(0.00, 0.00)	NA	NA
Immune System Disorders								
Amyloidosis	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA

Anaphylactic Reaction	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Drug Hypersensitivity	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Hypersensitivity	2	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Sarcoidosis	2	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Endocrine Disorders								
Diabetes Mellitus Inadequate	1	190	0	182	0.01	(-0.01, 0.02)	NA	NA
Goitre	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Hyperparathyroidism	0	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Hyperthyroidism	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Hypothyroidism	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Thyrotoxicosis	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Skin & Subcutaneous Tissue Disorders								
Angioneurotic Oedema	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Dermatitis	1	1649	1	1655	0.00	(0.00, 0.00)	NA	NA
Urticaria Nos	0	1649	2	1655	0.00	(0.00, 0.00)	NA	NA
Pregnancy, Puerperium And Prenatal Conditions								
Abortion	1	1649	0	1655	0.00	(0.00, 0.00)	NA	NA
Abortion spontaneous	2	1649	1	1655	0.00	(0.00, 0.00)	NA	NA

P-value represents significance of the heterogeneity between studies reporting the event within CSR only.

NA – Only one study reports the event and therefore can't calculate heterogeneity or P-value.

Table 32: Serious adverse events (SAEs) reported only in journal publication.

					Pooled estimates		Heterogeneity	
Event	**Orlistat events**	**Total randomised to orlistat**	**Placebo events**	**Total randomised to placebo**	**Risk Difference (RD)**	**95% CI**	**I^2 (%)**	**P - Value**
Nervous System Disorders								
Encephalomyelitis	0	352	1	181	-0.01	(-0.02, 0.01)	NA	NA

P-value represents significance of the heterogeneity between studies reporting the event within CSR or journal publication.
NA – Only one study reports the event and therefore can't calculate heterogeneity or P-value.

Appendix C – Copy of Survey Questionnaire from Chapter 5

A short survey of current practice: data-basing adverse events in UKCRC registered CTUs.

This is a very short survey and comprises of only nine questions and should take no longer than 5 minutes to complete. The aim is to identify the current practice in registered CTUs when data-basing AEs, and understand the use of those databases. These questions will inform my PhD simulation work and the results of the survey will be shared with the UKCRC Registered CTU Network at a Statistics Operational Group network meeting.

1. How are adverse events (AEs) data-based in your CTU?

    - Within a database specific to a single trial (Go to Question 6)
    - Within a database holding multiple trials related by disease/condition/treatment (Go to Question 2)
    - Within a database holding multiple trials of a diverse range of disease/condition/treatment (Go to Question 2)
    - Other, please specify……..

    ……
    ……
    ……..

2. Please describe how the database is used within the CTU?

    - Held by sponsor or other, with no use by the CTU
    - Monitoring for ongoing trials
    - Signal detection
    - Planning for new trials
    - Other, please specify…….

……
……
……

3. Approximately how many trials have contributed to the database?

.....
.....

4. Approximately how many AEs are contained in the database?

.....
.....

5. If the CTU does use a central database (including for the purposes of reconciliation), based on your experience please briefly describe:

    a. what the advantage for using this database are?

……
……
……
…….

    b. what the disadvantages for using this database are?

……
……
……
……

6. Have you considered using a central database containing AEs across multiple trials?

    - Yes, could see no benefit
    - Yes, future plans to do this
    - No, not aware of considering this

Comments:……
……
……
……

7. During safety monitoring what external data to the trial, have you used?

    - Central AE database
    - Published trial reports and studies including systematic reviews.
    - Clinical Practice Research Data-link (CPRD)/ General Practice Research Database (GPRD).
    - The Health Improvement Network (THIN)
    - Medical and Healthcare products Regulatory Agency (MHRA) yellow card data.
    - Medicines Monitoring Unit (MEMO)
    - Other please specify……
      ……
      ……
      ……

Comments on the value of approaches indicated:

……

……

……

……

8. Which of the following methods of signal detection have you used?
    - None
    - Gamma Poisson Shrinker (GPS)
    - Bayesian Confidence Propagation Neural Network (BCPNN)
    - Proportional Reporting Ratio (PRR)
    - Reporting Odds Ratio (ROR)
    - Other please specify…….
      ……
      ……
      ……

9. Would you be interested in exploring this topic further in a future UKCRC statistics operational group network meeting? [Y/N]

10. Would you be willing to present/talk? [Y/N]

Name of trials unit:

Person completing survey:

Role within the trials unit:

Years' experience in clinical research:

Email address:

Appendix D – Further Results from Chapter 7

SAS Code for simulation model:

```
%macro event;
 %do k=6   %to   30 ; p&k=0.0004            ; %end;
 %do l=31  %to   60 ; p&l=0.000322          ; %end;
 %do m=61  %to   90 ; p&m=0.0002            ; %end;
 %do n=91  %to  120 ; p&n=0.00004           ; %end;
 %do o=121 %to  150 ; p&o=0.000025          ; %end;
%mend event;

data event;
 p1=0.0004 ;  p2=0.0004 ;  p3=0.0004 ;  p4=0.0004 ;  p5=0.0004 ;
%event;
run;

%macro sim(froms,tos,drugnum,eventnum,signum,out);
* froms: simulation start point;
* tos: simulation end point;
* drugnum: number of drugs;
* eventnum: number of adverse events;
* signum: number of adverse drug reactions(true signals) per each drug;
* out: name of dataset;
proc datasets;
  delete _sim;
quit;
%do h=&froms. %to &tos.;
data nadrdata(keep=drugno eventno a tn);
  set event;
     array p{&eventnum.} p1-p&eventnum. ;
   %do i=1 %to &drugnum.;
      %let seed2 = %eval(&h.*&i.*&eventnum.*2 );
       drugno=%eval(&i);
       if                          1 <=drugno<=((15/30)*%eval(&drugnum))
                                                          then tn= 20000;
       if ((15/30)*%eval(&drugnum)) < drugno<=((22.5/30)*%eval(&drugnum))
                                                          then tn= 75000;
       if ((22.5/30)*%eval(&drugnum)) <
drugno<=((27.5/30)*%eval(&drugnum))
                                                          then tn= 150000;
       if ((27.5/30)*%eval(&drugnum)) < drugno<=(        %eval(&drugnum))
                                                          then tn= 300000;
         %do j=1 %to &eventnum.;
            eventno=%eval(&j);
            a=ranpoi(&seed2+&j,tn*p[&j]);
            output;
         %end;
   %end;
run;

data sig1(keep=drugno signo1-signo&signum.);
   array f{&eventnum.} f1-f&eventnum. ;
   array sig{&signum.} signo1-signo&signum.;
   %do i=1 %to &drugnum.;
       drugno=%eval(&i);
      %let seed3 = %eval(&h.*&i.*&eventnum.*3 );
         do g=1 to &signum.;
```

```
            sig[g]=0;
         end;
         do h=1 to &eventnum.;
            f[h]=0;
         end;
         do s=1 to 8;
            y=ceil(ranuni(&seed3.)*75);
            if f[y] = 0 then do;
               sig[s]=y;
               f[y]=1;
               end;
            else do;
               s=s-1;
               end;
         end;
         do t=9 to 16;
            y=75+ceil(ranuni(&seed3.)*75);
            if f[y] = 0 then do;
               sig[t]=y;
               f[y]=1;
               end;
            else do;
               t=t-1;
               end;
         end;
         output;
    %end;
 run;
proc transpose data=sig1 out=sig2(rename=(col1=eventno));
  by drugno;
run;
data sig3(keep=drugno eventno risk);
  set sig2;
    by drugno;
        count=1;
        if first.drugno then rsig=0;
      rsig + count;
      if rsig= 1 or rsig= 5 or rsig= 9 or rsig=13 or rsig=17 or
         rsig=21 or rsig=25 or rsig=29 or rsig=33 or rsig=37 or
         rsig=41 or rsig=45 or rsig=49 or rsig=53 or rsig=57 then risk=2;
      if rsig= 2 or rsig= 6 or rsig=10 or rsig=14 or rsig=18 or
         rsig=22 or rsig=26 or rsig=30 or rsig=34 or rsig=38 or
         rsig=42 or rsig=46 or rsig=50 or rsig=54 or rsig=58 then risk=3;
      if rsig= 3 or rsig= 7 or rsig=11 or rsig=15 or rsig=19 or
         rsig=23 or rsig=27 or rsig=31 or rsig=35 or rsig=39 or
         rsig=43 or rsig=47 or rsig=51 or rsig=55 or rsig=59 then risk=5;
      if rsig= 4 or rsig= 8 or rsig=12 or rsig=16 or rsig=20 or
         rsig=24 or rsig=28 or rsig=32 or rsig=36 or rsig=40 or
         rsig=44 or rsig=48 or rsig=52 or rsig=56 or rsig=60 then risk=10;
run;
data sig4;
  set sig3;
    if _n_=1 then set event;
run;
```

```
proc sort data=nadrdata out=sig5(keep=drugno tn) nodupkey;
  by drugno tn;
run;
data sig6;
  merge sig4 sig5;
  by drugno;
run;
data adrdata(keep=drugno eventno adr signal risk);
  set sig6;
    %let seed4 = %eval(&h.*&drugnum.*&eventnum.*4 );
    array p{&eventnum.} p1-p&eventnum. ;
    adr=ranpoi(&seed4,risk*tn*p[eventno]);
      signal=1;
run;
proc sort data=adrdata;
  by drugno eventno;
run;
data _sim0;
  merge nadrdata adrdata;
  by drugno eventno;
run;
data sim&h.(drop=tn adr);
  set _sim0;
   simno=%eval(&h);
   if signal=1 then a=adr;
   else             signal=0;
   if a=0 and signal=0  then delete;
   if risk=. then risk=1;
run;
proc append base=_sim data=sim&h.;
run;
proc datasets;
  delete nadrdata adrdata sig1-sig6 _sim0 sim&h;
quit;
%end;
data _sim1(keep=simno drugno ndrug);
  set _sim;
    by simno drugno;
    if first.drugno then ndrug=0;
    ndrug + a;
    if last.drugno then output;
run;
proc sort data=_sim out=_sim2;
  by simno eventno;
run;
data _sim3(keep=simno eventno nevent);
  set _sim2;
    by simno eventno;
    if first.eventno then nevent=0;
    nevent + a;
    if last.eventno then output;
run;
data _sim4(keep=simno n);
  set _sim2;
```

```
    by simno;
    if first.simno then n=0;
    n + a;
    if last.simno then output;
run;
data _sim5;
  merge _sim2 _sim4;
  by simno;
run;
data _sim6;
  merge _sim5 _sim3;
  by simno eventno;
run;
proc sort data=_sim6;
  by simno drugno;
run;
data &out;
  merge _sim6 _sim1;
  by simno drugno;
run;
proc datasets;
  delete _sim _sim1-_sim6;
quit;
%mend;

%sim(1,5,60,150,16,sim);
data sim;
  set sim;
  where a ge 1;
run;
```

Table 33: Data used for parameter selection in simulation model.

		AEs by background incidence[†]		
Substance	**Total number of different AEs reported over 1 year period (2012-2013)**	**Common events**	**Rare events**	**Number of prescriptions over 1 year period (2012-2013)**
Paclitaxel	236	222	11	105,474
Octocog alfa	76	47	12	1,089
Laronidase	51	40	0	2,009[#]
Orlistat	74	38	0	137,376
Leflunomide	438	228	132	43,824
Nelarabine	66	60	2	1,689
Fluticasone	16	9	5	131,184
Bevacizumab	99	61	2	95,668
Memantine	75	54	3	94,812
Rasagiline	47	41	0	972
Sugammadex	20	12	0	258[#]
Carglumic acid	2	2	0	646[#]
Mycophenolic acid	512	396	0	841[#]
Cetrorelix	10	10	0	186
Betaine	18	18	0	4,620
Docetaxel	1211	770	0	201,529
Parecoxib	78	51	6	10,668[#]
Emtricitabine	49	32	0	36,002
Repaglinide	144	18	78	15,024

Substance	Total number of different AEs reported over 1 year period (2012-2013)	Common events	Rare events	Number of prescriptions over 1 year period (2012-2013)
Rivastigmine	460	292	84	61,020
Toremifene	33	23	8	36,651
Degarelix	94	73	0	3,516
Teriparatide	52	44	7	38,588
Enfuvirtide	54	34	0	11,214*
Alipogene tiparvovec	38	37	0	2,511
Topotecan	192	144	30	32,407
Mecasermin	117	107	0	24,068
Axitinib	76	60	0	4,573
Interferon alfa-2b	244	155	54	10,863
Fosaprepitant	96	40	40	8,988*
Metformin hydrochloride, linagliptin	22	15	6	233,350
Palifermin	31	19	0	3,598*
Levetiracetam	576	448	120	283,872
Bimatoprost	41	40	0	185,100
Lutropin alfa	23	11	4	1,087
Galsulfase	44	25	0	3,411
Nepafenac	30	24	0	8,464
Belatacept	298	295	0	7,847*
Saxagliptin	25	21	3	70,356
Strontium ranelate	100	20	8	60,240
A/vietnam/1203/2004 (h5n1)	51	40	11	1,089
Peginterferon alfa-2b	596	466	72	84,573

Substance	Total number of different AEs reported over 1 year period (2012-2013)	Common events	Rare events	Number of prescriptions over 1 year period (2012-2013)
Dabigatran	66	48	3	26,640
Fenofibrate, pravastatin	85	63	14	181,964
A/indonesia/05/2005	47	31	9	423[#]
Ribavirin	1742	1337	215	48,414
Desirudin	42	33	2	1,443[#]
Conestat alfa	10	10	0	753[#]
Basiliximab	33	0	0	571
Fosamprenavir	39	16	1	1,248
Thiotepa	125	125	0	72,588
Bortezomib	659	265	198	44,342
Iloprost	32	22	0	643
Nelfinavir	52	19	12	1,644
Crizotinib	51	26	0	5,843[#]
Rivaroxaban	88	63	4	7,704
Ipilimumab	166	138	0	57,082
Trabectedin	66	51	0	21,204
Vemurafenib	59	50	0	15,643
Varicella vaccine (live)	26	10	1	8,421
Mean (Range)	150 (10, 1742)	-	-	53,262 (186, 283,872)
Source used	PROTECT ADR database	PROTECT ADR database		Health & social care information centre (HSCIC)[*]

[‡] Background incidences for AEs were coded in the database and separated into either common (including Uncommon, common and very common) or rare (including rare and very rare) events. Some events missing related to those that were not coded with a frequency.

The remaining events related to those that were unknown in classification

*Drugs we not centrally licensed in the UK so we approximated these prescription rates.

*Data here was from NHS GPS and pharmacies across the UK and published monthly for all UK licensed and dispensed drugs.

Table 34: Simulated scenarios to investigate rare events

					$PRR_{\alpha 2.5} > 1$			$IC_{\alpha 2.5} > 0$			$GPS_{\alpha 5} > 2$		
Scenario*	**$RR_{ij}^{s'}$**	**Exposure rates ¥**	**No. of drugs (D_j)**	**Event Incidence (E_i)**	**Sen. (SD)**	**FDR (SD)**	**PPV (SD)**	**Sen. (SD)**	**FDR (SD)**	**PPV (SD)**	**Sen. (SD)**	**FDR (SD)**	**PPV (SD)**
I_i = 1/250													
1	1.2	20,000 75,000 150,000 300,000	30 15 10 5	1/250	0.289 (0.025)	0.528 (0.031)	0.472 (0.031)	0.197 (0.022)	0.229 (0.025)	0.771 (0.025)	0.080 (0.019)	0.081 (0.024)	0.919 (0.024)
2	1.5	20,000 75,000 150,000 300,000	30 15 10 5	1/250	0.317 (0.023)	0.485 (0.030)	0.515 (0.030)	0.216 (0.021)	0.208 (0.022)	0.792 (0.022)	0.141 (0.021)	0.068 (0.023)	0.932 (0.023)
3	2.0	20,000 75,000 150,000 300,000	30 15 10 5	1/250	0.716 (0.020)	0.347 (0.030)	0.653 (0.030)	0.591 (0.017)	0.152 (0.021)	0.848 (0.021)	0.248 (0.020)	0.062 (0.020)	0.938 (0.020)
4	3.0	20,000 75,000 150,000 300,000	30 15 10 5	1/250	0.956 (0.018)	0.249 (0.028)	0.751 (0.028)	0.941 (0.015)	0.070 (0.020)	0.930 (0.020)	0.821 (0.014)	0.048 (0.017)	0.952 (0.017)
5	4.0	20,000 75,000 150,000 300,000	30 15 10 5	1/250	0.965 (0.015)	0.187 (0.027)	0.813 (0.027)	0.952 (0.012)	0.025 (0.019)	0.975 (0.019)	0.896 (0.011)	0.022 (0.018)	0.978 (0.018)

Scenario*	$RR_{ij}^{s'}$	Exposure rates ¥	No. of drugs (D_j)	Event Incidence (E_i)	Sen. (SD)	FDR (SD)	PPV (SD)	Sen. (SD)	FDR (SD)	PPV (SD)	Sen. (SD)	FDR (SD)	PPV (SD)
6	5.0	20,000 75,000 150,000 300,000	30 15 10 5	1/250	0.974 (0.015)	0.135 (0.024)	0.865 (0.024)	0.968 (0.011)	0.005 (0.016)	0.995 (0.016)	0.951 (0.012)	0.002 (0.015)	0.998 (0.015)
7	7.5	20,000 75,000 150,000 300,000	30 15 10 5	1/250	0.985 (0.013)	0.112 (0.021)	0.888 (0.021)	0.979 (0.008)	0.002 (0.014)	0.998 (0.014)	0.968 (0.007)	0.0007 (0.016)	0.999 (0.016)
8	10.0	20,000 75,000 150,000 300,000	30 15 10 5	1/250	0.998 (0.011)	0.092 (0.020)	0.908 (0.020)	0.998 (0.007)	0.0009 (0.013)	0.999 (0.013)	0.985 (0.005)	0.0003 (0.011)	1.000 (0.011)
I_i = 1/500													
9	1.2	20,000 75,000 150,000 300,000	30 15 10 5	1/500	0.264 (0.023)	0.509 (0.030)	0.491 (0.030)	0.180 (0.020)	0.214 (0.023)	0.786 (0.023)	0.062 (0.017)	0.061 (0.019)	0.939 (0.019)
10	1.5	20,000 75,000 150,000 300,000	30 15 10 5	1/500	0.298 (0.022)	0.465 (0.028)	0.535 (0.028)	0.207 (0.019)	0.203 (0.021)	0.797 (0.021)	0.132 (0.016)	0.059 (0.018)	0.941 (0.018)

Scenario*	$RR^{s/}_{ij}$	Exposure rates ¥	No. of drugs (D_j)	Event Incidence (E_i)	Sen. (SD)	FDR (SD)	PPV (SD)	Sen. (SD)	FDR (SD)	PPV (SD)	Sen. (SD)	FDR (SD)	PPV (SD)
11	2.0	20,000 75,000 150,000 300,000	30 15 10 5	1/500	0.675 (0.020)	0.343 (0.029)	0.657 (0.029)	0.564 (0.019)	0.120 (0.017)	0.880 (0.017)	0.213 (0.014)	0.050 (0.018)	0.950 (0.018)
12	3.0	20,000 75,000 150,000 300,000	30 15 10 5	1/500	0.920 (0.017)	0.238 (0.027)	0.762 (0.027)	0.893 (0.015)	0.051 (0.015)	0.949 (0.015)	0.804 (0.014)	0.035 (0.011)	0.965 (0.011)
13	4.0	20,000 75,000 150,000 300,000	30 15 10 5	1/500	0.946 (0.014)	0.172 (0.024)	0.828 (0.024)	0.941 (0.011)	0.010 (0.012)	0.990 (0.012)	0.871 (0.010)	0.009 (0.009)	0.991 (0.009)
14	5.0	20,000 75,000 150,000 300,000	30 15 10 5	1/500	0.961 (0.011)	0.116 (0.021)	0.884 (0.021)	0.948 (0.008)	0.003 (0.011)	0.997 (0.011)	0.932 (0.007)	0.0006 (0.008)	0.999 (0.008)
15	7.5	20,000 75,000 150,000 300,000	30 15 10 5	1/500	0.981 (0.009)	0.100 (0.018)	0.900 (0.018)	0.972 (0.006)	0.0009 (0.010)	0.999 (0.010)	0.969 (0.006)	0.0001 (0.008)	1.000 (0.008)
16	10.0	20,000 75,000 150,000 300,000	30 15 10 5	1/500	0.994 (0.007)	0.084 (0.015)	0.916 (0.015)	0.995 (0.005)	0.0007 (0.010)	0.999 (0.010)	0.983 (0.007)	0.00006 (0.008)	1.000 (0.008)

Scenario*	$RR^{s'}_{ij}$	Exposure rates ¥	No. of drugs (D_j)	Event Incidence (E_i)	Sen. (SD)	FDR (SD)	PPV (SD)	Sen. (SD)	FDR (SD)	PPV (SD)	Sen. (SD)	FDR (SD)	PPV (SD)
I_i = 1/1000													
17	1.2	20,000 75,000 150,000 300,000	30 15 10 5	1/1000	0.237 (0.020)	0.482 (0.025)	0.518 (0.025)	0.154 (0.017)	0.196 (0.021)	0.804 (0.021)	0.031 (0.016)	0.035 (0.017)	0.965 (0.017)
18	1.5	20,000 75,000 150,000 300,000	30 15 10 5	1/1000	0.275 (0.017)	0.438 (0.025)	0.562 (0.025)	0.181 (0.014)	0.185 (0.016)	0.815 (0.016)	0.101 (0.014)	0.033 (0.018)	0.967 (0.018)
19	2.0	20,000 75,000 150,000 300,000	30 15 10 5	1/1000	0.652 (0.014)	0.316 (0.024)	0.684 (0.024)	0.538 (0.010)	0.102 (0.014)	0.898 (0.014)	0.182 (0.011)	0.024 (0.015)	0.976 (0.015)
20	3.0	20,000 75,000 150,000 300,000	30 15 10 5	1/1000	0.897 (0.014)	0.211 (0.021)	0.789 (0.021)	0.867 (0.009)	0.033 (0.011)	0.967 (0.011)	0.773 (0.008)	0.015 (0.010)	0.985 (0.010)
21	4.0	20,000 75,000 150,000 300,000	30 15 10 5	1/1000	0.923 (0.010)	0.145 (0.018)	0.855 (0.018)	0.915 (0.007)	0.005 (0.012)	0.995 (0.012)	0.840 (0.006)	0.0008 (0.008)	0.999 (0.008)

Scenario*	$RR_{ij}^{s'}$	Exposure rates ¥	No. of drugs (D_j)	Event Incidence (E_i)	Sen. (SD)	FDR (SD)	PPV (SD)	Sen. (SD)	FDR (SD)	PPV (SD)	Sen. (SD)	FDR (SD)	PPV (SD)
22	5.0	20,000 75,000 150,000 300,000	30 15 10 5	1/1000	0.938 (0.008)	0.089 (0.015)	0.911 (0.015)	0.922 (0.005)	0.001 (0.009)	0.999 (0.009)	0.901 (0.004)	0.0002 (0.006)	1.000 (0.006)
23	7.5	20,000 75,000 150,000 300,000	30 15 10 5	1/1000	0.958 (0.005)	0.073 (0.010)	0.927 (0.010)	0.946 (0.004)	0.0002 (0.007)	1.000 (0.007)	0.938 (0.004)	0.0001 (0.006)	1.000 (0.006)
24	10.0	20,000 75,000 150,000 300,000	30 15 10 5	1/1000	0.971 (0.004)	0.057 (0.007)	0.943 (0.007)	0.969 (0.003)	0.00006 (0.008)	1.000 (0.008)	0.952 (0.005)	0.00002 (0.005)	1.000 (0.005)

*Each scenario consists of 1000 simulated datasets.

¥ For each scenario the reporting probabilities were the same.

Table 35: Scenarios for different sized harms databases in clinical trial units.

						$PRR_{O2.5} > 1$			$IC_{O2.5} > 0$			$GPS_{O5} > 2$		
Scenario*	**No. of drugs (D_j)**	**Drugs - Exp**	**Exposure rates (T_j)**	**No. of events (E_i) ¥**	**Event Inc**	**Sen. (SD)**	**FDR (SD)**	**PPV (SD)**	**Sen. (SD)**	**FDR (SD)**	**PPV (SD)**	**Sen. (SD)**	**FDR (SD)**	**PPV (SD)**
I_i = 1/250														
1	60	30 15 10 5	20,000 75,000 150,000 300,000	150	1/250	0.513 (0.014)	0.141 (0.018)	0.859 (0.018)	0.399 (0.012)	0.007 (0.015)	0.952 (0.015)	0.221 (0.011)	0.0003 (0.00008)	1.000 (0.00008)
2	40	20 10 8 2	10,000 50,000 100,000 250,000	120	1/250	0.426 (0.012)	0.248 (0.018)	0.752 (0.018)	0.305 (0.012)	0.037 (0.015)	0.949 (0.015)	0.173 (0.012)	0.006 (0.005)	0.994 (0.005)
3	30	15 9 4 2	5,000 25,000 50,000 150,000	100	1/250	0.335 (0.012)	0.349 (0.015)	0.651 (0.015)	0.213 (0.010)	0.054 (0.014)	0.949 (0.014)	0.105 (0.013)	0.010 (0.009)	0.989 (0.009)
4	20	10 6 3 1	1,000 5,000 25,000 100,000	80	1/250	0.257 (0.010)	0.461 (0.014)	0.539 (0.014)	0.155 (0.008)	0.076 (0.014)	0.948 (0.014)	0.070 (0.010)	0.025 (0.008)	0.975 (0.008)
5	10	4 3 2 1	500 1,000 5,000 20,000	50	1/250	0.168 (0.009)	0.592 (0.015)	0.408 (0.015)	0.098 (0.005)	0.113 (0.010)	0.946 (0.010)	0.005 (0.008)	0.049 (0.005)	0.951 (0.005)
I_i = 1/500														

Scenario*	No. of drugs (D_j)	Drugs - Exp	Exposure rates (T_j)	No. of events (E_j) ¥	Event Inc	Sen. (SD)	FDR (SD)	PPV (SD)	Sen. (SD)	FDR (SD)	PPV (SD)	Sen. (SD)	FDR (SD)	PPV (SD)
6	60	30 15 10 5	20,000 75,000 150,000 300,000	150	1/500	0.469 (0.013)	0.127 (0.016)	0.873 (0.016)	0.350 (0.010)	0.006 (0.014)	0.994 (0.014)	0.213 (0.010)	0.0002 (0.0001)	1.000 (0.0001)
7	40	20 10 8 2	10,000 50,000 100,000 250,000	120	1/500	0.298 (0.011)	0.221 (0.016)	0.779 (0.016)	0.248 (0.008)	0.034 (0.013)	0.966 (0.013)	0.152 (0.010)	0.004 (0.002)	0.996 (0.002)
8	30	15 9 4 2	5,000 25,000 50,000 150,000	100	1/500	0.193 (0.010)	0.318 (0.014)	0.682 (0.014)	0.154 (0.006)	0.050 (0.011)	0.949 (0.011)	0.087 (0.008)	0.008 (0.001)	0.992 (0.001)
9	20	10 6 3 1	1,000 5,000 25,000 100,000	80	1/500	0.114 (0.008)	0.429 (0.010)	0.571 (0.010)	0.097 (0.007)	0.071 (0.009)	0.929 (0.009)	0.048 (0.006)	0.021 (0.002)	0.979 (0.002)
10	10	4 3 2 1	500 1,000 5,000 20,000	50	1/500	0.076 (0.005)	0.521 (0.008)	0.479 (0.008)	0.043 (0.006)	0.101 (0.004)	0.899 (0.004)	0.012 (0.005)	0.042 (0.0009)	0.958 (0.0009)
I_i = 1/1000														

Scenario*	No. of drugs (D_j)	Drugs - Exp	Exposure rates (T_j)	No. of events (E) ¥	Event Inc	Sen. (SD)	FDR (SD)	PPV (SD)	Sen. (SD)	FDR (SD)	PPV (SD)	Sen. (SD)	FDR (SD)	PPV (SD)
11	60	30 15 10 5	20,000 75,000 150,000 300,000	150	1/1000	0.425 (0.012)	0.105 (0.014)	0.895 (0.014)	0.306 (0.010)	0.004 (0.001)	0.996 (0.001)	0.198 (0.009)	0.0002 (0.00007)	1.000 (0.00007)
12	40	20 10 8 2	10,000 50,000 100,000 250,000	120	1/1000	0.254 (0.010)	0.197 (0.012)	0.803 (0.012)	0.204 (0.009)	0.029 (0.011)	0.971 (0.011)	0.137 (0.007)	0.003 (0.001)	0.997 (0.001)
13	30	15 9 4 2	5,000 25,000 50,000 150,000	100	1/1000	0.149 (0.010)	0.283 (0.010)	0.717 (0.010)	0.110 (0.009)	0.045 (0.008)	0.955 (0.008)	0.072 (0.006)	0.007 (0.002)	0.993 (0.002)
14	20	10 6 3 1	1,000 5,000 25,000 100,000	80	1/1000	0.070 (0.008)	0.391 (0.008)	0.609 (0.008)	0.053 (0.005)	0.068 (0.006)	0.932 (0.006)	0.033 (0.003)	0.018 (0.001)	0.982 (0.001)
15	10	4 3 2 1	500 1,000 5,000 20,000	50	1/1000	0.032 (0.002)	0.502 (0.005)	0.498 (0.005)	0.025 (0.001)	0.110 (0.003)	0.890 (0.003)	0.010 (0.001)	0.038 (0.0004)	0.962 (0.0004)

*Each scenario consisted of 1000 simulations.

¥ For each scenario the reporting probabilities were the same.

Appendix E – Publications in this Thesis