An appraisal of health datasets to enhance the surveillance of Lyme disease in the United Kingdom

UNIVERSITY OF LIVERPOOL

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

Doctor in Philosophy

By

John Stamford Parker Tulloch

January 2019

'Infectious disease is one of the few genuine adventures left in the world. The dragons are all dead and the lance grows rusty in the chimney corner... About the only sporting proposition that remains unimpaired by the relentless domestication of a once free-living human species is the war against those ferocious little fellow creatures, which lurk in dark corners and stalk us in the bodies of rats, mice and all kinds of domestic animals; which fly and crawl with the insects, and waylay us in our food and drink and even in our love'

Hans Zinsser – 'Rats, Lice and History' (1935)

Abstract:

An appraisal of health datasets to enhance the surveillance of Lyme disease in the United Kingdom – J.S.P. Tulloch

Lyme disease is a tick-borne disease of increasing global public health interest. Clinical presentation is varied, posing challenges for case definition. Currently national incidence figures for the United Kingdom (UK) are derived from two-tier confirmatory laboratory diagnostic results. These figures have the potential to underestimate incidence as clinical cases managed without diagnostic investigation are unrecorded.

This thesis aimed to identify and evaluate a variety of datasets for their ability to describe the incidence and sociodemographics of Lyme disease cases in the UK, and to assess whether they could be utilised in future national surveillance programmes. The datasets analysed were: Public Health England (PHE)'s Lyme disease diagnostic laboratory, PHE's laboratory surveillance system, hospital episode statistics data for England and Wales, an electronic health records database of primary care in the UK, Twitter, and the Small Animals Veterinary Surveillance Network (SAVSNET).

A generalised Lyme disease population could be described from these data. This population had a bimodal age distribution, was predominately white, was from rural areas, and increasingly from areas with lower societal deprivation. Geographic distribution of cases could be described for England and Wales and showed the highest incidence of disease in southern central to south western England. These data showed an increasing incidence of Lyme disease. The relative incidence of Lyme disease cases varied between datasets, with the primary care data having the largest incidence of 4.42 per 100,000 person-years (95% CI 4.23-4.67). Multiplication factors were described between the three datasets of routinely collected health care data. The most important being a multiplication factor of 2.35 (95% CI 1.81-2.88) between laboratory-confirmed incidence and primary care incidence in England and Wales.

The results from this thesis start to describe the epidemiological picture of Lyme disease in the UK; specifically identified as a research gap by the NICE guidelines. They will provide a platform for the many unanswered questions about the changing landscape of Lyme disease in the UK. It was concluded that a combination of health datasets could be used for future Lyme disease surveillance systems in the UK. Ideally this would include laboratory and primary care data. Until this is in place, the multiplication factor can be used to estimate the national incidence of Lyme disease and the potential burden it places on the National Health Service and the patients it afflicts.

Acknowledgments

This work was funded by the National Institute of Health Research Health Protection Research Unit in Emerging and Zoonotic Infections and was supported by the University of Liverpool and Public Health England (PHE). I want to thank these organisations for giving me the chance to explore such an interesting field of research, and for all the support and opportunities they have provided over the last three years.

I must thank the patients whose health data has been integral to this thesis. Without it this research would not have been possible. I hope it answers some of their many questions about Lyme disease, and helps them understand why access to patient data is so critical to the future of epidemiological research.

The team at the Rare and Imported Pathogens Laboratory have been invaluable to this thesis. Their expertise, the collaborations proffered, and their guidance in navigating what at times is a very emotive and politically charged field has been an incredible help. In particular, Dr Amanda Semper and Dr Tim Brooks, have gone above and beyond with their support.

During the planning and analysis stages of this thesis I worked alongside the Field Epidemiology Service at PHE North West, Liverpool. They have been incredible teachers and their humour and kindness made me feel part of their team. Special thanks must go to Dr Paul Cleary and Dr Valerie Decraene for their epidemiological knowledge and shared train journeys, Richard Dunn and Michelle Man for their SQL tutoring, Tom Inns for many a lunch time walk, and Elaine Mooney for being my desk buddy.

The infectious disease team at the Royal Liverpool University Hospital have provided a huge amount of support relating to all clinical aspects of Lyme disease. Dr Mike Beadsworth and Dr Nick Beeching have been a mine of information, have provided me with some great opportunities, and been involved with many an entertaining conversation. I apologise to the registrars, who were often concerned and confused by my veterinary presence on the ward.

My fellow PhD students based in Leahurst and at the Farr Institute have become close friends. They are great fun and together we've helped each other deal with, what at times have seemed to be, insurmountable problems. The shared PhD experience has been an enlightening one, both professionally and personally, and I'm sure that this special cohort will all be future leaders in their respective fields.

I would like to give special thanks to my supervisors for all the assistance and opportunities that they have given me. Prof Alan Radford, Dr Roberto Vivancos, Prof Rob Christley, and Dr Jenny Warner, have all, in their own manner, shown me the way to becoming a better researcher. They have been an inspiring bunch, and I hope that I will continue to work and socialise with them as colleagues and friends.

Finally, my biggest thanks and love must go to my friends and family. They have offered unquestioned support, despite what appeared to them to be a rather bold/crazy change in career direction. But most of all, I must thank Rachael, who has been receiving the brunt of my excitableness on the good days, and inevitable moaning on the bad days. She has given me such incredible support and love and has helped me maintain my drive and focus. I will remain eternally grateful for this.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Accident and emergency – A&E

Acrodermatitis chronica atrophicans – ACA

Augmented Dickey-Fuller test – ADF test

Admitted patient care – APC

Application Programming Interface - API

Area of outstanding natural beauty – AONB

British Infection Association - BIA

Central Statistics Office – CSO

Clinical commissioning group - CCG

Clinical Practice Research Datalink – CPRD

Clinical Research Network – CRN

Continuing professional development – CPD

Department of Health and Social Care - DHSC

Dynamic time-warping - DTW

Electronic health record – EHR

English Indices of Deprivation - EID

Enzyme-linked immunosorbent assay - ELISA

Erythema migrans – EM

European Centre for Disease Prevention and Control - ECDC

Evidence for Policy and Practice Information – EPPI

Exploratory spatial data analysis - ESDA

General practitioners – GP

Great Britain – GB

Health Protection Surveillance Center – HPSC

Health Research Authority - HRA

Hospital Episode Statistics – HES

Index of Multiple Deprivation – IMD

Infectious intestinal disease - IID

Inflammatory bowel disease – IBS

Integrated Research Application System - IRAS

International Classification of Diseases, 10th Revision - ICD-10

Laboratory information management system - LIMS

Liverpool School of Tropical Medicine – LSTM

Local indicators of spatial association - LISA

Lower super output area - LSOA

National Health Service – NHS

NHS Research Ethics Committee - REC

National Institute for Health and Care Excellence – NICE

National Institute for Public Health and the Environment - RIVM

Office for National Statistics - ONS

Outpatient attendance - OA

Patient Episode Database for Wales – PEDW

Positive predictive values - PPV

Post-treatment Lyme disease syndrome - PTLDS

Practice management systems - PMS

Public Health England - PHE

Quality and Outcomes Framework - QOF

Rare and Imported Pathogens Laboratory - RIPL

Real-time syndromic surveillance team - ReSST

Royal College of General Practitioners - RCGP

Royal Liverpool and Boradgreen University Hospitals NHS Trust - RLBUHT

The Second Generation Surveillance System - SGSS

The Small Animal Veterinary Surveillance Network - SAVSNET

The Health Improvement Network – THIN

Tick-borne diseases – TBD

Tick Surveillance Scheme - TSS

United Kingdom - UK

United States of America – USA

Welsh Index of Multiple Deprivation - WIMD

World Health Organisation – WHO

Chapter 1 An Introduction

> ***'it is believed that Lyme arthritis extends beyond the three communities studied here; but how far beyond is not known'*** **Allen Steere** ***[1]***

In 1977 a group of American clinicians described 'a previously unrecognized clinical entity' and named it "Lyme arthritis" [1]. Lyme disease had entered the medical lexicon and the general public's consciousness. Soon after this case description, Willy Burgdorfer isolated a spirochaete as the causal pathogen [2]. In his honour, this spirochaetal genospecies complex was named *Borrelia burgdorferi* sensu lato. Since Lyme disease's 'discovery', interest grew in this 'new' disease and soon cases were being identified in the United Kingdom, the first cases being reported in the early 1980's, with concerns being raised that we were on the brink of a new epidemic [3–8].

After the emergence of this pathogen onto the world stage, cases have steadily risen in number across the Northern Hemisphere [9,10]. The dawn of the social media age has led to a jump in the public's awareness, with media stories, health scares, and subsequent cases spreading in a meme-like manner around the globe [11–15]. Despite the relatively recent recognition of this disease, cases had been seen in Europe prior to it being identified in Lyme, Connecticut. In fact, the first clinical description of the pathognomonic erythema migrans (EM) rash, was by a Swedish dermatologist, Arvid Afzelius, at a Swedish Dermatology conference in 1909 [16]. In the 1950's, Hollström described a spirochaete being isolated from the EM lesion, and patients being successfully treated with penicillin [17], something which some clinicians still recommend for specific presentations of the disease today [18].

Recent research indicates that *B. burgdorferi* has infected humans since at least the Copper age (between the 5th and 3rd millennia BCE) [19], and has existed in the global environment for around 60,000 years [20]. The authors of the latter paper conclude that;

> '*the recent emergence of human Lyme disease probably reflects ecological change – climate change and land use changes over the past century – rather than evolutionary change of the bacterium*.'

Despite the history and high profile nature of this disease within the general public and media's minds, little epidemiological knowledge exists about Lyme disease within the United Kingdom (UK). The aim of this thesis is to utilise health data to fill in some of the gaps about Lyme disease in the UK, specifically in terms of incidence and patient demographics.

1.1 What is Lyme disease?

1.1.1 Pathogen, vectors and hosts

The causal pathogen of Lyme disease is *B.burgdorferi* sensu lato, which includes at least twenty different genospecies [21–23]. The three main genospecies responsible for disease in humans are *B.burgdorferi* sensu stricto*, Borrelia afzelii* and *Borrelia garinnii* [22]. They rely on a variety of tick species to act as vectors to transmit the pathogen to humans and other species. These ticks all belong to the *Ixodes* genus*, Ixodes ricinus* is the predominant tick species in the UK (Fig. 1.1) [22,24].

Figure 1.1 A female I.ricinus questing. ©ECDC/Guy Hendrickx.

The enzootic cycle of *B.burgdorferi* is complex and varies depending upon geography, climate and local ecology [25,26]. *Ixodes ricinus* has been found on a wide range of host species, from deer, to small mammals and birds [24,27,28]. Within the tick life cycle, there is no vertical transmission of the spirochaetes, and the perpetuation of Lyme disease is reliant on horizontal transmission [21]. Therefore, at a local level, infected nymphs must feed and infect a host animal in spring, with subsequent uninfected larvae becoming infected by feeding on these reservoir hosts, which then moult to become infected nymphs the subsequent year. Some animals, such as deer and humans, will only act as dead end hosts in this cycle (Fig. 1.2).

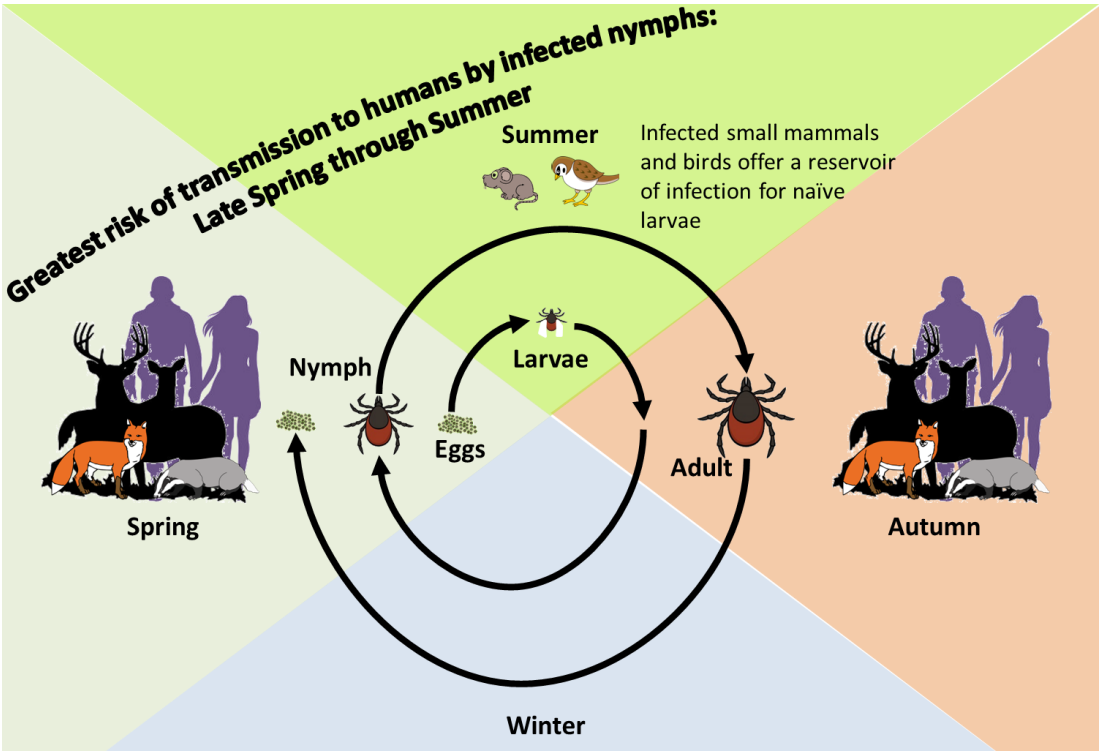


Figure 1.2 The life cycle of Ixodes ticks and their relationship with reservoir and dead-end hosts.

This life cycle can take up to three years to complete [21,29]. Only the nymph and adult life cycle stages can feed on humans [21,29]. They identify potential blood meals by detecting hosts through a process called questing, during which they wait on the edge of leaves or grass with their third and fourth pairs of legs waiting to grasp on to a host, whom they detect through odours, heat, moisture and vibrations [30]. Lyme borrelia reside within a tick's midgut and undergo phenotypic change on exposure to a blood meal. This enables the spirochaetes to increase in number and populate the tick's salivary gland [21]. To transmit *Borrelia spp* they must feed on the host's blood. They appear to have to feed for at least 24 hours to transmit the pathogen, with 10% transmission at 48 hours and 70% by 72 hours of attachment [31]. In the Netherlands it has been estimated that 2.6% of tick bites (with unknown Lyme status) result in the development of Lyme borreliosis [32]. This risk increased if the tick was engorged at time of removal and with the length of time attached increasing [32]. Once *B.burgdorferi* has been transmitted into the host's dermis, it will multiply locally and be disseminated via the blood and lymphatic vessels [21].

1.1.2 Pathogenesis, clinical presentations and epidemiology

Upon the inoculation of *B.burgdorferi* into a human host an innate and adaptive immune response is elicited. The majority of the resultant tissue damage to the host is due to this inflammatory process. The intensity of this varies depending upon the mix of genospecies present in the tick's saliva.

Early dissemination is typified by erythema migrans (EM), Lyme neuroborreliosis or Lyme carditis. The type of erythema migrans lesion varies depending upon the genospecies involved. *B.burgdorferi* sensu stricto (predominantly found in North America) usually presents with multiple rapidly expanding rashes and is accompanied by flu-like symptoms. *B.afzelii* and *B.garinii* (most prevalent in Europe) tend to expand slower. *B.afzelii* is also associated with the rare manifestation of Borrelial lymphocytoma. *B.garinii* can present with a solitary lesion and can be associated with a burning sensation [18,21,23].

Between days to weeks *B.burgdorferi* sensu stricto disseminates to multiple skin sites, causing multiple EM lesions, or to other organs. It appears to have a predilection to the peripheral and central nervous system (PNS and CNS), the heart and joints. *B.afzelii* tends not to disseminate and can persist at the same skin site for a long time. *B.garinii* is very neurotropic and will cause PNS and CNS abnormalities [18,21,23].

The type of Lyme neuroborreliosis presentation appears to vary by genospecies. *B.burgdorferi* sensu stricto typically leads to lymphocytic meningitis with headaches and neck stiffness, neuropathies of the cranial nerves, and sensory or motor radiculopathies. *B.garinii* presents as Bannwarth syndrome, a lymphocytic meningoradiculitis. *B.afzelii* causes less specific neurological symptoms such as headaches, dizziness, and memory or concentrations disturbances [18,21,23].

Late dissemination of Lyme borrelia infection is characterised by arthritis, acrodermatitis chronica atrophicans (ACA), and rarer neurological conditions. The main genospecies associated with arthritis is *B.burgdorferi* sensu stricto. Symptoms appear on average 6 months post infection and are typified by swelling of the large joints and a persistent synovitis that can persist up to 4-5 years. ACA is associated predominately with *B.afzelii* and is seen more frequently in older women, though the reasons for this are unknown [18,21,23].

In the United States between 2001 and 2016, 70% of 253,690 cases showed erythema migrans, 30% showed arthritis, 9% Bell's Palsy, 4% radiculoneuropathy, 4% meningitis or encephalitis, and 1% showed cardiac symptoms [33]. There are no nationwide prevalence studies of clinical manifestations of Lyme disease in the UK. A few UK studies have identified the prevalence of *B.burgdorferi* genospecies found in *I.ricinus* ticks, however most were limited to localised tick collections (from one habitat or one region) or from ticks found on a specific city. The prevalence of *B.burgdorferi* s.l. carrying ticks ranged from 1.8-18.1%. Of the borrelia positive ticks between 10.5% and 60% were carrying *B.garinii,* 1.6-48% were carrying

B.afzelii, and 0-25.5% were carrying *B.burgdorferi* sensu stricto [34–37]. The largest study, based in Scotland, surveyed over 2200 ticks and found 5.6% positive for *B.burgdorferi* s.l.. Of these 48% were *B.afzelii*, 36% were *B.garinii,* 7% were *B.burgdorferi* sensu stricto, and 14% had a mix of genospecies [38]. One could hypothesis that, due to the genospecies mix, there would be less presentations of arthritis and more neurological presentations in the UK compared to the USA. Research is needed to establish the prevalence of different clinical presentations of Lyme disease in the UK.

The very varied and subtle clinical presentation of Lyme disease presents a challenge for clinicians and researchers when creating a case definition. In this thesis, the case definition will be based on the recently published National Institute for Health and Care Excellence (NICE) guidelines [23]. This was chosen as the research committee who wrote the guidelines thoroughly reviewed all the scientific and medical literature, have many years of clinical and personal experience, and have applied this knowledge to make UK specific conclusions. Outlined below is a brief summary of the clinical presentations of Lyme disease outlined by NICE [23].

- **Erythema migrans** - A pathognomonic rash that usually arises a few days to four weeks post-exposure and lasts for several weeks. It is a red rash, which increases in size and can present with a central clearing, thus presenting with a textbook 'bull's eye' rash (Fig. 1.3). Although it usually appears near the tick bite site and is not hot, itchy or painful, there can be multiple expanding rashes, away from the tick bite site. This latter feature is unlike a tick bite reaction, which it could be easily confused with.

Figure 1.3 An erythema migrans rash. Image provided by the CDC.

- **Non-focal symptoms –** These include: flu-like symptoms (pyrexia, sweats, lymphadenopathy, fatigue), neck pain and stiffness, joint or muscle pain, cognitive impairment, headache, and paraesthesia. These symptoms matched with a history of possible tick exposure, increases the likelihood that a patient is infected with Lyme disease.
- **Cranial nerves and peripheral nervous system symptoms** – These include: facial and cranial nerve palsies (including Bell's palsy), mononeuritis multiplex, and unexplained radiculopathies.
- **Central nervous system symptoms –** These include: meningitis, encephalitis, neuropsychiatric presentations, and unexplained white matter changes on brain imaging.
- **Other symptoms –** These include: inflammatory arthritis often affecting multiple joints, cardiac problems (heart block and pericarditis), ocular symptoms (uveitis, keratitis), and other dermatological presentations (acrodermatitis chronica atrophicans (ACA), lymphocytoma).

The age distribution of cases is bimodal, with peaks in pre-adolescence and late middle age [21,23,29,39]. The sex ratio is varied between nations and datasets analysed.

In areas known to have *Ixodes* ticks, the risk of human infection is associated with: the abundance of ticks, the prevalence of *B.burgdorferi* infected ticks, the presence of a human in a tick habitat, and human behaviours that could increase the chance of being bitten [22,23]. This would include occupational and recreational habits.

1.1.3 Diagnosis

The NICE guidelines and other international guidelines all commonly state that clinicians should 'Diagnose and treat Lyme disease without laboratory testing in people with erythema migrans' [23,39–44]. This is due to the high specificity (88-99%) of this presentation [23]. If any of the other clinical presentations arise with a history suggestive of tick exposure, but in the absence of an EM, then a two-tier testing system is recommended by NICE [23]. The initial screening test is an enzyme-linked immunosorbent assay (ELISA) for IgM and IgG antibodies, based on C6 peptide or an equivalent VlsE antigen. If this test is positive or equivocal a confirmatory immunoblot should be performed. Timing of these tests is critical for interpretation, as a humoral immune response may not be strong enough to be diagnostic if the test is performed at symptom onset. It is recommended that diagnostic tests are performed at least 4 weeks after symptom onset and then potentially repeated 4 to 6 weeks

later. If signs persist for 12 weeks after a negative ELISA, an immunoblot should be performed. However, there is a consensus that patients with non-specific subjective symptoms are not recommended to undergo diagnostic testing because of a low positive predictive value [44].

Despite what appears to be a straight-forward diagnostic pathway there are some limitations. Firstly, the decision that an EM rash is pathognomonic only remains valid if the clinician recognises that the rash is an EM rash. An American study showed that 72% of clinicians are unable to correctly identify an EM over other common rash lesions in an ambulatory population [45,46]. Similar work has not been conducted in the UK. These results suggest that there is the potential for a high number of false negatives and a lot of missed early cases. The degree of missed diagnoses is likely to be highly variable due to the knowledge and experience of the clinician involved and the frequency and volume of Lyme disease cases seen in their clinic.

The diagnostic recommendations currently all revolve around identifying a humoural response in a patient rather than identifying the presence of a pathogen. Methodologies proving the presence of the spirochaete involve culture and PCR. Due to the low bacterial load in blood it is very challenging to find the spirochaete in high enough numbers for diagnosis. They can be found in higher concentrations at the skin lesion, in synovial fluid and in cerebrospinal fluid. However, borrelia DNA persist after spirochaetal killing and so identification of the pathogen may not demonstrate an active infection. Results from these tests are highly variable with suboptimal sensitivities and are only utilised in research settings [18,21,23,44].

Basing diagnosis on an immune response has its own set of limitations. It assumes that a patient produces an immune response and that this response is related to a current, rather than a past, infection. It also assumes that a diagnostic test is performed at the correct time to identify the immune response. Fifty percent of patients presenting with an EM rash are initially seronegative, but by 6-8 weeks after the presentation of clinical signs 99% will become seropositive [44]. Therefore, cases could be missed if the diagnostic tests are performed too early.

When the tests, outlined above, are utilised at the correct time the average sensitivity across Europe has been reported at 80%, with an average specificity of around 95% [41,44]. To interpret the results of a test appropriately the background seroprevalence is needed. In the

UK this is unknown, but the Scottish population has an estimated seroprevalence of 4.2%, with regional seroprevalences ranging from 0% to 8.6% [47]. It is known that higher risk populations, such as hunters and forest workers, can have a seroprevalence of greater than 20% [41,44]. When the background seroprevalence is low it makes interpreting a positive diagnostic test result harder as the positive predictive value will also be low. Using the above figures (prevalence = 4.2%, sensitivity = 80%, specificity = 95%) a positive predictive value of 41.2% and a negative predictive value of 99.1% can be calculated. This means that, regardless of clinical presentation, only 41.2% of positive results are likely to represent infected individuals, where as 99.1% of test negative results correctly represent uninfected individuals. This example is only hypothetical and should be treated with a certain degree of caution, but it highlights why current recommendations suggest that serological testing is only requested when the clinical symptoms and case history are supportive of Lyme disease [23,41,44]. By adopting these guidelines, the likelihood of false positive results is reduced in patients with no Lyme specific symptoms and a low pre-test probability of Lyme disease.

The diagnostic pathway for Lyme disease has some challenges. It is vital that clinicians at all stages of the diagnostic process, from primary care clinics to diagnostic laboratories, have as thorough clinical history as possible. This enables contextualisation of the diagnostic results and would lead to an appropriate interpretation and subsequent management of the patient.

1.1.4 Management and Treatment

For the majority of cases, which present with erythema migrans as the primary clinical complaint, patients can be managed in primary care. Those presenting with symptoms suggestive of a central nervous symptom infection or complete heart block are recommended to be referred under standard clinical emergency referral practices. For children, all presentations barring erythema migrans should be discussed with a specialist [23].

Antibiotic therapy is the recommended course of treatment for all cases. Depending on the clinical presentation and age of the patient, the active ingredients used are: doxycycline, amoxicillin, azithromycin, or intravenous ceftriaxone. Use of non-antibiotic therapy is encouraged for certain symptoms, such as depression, chronic pain and fatigue [23].

1.1.5 Prognosis and controversies

Persistent symptoms can occur after a course of antibiotics either due to re-infection, an interruption of the antibiotic course, or organ damage caused by Lyme disease. Some

patients can have prolonged arthritis as the inflammatory process during infection which can lead to the development of degenerative arthritis [22,23]. Patients with neuroborreliosis can have residual motor or sensory deficits. This is due to permanent nerve damage or the lengthened healing process of nervous tissue. ACA can also have prolonged symptoms after treatment, similar to other sequelae of long term skin infections. These include, skin atrophy and subluxation of small joints. However, the classical presentations of disease normally resolve after the completion of an antibiotic course [18]. Any non-specific subjective symptoms such as fatigue, cognitive impairment and musculoskeletal pain should also resolve.

Some patients report symptoms for more than six months after treatment completion [48–57]. These can last in excess of ten years and have been described as either 'post-treatment Lyme disease syndrome' (PTLDS) or 'chronic Lyme disease'. There is much debate about whether these symptoms are more common with Lyme disease compared to healthy controls or other infections and how they should be managed [48–57]. This cohort of patients often present with similar symptoms to other infectious diseases with post-infectious/latent sequelae. Notable examples include chronic Q fever [58–61] and infectious mononucleosis [62].

The pathogenesis and management of this presentation of disease for this subset of patients remains unclear, which can lead to great anger, frustration and tension between patients and clinicians. It is outside the scope of this thesis to discuss such a large and controversially debated topic. Nevertheless, regardless of disease pathogenesis this cohort of patients are ill and should not be dismissed. They deserve the most appropriate management and treatment based on the current scientific evidence base.

1.2 What is the situation globally?

As part of the Department of Health's commissioned review into Lyme disease (section 1.3), a systematic review was performed into Lyme disease surveillance programmes that occur globally [9]. Before considering their findings, it must be noted that current evidence suggests that endemic Lyme disease cases can only occur where the *Ixodes* tick vectors live. This is restricted to the Northern hemisphere. Countries where *Ixodes spp* have been found in wild habitats are displayed in Fig. 1.4 [18,63].

Figure 1.4 The known distribution of the tick vectors of Lyme disease, adapted from figures in [18,55] (The southern hemisphere hasn't been included, as no competent tick vectors have been found there).

Note that only a handful of countries in the Northern hemisphere do not have *Ixodes* ticks. This is may be a result of inhospitable habitats for the ticks (i.e. the ticks are truly absent), or a lack of tick surveillance in these countries (i.e. the ticks have not been reported). A Department of Health review of surveillance systems in Europe and North America [9] reported that, of the 34 countries assessed, 28 had national Lyme disease surveillance systems, 25 of which publicly reported surveillance figures (Fig. 1.5 and 1.6).

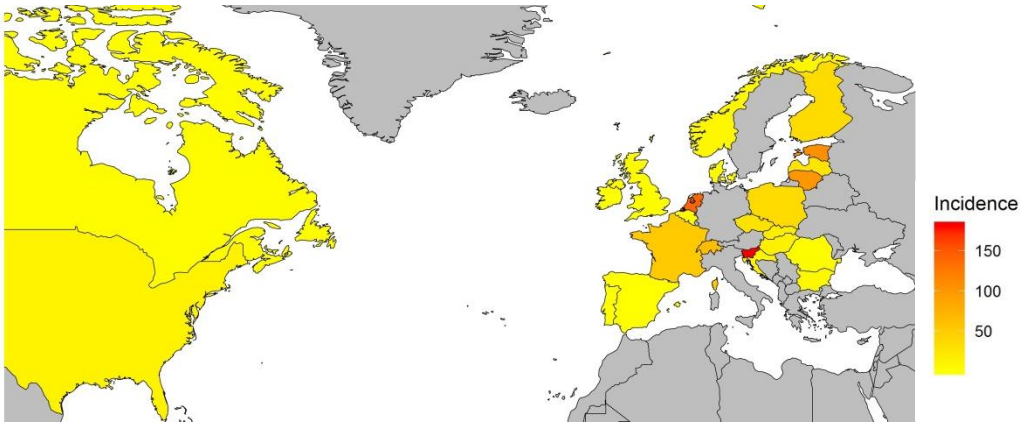


Figure 1.5 The 2015 annual incidence (cases per 100, 000 population) of countries with national surveillance systems for Lyme disease in place (countries in grey have no national surveillance systems). Figures adapted from tabular results in [9].

As one can see, there is a large range in incidence figures, with Slovenia having the highest reported national incidence (181 cases per 100,000 per year) and many countries having an incidence of less than one case per 100,000 per year [9]. Not all countries have national surveillance systems in place (for example, Germany and Austria), but have regional surveillance systems. These countries were therefore not included in this systematic review.

Surveillance data from the United States of America (USA) describes Lyme disease incidence rising from 6.7 cases per 100,000 population in 2004 to 8.1 in 2016 [29,64,65]. Surveillance methods in the USA vary between states and so it is likely that there is a large underestimate of cases [64,65]. There is also a huge range between states, with Louisiana having an incidence of zero and Maine having the highest at 86.4 [65]. This variability is likely due to

the differences in surveillance methods, the differing tick species and their relative abundance in each state, and the prevalence of *Borrelia*-carrying ticks. Nelson et al attempted to get a more representative incidence figure by looking at medical insurance claims; they calculated an incidence of 106.6 cases per 100,000 per year [64]. In 1990, less than 5,000 cases were reported from Europe to the World Health Organisation (WHO) Centralized Information System for Infectious Diseases; this has risen to over 35,000 cases in 2010 [10]. A population-weighted incidence rate in Western Europe has been estimated at 22.05 cases per 100,000 person-years [66]. Like the USA, there is a huge range in annual incidence captured by national surveillance systems in Europe (Fig. 1.5 and 1.6). This is likely due to the same reasons mentioned above. Overall there appears to be an increase in incidence both in Europe and North America over the last twenty years. Reasons for this are likely to include a mixture of: the different national surveillance systems, the changing case definitions between countries and globally, the increased awareness of the disease by the general public and clinicians, and a real increase in disease.

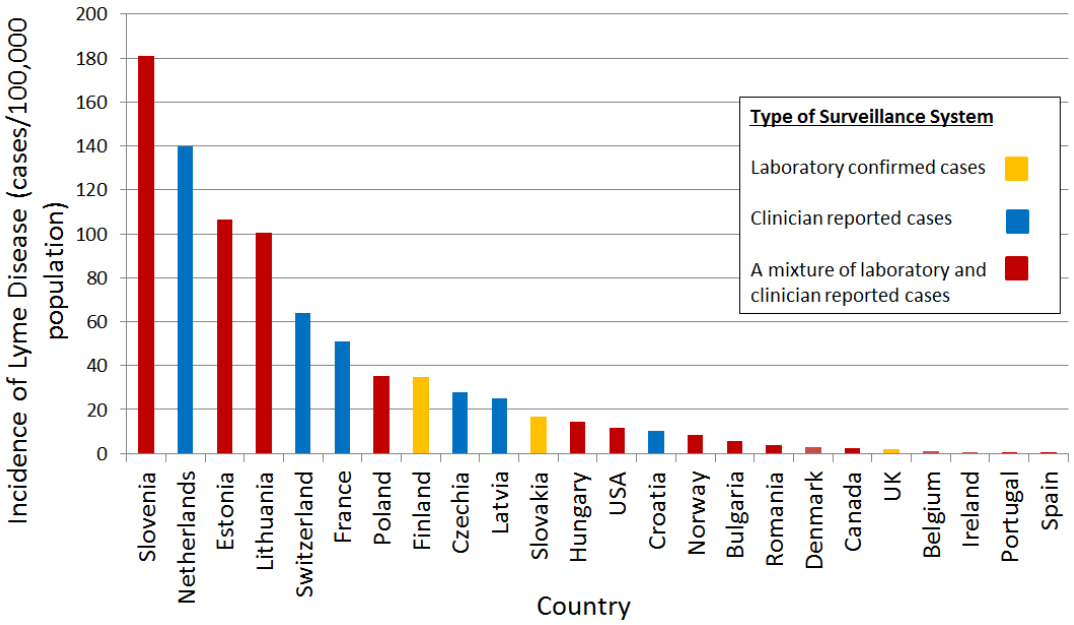


Figure 1.6 The 2015 annual incidence (cases per 100,000 population) of national surveillance systems for Lyme disease, stratified by surveillance system type. Figures adapted from tabular results in [9].

Throughout North America and Europe, there are a variety of surveillance systems in place (Fig. 1.6). Only three surveillance systems, including the UK, rely solely on laboratory-confirmed cases. The remainder either base their figures solely on clinician reported cases or a combination of confirmed cases and clinician reported cases. Seventy-five percent (n=21) of these systems require mandatory notifications; however the case definitions vary widely between the systems [9]. It was noted that the level of clinical detail needed to define

a case varied considerably, which was reflected in the precision of the incidence estimate. As a result, direct comparison of incidence between countries remains challenging. Some countries have no national surveillance (Sweden, Austria, Germany), which is surprising as they have reported high levels of Lyme disease. For example, some German states have an incidence as high as 74.8 cases per 100,000 [67], but surveillance is led by each state rather than at a national level.

Lorenc et al's systematic review found sixteen papers that attempted to compare different surveillance data to determine the completeness of surveillance systems [9]. Comparisons were only included in the analysis if they used consistent case definitions. A median estimate of the percentage of unreported cases was calculated at 30%, with a range of 10-120%. The studies were all performed in countries with mandatory clinician reporting; only five studies were based in Europe (Belgium, Denmark, Germany, Norway, and Slovakia). Data comparing laboratory data and clinician data were very limited. With a huge range of unreported cases, one could consider that the surveillance systems are performing badly. However, patient demographics, geographic location, and incidence trends remained constant over time in each of the studies. This suggests that each system is collecting a representative sample of the overall Lyme disease affected population, but not in the same consistent proportions. This led the authors to conclude that the introduction of mandatory clinician notification would increase the number of cases captured by surveillance, but it would not necessarily be a more reliable methodology than laboratory-confirmed cases alone. The systematic review succinctly concluded that,

> '*a combination of methods gives more complete coverage in terms of the identification of cases than any single method alone, but also that no combination can guarantee full coverage of all cases. Whether such a combination of systems gives a better representation of overall trends cannot be determined from the available data '* [9]

1.3 What is the situation in the United Kingdom?

In November 2015, National Health Service (NHS) England commissioned NICE to develop clinical guidelines for Lyme disease. In May 2016, the Department of Health commissioned three reviews into the diagnosis, treatment and management of Lyme disease in the UK [68]. These were commissioned due to the growing call for UK-specific guidance and the increase in public and parliamentary lobbying of the UK government, as evident by recent parliamentary records. Between May 2014 and the end of 2017 there were 58 written

parliamentary questions to the Department of Health (Fig. 1.7), these peaked in 2016 when both the NICE guidelines and systematic reviews had been commissioned [68]. During this period there were two Commons Chambers debates, a House of Lords Grand Committee and four minor Commons debates, with at least four general public petitions to the government relating to Lyme disease with over 12,000 signatures [69].

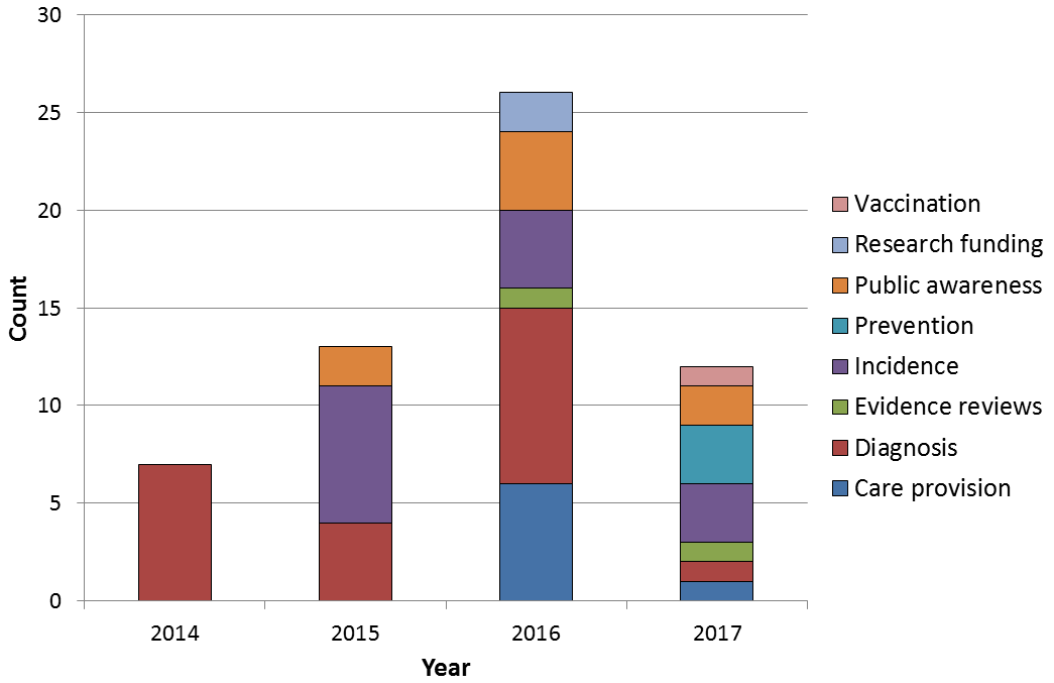


Figure 1.7 The number of written parliamentary questions by theme between May 2014 and December 2017. Based upon parliamentary records [60].

The main themes identified in these questions were around diagnosis (36%, n=21), incidence (24%, n=14), general public and medical practitioner awareness (14%, n=8), and NHS care provision (12%, n=7).

To place this level of concern in to context, we can compare Lyme disease to *Campylobacter*. *Campylobacter* is a zoonotic infection, which has a much higher reported incidence (90.8 cases per 100,000 in 2016), and has the potential for great political concern because of its impact on food safety. However, during the same time period there were only 15 written questions, and 15 verbal mentions in parliament, there were no parliamentary debates or petitions by the general public [68–70].

Further evidence for the general public's interest in Lyme disease can be observed in the increase in relative internet searches using the Google search engine in the UK (Fig. 1.8) [71]. These show a gradual increase until the summer and autumn of 2015, when there is a large

spike in searches. This is just before the announcement of the commissioning of NICE guidelines.

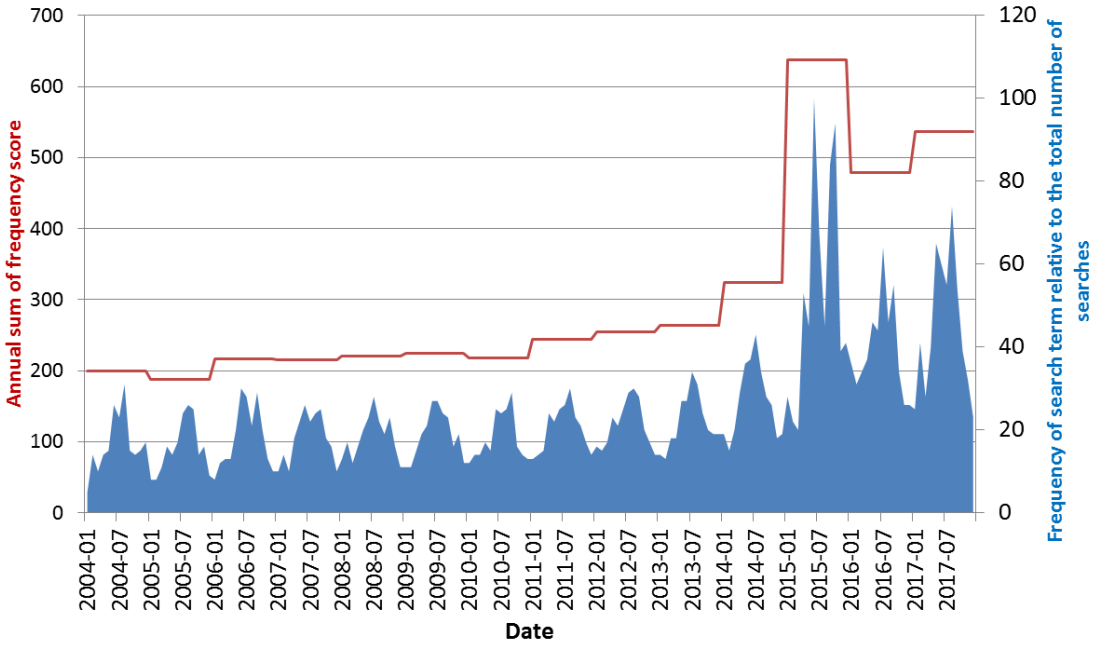


Figure 1.8 The frequency score and annual sum of Google searches in the United Kingdom for 'Lyme disease'. Based upon data provided by Google Trends [63].

This overwhelming increase in awareness of the general public and frequent debates and questions in parliament understandably led to the NICE guidelines and the Evidence for Policy and Practice Information (EPPI) systematic reviews being commissioned by the Department of Health. EPPI published their findings in November 2017 through five projects with the following scopes:

1) The nature and extent of research evidence on Lyme disease in humans [22]
2) The incidence of Lyme disease in the UK and the type of surveillance systems in different countries [9].
3) Patient, clinician and researcher experiences of Lyme disease diagnosis [72]
4) Patient, clinician and researcher experiences of Lyme disease treatment [73]
5) The effectiveness of different approaches for preventing Lyme disease [74]

The second of these projects detailed the current epidemiological situation in the UK, and is most pertinent to the current thesis [9], finding eleven studies that detailed the incidence of Lyme disease in the UK. They were all based on official surveillance reports from the devolved nations of the UK, which are based on laboratory-confirmed cases only. They described a peak incidence in later middle age (40-64 years), with similar rates in men and women, and a large variation in incidence between nations (Fig. 1.9).

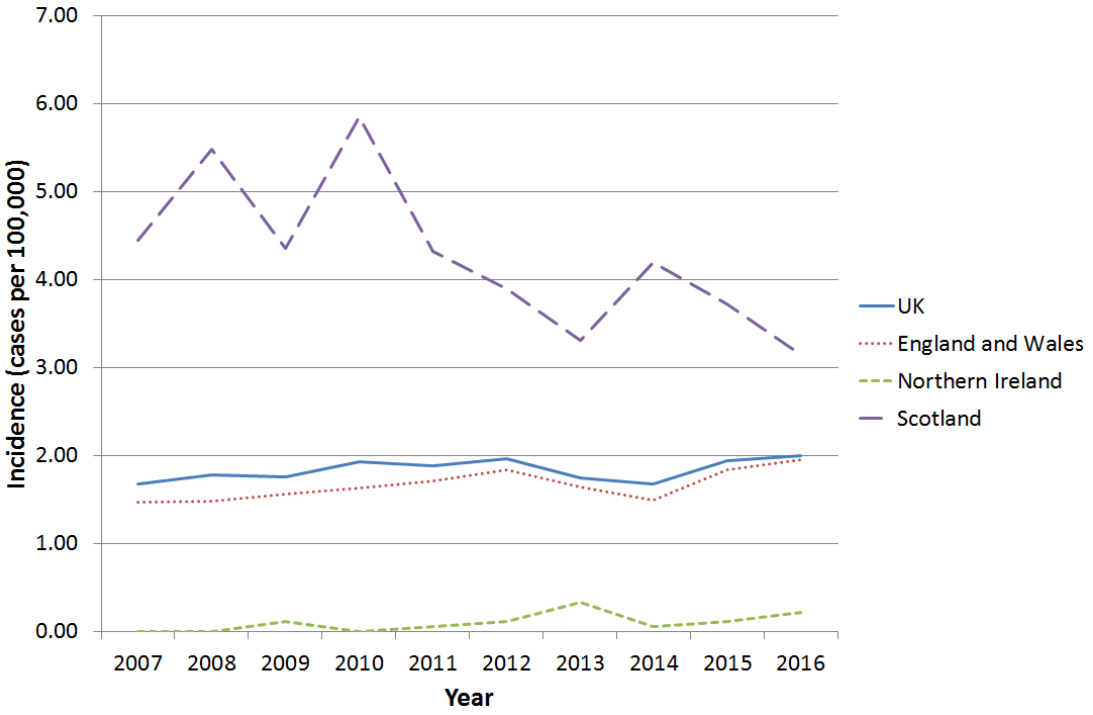


Figure 1.9 Government published Lyme disease incidence figures based on laboratory-confirmed cases (2007-2016). NB. Reporting standards in Scotland changed in 2012. Figure derived from Public Health England's Zoonosis Report [62].

UK incidence figures have risen from 1.67 cases per 100,000 in 2007 to 2.00 in 2016. This rise has been reflected in England and Wales (1.47 to 1.95), and Northern Ireland (0 to 0.21) [70]. In Scotland cases have remained relatively stable between 3 and 4 cases per 100,000, or the incidence may in fact be decreasing; in 2016 the incidence was 3.15 per 100,000. However, there is evidence that disease hot spots exist and national figures mask potentially significant geographical variation. In Scotland hot spots have been explicitly defined in the Highlands [75,76]. In England, hotspots are predominantly anecdotal, and no geographical surveillance data is published. With this lack of evidence the authors of the review conclude that,

> '*Given the absence of geographically comparable routine data, we cannot determine how much of the discrepancy is due to geographical variation and how much to under-reporting or under-diagnosis of Lyme disease. It seems likely that there is some combination of both these factors*.'[9]

Since these surveillance reports only use laboratory-confirmed cases; cases which do not require laboratory confirmation (predominantly erythema migrans) will not be captured and these surveillance figures will represent an underestimate. The report stated,

'We do not know how many cases are diagnosed and treated in the UK on the basis of clinical symptoms alone, without a diagnostic test being ordered. Hence, it is challenging to establish the true extent of under-reporting.' [9]

The report's authors also discussed how little is known about the reasons behind the rise in cases and whether it is reflective of the true incidence or a result of changes in reporting practice, awareness of clinicians and the general public, clinical practice, and healthcare-seeking behaviours in patients.

The following conclusions were made:

- The incidence figures in the UK are likely an underestimate.
- Current UK data is currently unable to identify disease hotspots, especially important as publications suggest there is a large variation in localised incidence.
- No contextual information is provided on location, demographics and clinical presentation of cases.

The authors of the report suggested several surveillance policy options from their findings:

a) Maintain the existing system with no change
b) Introduce mandatory clinician reporting for all Lyme disease cases
c) Introduce mandatory clinician reporting for late or disseminated Lyme disease cases
d) Include Lyme disease in clinician sentinel networks
e) Introduce enhanced surveillance using clinician questionnaires.

NICE is an executive non-departmental public body of the UK's Department of Health. It publishes guidelines on clinical practice, specifically guidance on the appropriate treatment and care of persons with a specific disease or condition. The NICE guidelines for Lyme disease were published in April 2018 [23]. Alongside recommendations of awareness, diagnosis, management and information to be provided to people affected by Lyme disease, they also published recommendations for research, including research into the clinical epidemiology of Lyme disease in the UK. They ask and state;

'What are the incidence, presenting features, management and outcome of Lyme disease…in the UK?'

'There is a lack of robust epidemiological data on Lyme disease in the UK, particularly in people who are immunocompromised or pregnant. A large clinico-epidemiological study to collect data on incidence, presenting clinical features,

management and outcome of Lyme disease in community and hospital settings in the UK would generate population-based statistics. These statistics would enable interventions such as antibiotic treatment and service improvements to be assessed properly, and for services to be tailored so they best serve people with Lyme disease; this was felt to be of high priority. There is no current requirement to notify cases of Lyme disease, therefore, current data are likely to under-estimate the number of people who are seen and treated in the community without serological testing. The morbidity of those who are not rapidly diagnosed and those who seek and receive non-standardised care outside the NHS would justify the costs of this large study.'

1.4 Summary and aims

Little is known about the epidemiological situation of Lyme disease in the United Kingdom, specifically around the incidence, sociodemographics, clinical presentation, and management of Lyme disease cases. The government, non-departmental public bodies, and patient groups have all called for this situation to be redressed. If this enhanced data were to be analysed it would potentially impact on surveillance policy and enable the assessment of any future health interventions. The aims of this thesis, conceived before the NICE guidelines were published, were to appraise health datasets for their ability to describe the epidemiology of Lyme disease in the United Kingdom, and to evaluate their potential for Lyme disease surveillance.

This will be achieved through the following research questions:

- For each health dataset appraised (Chapters 3-8);
    - What is the incidence of Lyme disease?
    - What are the sociodemographics of the patient population?
    - Can any geographical hotspots be identified?
    - Can any data about patient presentation and management be extracted?
    - Is there any additional information about the epidemiology and management of cases which is unique to this dataset?
- Using the current laboratory-confirmed based surveillance system (RIPL) as a reference point, how does each of the other dataset compare? How complete is each dataset, and is there a stable multiplication factor that can be applied to RIPL, that can provide an improved overall annual incidence estimate? (Chapter 9)
- Based on the datasets analysed, what policy described by EPPI should the public health authorities of the United Kingdom adopt? (Chapter 10)

1.5 Thesis Outline

This thesis is a research paper style thesis, that has two papers published (Chapter3 and 8), three in peer-review (Chapters 4, 5, 7), two additional non-published results chapters (Chapters 6, and 9), and a dataset selection chapter (Chapter 2). To maintain consistency in style, chapters 6 and 9 have been written in the style of research papers. In accordance with University of Liverpool guidelines, each chapter containing a published/in peer-review paper will have an introductory and discussion section explaining how it links to preceding and following chapters.

The structure of this thesis follows the design of the hypothesised surveillance pyramid, displayed in Fig. 2.1, by initially exploring datasets at the top of the surveillance pyramid, and with each chapter descending to a lower level in the pyramid.

Chapter 2 explains the decisions behind selecting the datasets that were chosen for analysis.

Chapter 3 describes the incidence and sociodemographics of laboratory-confirmed Lyme disease patients in England and Wales' national Lyme reference laboratory (RIPL), and PHE's laboratory surveillance system (SGSS).

Chapter 4 describes the incidence and sociodemographics of Lyme disease patients identified through ICD-10 codes in English (HES) and Welsh (PEDW) hospitals.

Chapter 5 describes the incidence and sociodemographics of Lyme disease patients identified through Read codes in a primary care electronic health record dataset (THIN).

Chapter 6 describes a novel blinded questionnaire study of primary care clinicians to identify coding behaving in relation to Lyme disease, as a methodology to validate the Read codes utilised to produce the results in Chapter 5.

Chapter 7 describes the spatial and temporal incidence of Twitter users who tweet about Lyme disease in the UK and the Republic of Ireland.

Chapter 8 describes the spatial and temporal incidence of ticks identified in companion animal electronic health records (SAVSNET).

Chapter 9 compares each dataset, their degree of agreement and completeness.

The thesis concludes with a discussion in **Chapter 10,** drawing personal conclusions and recommendations about what surveillance strategy and policy the British government could adopt in the future.

Chapter 2 Dataset selection for analysis

When assessing which health dataset may capture information on Lyme disease patients, and prove beneficial for this research, it was useful to construct a surveillance pyramid (Fig. 2.1).

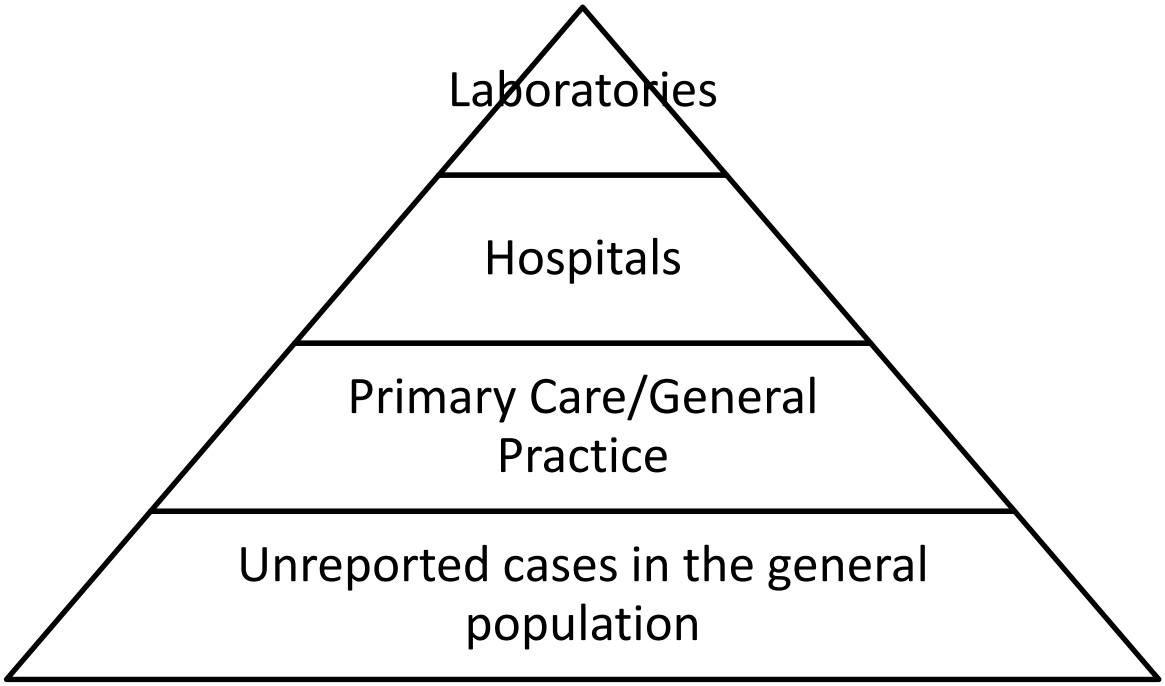


Figure 2.1 A proposed surveillance pyramid for Lyme disease in the United Kingdom

This pyramid displays where patients may interact with the NHS in relation to Lyme disease, with each level being relatively proportional in size to the suspected number of patients at each level. At each of these points health records may be generated that if assessed could help answer the aims of this thesis. Below is a description of the datasets available, and the reasoning behind why certain datasets were used in analysis. It must be first noted that this research took place within the National Institute for Health Research Health Protection Research Unit in Emerging and Zoonotic Infections, which is a partnership between the University of Liverpool, the Liverpool School of Tropical Medicine (LSTM), and Public Health England (PHE). This means that certain datasets are directly accessible through PHE, and as such were prioritised to minimise project costs. Due to the project being funded through PHE, it was important that the predominant focus was exploring datasets that described the situation in England and Wales, however, if information was collected about Scotland and Northern Ireland this was also considered beneficial. It was also beneficial to access data that could easily be made available to PHE, as any adaptations to PHE's surveillance systems as a result of this thesis would require on-going access to the same data.

2.1 Laboratory datasets

In the 2010 Health Protection (Notification) Regulations [77], *Borrelia* spp are classified as a 'Schedule 2 – Causative Agent'. Thus, if any of the species of the *Borrelia* genus capable of causing Lyme disease are identified by a diagnostic laboratory, they should be notified to PHE [78]. The system that captures this reported data is The Second Generation Surveillance System (SGSS). This represents the official pathway for any laboratories (NHS or private) across England, Wales and Northern Ireland to notify PHE of any positive identification of *Borrelia burgdorferi*. SGSS performs surveillance for many other pathogens that can be identified through routine laboratory diagnostics. A recent example of its surveillance activities include a study exploring the relationship between laboratory-confirmed cases of *Campylobacter* and *Cryptosporidium* and daily weather parameters [79].

However, the national Lyme disease surveillance figures in the 'Zoonoses Report' are not based on SGSS reports but instead come from two reference laboratories [70]. These are the Rare and Imported Pathogens Laboratory (RIPL) based at Porton Down, England [80], and the National Lyme Borreliosis Testing Service based in Inverness, Scotland [81,82]. RIPL performs critical laboratory work for the British government; examples of which include performing infectious disease seroprevalence studies of British military personnel in Afghanistan [83], and investigating outbreaks of hantavirus, a potentially emerging public health threat in the UK [84].

Both of the reference laboratories use the two-tier diagnostic protocol described in Chapter 1 and recommended by the NICE guidelines [23]. RIPL provides services for England and Wales, whilst Inverness provides services for Scotland. It is unclear where Northern Ireland's official figures originate from as no laboratory is specifically mentioned in reference to diagnostic sample submission [85]. As this project is supported by PHE, only the RIPL database was used for further analysis. It must be noted that some private laboratories will perform Lyme disease diagnostics; some following NICE guidelines and others not. Results from these will only appear in the SGSS dataset if these laboratories are following the Health Protection Regulations. Any patients that have diagnostic tests performed abroad won't be captured by the two datasets included in this analysis. RIPL and SGSS datasets will be discussed in Chapter 3.

2.2 Hospital datasets

Information about patient care during hospital emergency department visits, inpatients admittance and procedures, and outpatient clinics varies between the four nations of the

UK. All have the potential to describe Lyme disease patients and their care by utilising International Classification of Diseases, 10th Revision (ICD-10) codes [86]. In England, Hospital Episode Statistics (HES) captures data from NHS England emergency departments, admissions, and outpatients [87]. HES were developed to primarily calculate payments, monitor trends in NHS hospital activity patterns, assess effective delivery of care and support policy and accountability. As such, HES was never envisioned to be used as an epidemiological research dataset; nevertheless, this is now one of its main secondary uses [87–91]. In Wales, Patient Episode Database for Wales (PEDW) has collected data on NHS Wales hospital admissions since 1997 [92,93]. No information is collected on emergency departments and outpatient departments. Scotland collects NHS Scotland hospital data on inpatient admissions, inpatient day cases and outpatient, and is published as 'Acute Hospital Activity and NHS Beds Information' [94]. No data is collected from emergency departments. Northern Ireland collects data on outpatients, admissions and day cases as hospital activity statistics [95]. It also collects data on emergency care activity [96]. Both HES and PEDW have been used previously as research datasets. A recent systematic review reviewed 148 research articles that used HES as their data source [88]. Many of these were exploring specific treatment and outcome trends, and 11.5% were exploring the epidemiology of a specific disease or condition. Examples include, a study exploring the admission trends of adverse drug reactions for a ten year period [97], and a study exploring the prevalence and associated mortality of lower limb amputation [98]. Likewise, PEDW has been used as a data source for many epidemiological studies including a study exploring the demographic and socioeconomic inequalities of emergency admission for violence [99], and another describing the rising trend of obstetric anal sphincter injuries [100]. Due to the ease of data access through PHE, the HES and PEDW datasets were taken forward for further analysis. Similarly to the laboratory datasets there are no publicly available datasets that explore information about patient care in private hospitals. These datasets will be discussed in Chapter 4.

2.3 Primary care datasets

Since the 1980s, patient primary care clinical records have been increasingly collected and stored on computer databases rather than traditional paper records [101]. To support this digitalisation, Read codes were developed. Read codes are a hierarchical standard clinical terminology system that encode multiple patient details including; clinical signs, symptoms, laboratory tests and results, diagnoses, therapeutics, surgical procedures, demographics, and a variety of administrative items [101]. These codes are now used by virtually all general practitioners (GPs) in the UK and have enabled the development of the Quality and

Outcomes Framework (QOF), GP commissioning, and GP fundholding; all of which are integral parts of modern UK primary care. Practice management systems (PMS) are computer software used for the day-to-day administration of a GP practice but also the recording of clinical records. In 2011, seven PMSs held 99% of the market share in England [102,103]; equivalent figures are not available for the rest of the UK. The PMSs are; EMIS systems (54.7% share of practices), Vision V.3 (18.1%), SystmOne (17.8%), and the remaining 9.4% is made up of Synergy, Practice Manager, Premiere and EMIS Web [102,103]. For all the databases, information on geographic coverage is restricted to England. The use of PMSs has enabled the creation of large databases of primary care electronic health records (EHRs), which by utilising patient's Read codes can be used for medical research. The owners of the PMSs have enabled the use of these EHR databases for the research community, the four with the greatest number of patient records being the Clinical Practice Research Datalink (CPRD), The Health Improvement Network (THIN), QResearch and ResearchOne. All of these had the potential to be used in this project, below is a brief summary of each system.

CPRD [104] – This database has existed since 1987 (previously the General Practice Research Database), and in 2017 covered around 8% of the UK population, with over 700 contributing practices and heath records from over 17 million patients in its database. It primarily utilises data from Vision and is in the process of adding EMIS and SystmOne practices. The strength of CPRD lies in the fact that it is the best established as a research tool with over 1,800 publications. However, along with THIN, it has one of the most restricted geographical distributions with a lack of significant coverage in the North and East of England [102], with most practices in London, the south and West Midlands. Many studies have explored CPRD data, notable examples include: survival analysis of patients with liver cirrhosis [105], validation of Read codes to identify patients with chronic obstructive pulmonary disease [106], and validation of cases of suicide and self-harm [107]. At the time of planning and design of the work presented in this thesis, the University of Liverpool or PHE did not have a license to access the data, and a data extract just for this project was beyond the realms of the project budget.

THIN [108] – This was established in 2003 and collects records from Vision with around a 60% overlap with CPRD. In 2015 it covered around 6% of the UK population, with over 500 practices and over 11 million patient heath records. It is the second most commonly used database in terms of research with over 500 publications. Two studies using THIN include; a 2013 study exploring gender differences in consultations in UK primary care [109] and a large

study exploring the incidence, prevalence and treatments of patients diagnosed with type 2 diabetes mellitus [110]. The same issues with geographical distribution as CPRD remain. The biggest advantage of this database was that PHE holds an institute THIN license, and so access to the database would have no cost for this project.

QResearch [111] – This is the largest UK database as it collects data from the EMIS system. In 2017 it contained around 1,500 practices and over 30 million patient health records. QResearch has the best geographic coverage with only small isolated regions with no coverage. It is partnered with the University of Nottingham, who is the sole data controller and point of access. Over 200 papers have been published, but the majority have been authored by three University of Nottingham researchers [112]. Examples include: a study exploring the adverse outcomes of antidepressant use in an adult non-geriatric population [113], and a study describing the risk of pneumonia in patients who have been prescribed statins [114]. Dataset extracts are limited to 100,000 patients to non-University of Nottingham users, and a cost would be involved to receive this extract.

ResearchOne [115] – This collects data from the SystmOne PMS and is held in partnership with the University of Leeds . It holds over 15 million active records from primary care from over 2000 GP practices. It has the second greatest geographical coverage; however, there is no presence in parts of the North West, West Midlands, London and the South East. It is a relative new comer as an EHR database, with less than 50 publications and so there is less familiarity with how to utilise its data. Examples of its usage include a paper showing the increased use of opiates and a resultant premature mortality among patients with inflammatory bowel disease [116], and a paper estimating primary care attendance after vaccination [117]. The newness of this dataset has resulted in it being the cheapest cost for data extraction out of the databases identified.

It must be noted that all of these databases have been deemed demographically representative of the UK general population [117–120]. At the time of choosing the appropriate database to analyse for this thesis, the research on geographic coverage had not been published and so the decision was based on ease of access, cost and population coverage. THIN came top in the first two out of these three categories as it sits on the PHE servers, and would be cost neutral to the project. We therefore took THIN forward for analysis. This dataset will be discussed in Chapters 5 and 6.

2.4 Datasets describing unreported cases in the general population

The estimation of the incidence of community-level infections that do not present to a health care professional will always be a challenge. To date, large studies have been performed in disease areas that have much higher morbidity and mortality, and higher general public knowledge and awareness than Lyme disease. This includes infectious intestinal disease, acute respiratory infections, and influenza type illnesses. For infectious intestinal disease a UK population cohort study was performed following 6,836 participants for a year to understand rates of disease in the community [121]. This was then compared GP presentation rates and national surveillance figures. They found 274 (253-295) cases per 1000 person-years in the community compared to 17.7 (14.4-21.8) in general practice and 1.9 reported to national surveillance. This resulted in a community to national surveillance incidence ratio of 147 (136-158) and 9.5 (7.7 to 11.7) between GP and national surveillance. The MoSAIC study was a similar design but following 789 New Yorkers for acute respiratory infections and influenza like illnesses [122]. Only 23.5% of participants who had a disease episode had an associated medical visit, 16.6% to primary care and 4.4% to a hospital emergency department. Both of these studies demonstrate that there is the potential that for any given condition there is a large portion of the population that won't seek professional health care help and advice. Therefore, relying on an incidence figure based solely on a primary care, hospital or laboratory diagnosis is almost always going to underestimate the true incidence of disease.

Estimation of the size of this population in regards to Lyme disease poses a challenge. Lyme disease has a much lower incidence than either gastrointestinal or respiratory disease described above, and a cohort study would therefore need a very large study population and long follow-up period to obtain enough cases. This was deemed beyond the realms of possibility for this thesis both in terms of time and cost. Instead, it was decided to see whether there were datasets that could be used as a proxy to estimating incidence and seasonal trends.

In recent years the increasing global access and utilisation of the internet has led to a huge increase in using internet data as potential sources for disease surveillance and outbreak prediction. The most infamous of these has been Google flu trends [71,123]. However, after its initial well-trumpeted success, errors soon appeared in its outbreak detection. Suggested reasons for this included its vulnerability to overfitting seasonal terms unrelated to flu, and that it did not account for changes in search behaviours over time. Other criticisms were a lack of transparency and replicability with a lot of the research, such that, rather than being

a substitute for more traditional surveillance, it should be seen as a complement [124]. To do research on Google data with a degree of geographical resolution is expensive. Due to the expense and the lack of clear methodology and transparency, it was decided not to explore this dataset in this thesis. However, more accessible internet data sources are social media [125–129].

Social media are websites and applications that allow users to share and create content, and to participate in social networking. Millions of users in the UK use these platforms to network over a wide variety of topics, just one of which may be Lyme disease. In 2017, the most popular UK platforms in terms of users were Facebook (32 million users, 48% of the UK population), YouTube (19.1 million, 29%), and Twitter (20 million, 30%) [130]. These platforms all offer the opportunity to utilise their users' data for research and marketing purposes, and they have all had their potential utility for disease surveillance explored [125–129]. Unless fees are paid, Facebook currently only allows users to explore data related to their own social network. It was therefore ruled out as a platform to explore. YouTube is a mixed media platform primarily based on user-uploaded videos and related comments. This poses a large challenge for research, as hours of video content would need to be watched to draw any conclusions; this was considered too time consuming and beyond the scope of this project [131,132]. Twitter allows users to post and interact with messages of 280 characters (Tweets) that are created by the users. Twitter has enabled users to create databases on all tweets defined by a search term and geographical area. A number of projects have successfully used this platform for disease surveillance and measuring the concern of the general public [129,133–136]. These have included a study exploring the spatiotemporal nature of tweets in relation to the 2012-2013 influenza season in New York [134], and another using Twitter to explore the general publics' concerns and ability to monitor disease activity in a national H1N1 influenza outbreak in the USA [135]. With well-recognised methodology and free access, it provides an excellent resource to use as a database, and will be discussed in Chapter 6.

It has been proposed by numerous authors that companion animals, and dogs in particular, can be used for sentinels for both tick distribution and Lyme disease incidence [34,137–142]. In 2008 the University of Liverpool launched the Small Animal Veterinary Surveillance Network (SAVSNET) [143]. At the time of writing, this surveillance scheme collects EHRs from over 500 veterinary primary care clinics across the UK and laboratory data from most of the largest veterinary diagnostic laboratories in the UK. SAVSNET has previously been used to

describe diseases with a zoonotic potential, in particular salmonellosis [144] and *Babesia spp* [145]. SAVSNET provides the perfect platform to examine whether companion animals infested with ticks can act as a proxy for Lyme disease risk. The SAVSNET data is my final dataset to be analysed and will be discussed in Chapter 8.

2.5 Ethical requirements of data access

During the selection of the above datasets consideration had to be taken about data access and the ethical requirements needed to enable this. Within the NHS, scientific studies and data analysis can be defined as research, service evaluation, clinical audit or surveillance [146,147]. Each of these have varying levels of ethical approval, it is therefore critical to understand how the studies outlined in this thesis are defined. The key determinants to understanding whether a study is defined as research are intent, intervention, allocation, and randomisation [146]. The primary intent of research is to derive generalisable new knowledge, for audit and service evaluation it is to measure standard of care, and for surveillance it is to investigate health issues in a population to improve population health or to investigate outbreak or incidence to help in disease control and prevention. Only research involves an intervention, an allocation of treatment or service by protocol, and randomisation. Using these criteria and in discussion with PHE's Research Ethics and Governance Group, including the local Caldicott Guardian, the analysis for each dataset was categorised accordingly.

RIPL and SGSS collect anonymised patient data for public surveillance under The Health Protection Legislation (England) Guidance 2010 [148]. The analysis of these data was for disease surveillance purposes and there was no intervention, allocation or randomisation. It was therefore defined as surveillance and no ethical approval was needed. These data were provided to the author securely and anonymised in summary tables over the PHE network and were analysed and stored on PHE's secure servers.

HES and PEDW collect anonymised patient data. The analysis was for disease surveillance purposes and there was no intervention, allocation or randomisation. The analysis of these data was therefore defined as surveillance and no ethical approval was needed. HES pseudo-anonymised data can be accessed internally within PHE through the application of a data access form. This ensures that, in this instance, data can only be used for health surveillance and mapping patient pathways. Data can only be accessed through a PHE network. The applicant must have completed NHS data security training and must understand how to maintain patient confidentiality. The application includes a summary of the project and if

probity issues are raised data access would not be granted. Data access was provided for this project. To access PEDW a data access request was sent to NHS Wales Informatics Service [92] clearly stating the data was being analysed for surveillance purposes, the intent to publish the data, and that the data would be stored securely on PHE servers. Access was duly provided, and NHS Wales was acknowledged on the resultant publication.

PHE have a data access agreement in place with IMS Health, the data owners of THIN, to allow pseudo-anonymised non-linked data to sit on their servers. To access the THIN dataset a data access request and study protocol must be submitted to IMS Health's independent Scientific Review Committee. This committee treat all requests as research and as such they go through a rigorous scientific and ethical review process. They are given this remit by the NHS Multi-centre Research Ethical Committee (MERC) and is certified by the Information Commissioner's Office [108]. Access, analysis and publication of the results of the data was granted under the reference number 16THIN103.

In chapter 6 a mixed methods questionnaire was utilised to validate the Lyme disease associated Read codes used in the analysis of THIN data. The questionnaires were targeted at primary care clinicians. Any scientific study involving clinicians is deemed as research and ethical approval was therefore sought. The project was granted REC (NHS Research Ethics Committee) and HRA (Health Research Authority) approval under the IRAS (Integrated Research Application System) project ID: 208815.

All the data provided through Twitter's API (Application Programming Interface) is publicly available. The University of Liverpool's Research Support Office [149] clearly stated that any research that 'involves information freely available in the public domain' does not require research ethics approval. Therefore, research ethics approval was not sought for the methodology described in chapter 7.

The overall ethical approval for the SAVSNET research project was provided by the University of Liverpool research committee (RETH00964). To apply for data access for specific projects a request and project protocol must be sent to the SAVSNET Data Access and Publication Panel. They review the methodology and any resultant publication to ensure robust analysis and to maintain anonymity of any potentially identifiable person. Permission to analyse the data as described in this thesis and the publication were approved by the Data Access and Publication Panel (201601AR002).

2.6 Conclusion

It was decided to take seven datasets forward for analysis. RIPL and SGSS were chosen as laboratory datasets (Chapter 3). HES and PEDW were chosen as hospital datasets (Chapter 4). THIN was chosen as the primary care dataset (Chapter 5). Finally, Twitter (Chapter 7) and SAVSNET (Chapter 8) were chosen as datasets representing unreported cases in the general population.

Chapter 3 Surveillance of Lyme disease in laboratory datasets

Parts of this chapter, relating to the Rare and Imported Pathogens Laboratory data, are currently in press as: Tulloch, J., Semper, A., Brooks, T., Russell, K., Halsby, K., Christley, R., Radford, A., Vivancos, R., Warner, J. (2019). The demographics and geographic distribution of laboratory-confirmed Lyme disease cases in England and Wales (2013-2016): an ecological study *(BMJ Open)*.

3.1 Introduction

In the United Kingdom (UK) Lyme disease is currently not a notifiable disease, but laboratory-confirmed *Borrelia* spp. are notifiable causative organisms [77]. In the 2010 Health Protection (Notification) Regulations [77], *Borrelia* spp are classified as a 'Schedule 2 – Causative Agent'. By obeying Provisions 4.1 of this legislature, 'the operator of a diagnostic laboratory must notify the Health Protection Agency ([sic] Public Health England (PHE), from 2013) in accordance with this regulation where the diagnostic laboratory identifies a causative agent in a human sample', and 'notification must be provided in writing within 7 days'. Thus, if any species of the *Borrelia* genus capable of causing Lyme disease were identified by a diagnostic laboratory, it should be notified to PHE [78].

The Second Generation Surveillance System (SGSS) is a tool developed by PHE, for the electronic reporting, storage and management of information regarding laboratory notifications [150]. It is the recommended method for capturing infectious disease laboratory surveillance data across England, Wales and Northern Ireland. Information captured via SGSS are summarised in Table 3.1.

Table 3.1 Information captured via SGSS, mandatory fields (according to The Health Protection (Notification) Regulations 2010) (69) are marked with an *.

Field	More details
Source lab	
Reference lab	
Reporting lab and address *	Usually source lab, sometimes reference lab
Patient identification *	• Patient's surname and initial * • Patient's hospital number • Patient's NHS number *
Date of Birth (DOB)*	If DOB is unknown, then patient's age
Sex *	
Organism *	Full organism name
Date of onset	Date of onset of illness caused by organism reported
Specimen type(s) *	
Specimen date(s) *	
Identification method	Method used to identify the organism
Patient's home address/residence including postcode *	
Ethnicity *	

SGSS, therefore, represents the official pathway for laboratories across England and Wales to notify PHE of any positive identification of *Borrelia burgdorferi*; some cases from Northern Ireland are also recorded.

Public Health England's current Lyme borreliosis reference diagnostic service has been based at the Rare and Imported Pathogens Laboratory (RIPL) since 2012 [23,151]. It provides specialist advice and diagnostics for Lyme disease to the National Health Service (NHS) in England and Wales. For Lyme disease it offers the current best practice diagnostic protocol for disease confirmation [23,41,42,44,152]. This uses a two-tier methodology: an initial C6 antigen-based ELISA screening test, followed by a confirmatory Western blot. RIPL offers advice from consultants with years of experience with Lyme disease, and is the reference laboratory for England and Wales. As RIPL forms part of PHE, disease notification is an

automated process. The RIPL laboratory-confirmed cases are used to construct PHE's England and Wales Lyme disease surveillance figures [153].

Upon receiving a patient's tissue sample, NHS and private diagnostic laboratories have a choice to make. They can send all their suspect Lyme disease samples directly to RIPL, perform screening tests and send onwards only confirmatory samples, or do both screening and confirmatory tests themselves. In theory these referring laboratories should only report, through SGSS, positives from their own confirmatory tests. However, it is unknown how they report, which of the tests they report, and which tests they perform. Without understanding what the incidence and demographics of RIPL confirmed Lyme disease cases, and SGSS notifications are, and how they compare, it would be challenging to get an understanding of the laboratory-confirmed incidence of Lyme disease in England and Wales.

Historical information regarding the demographics, geographic distribution, and socioeconomic status of Lyme disease cases in England and Wales is limited. Laboratory surveillance data published in 2000 describe an equal sex ratio at all ages; however, numbers were not provided and statistical comparison was not performed [154]. The authors of this paper described a bimodal age distribution with peaks in childhood and at 45-64 years old. They also described a tendency for cases in southern England, especially around the New Forest. However, this data may not reflect the current distribution of Lyme disease cases in England and Wales. More current data is urgently needed to enable targeted public health messaging and intervention strategies.

Work to explore the association between socioeconomic status and Lyme disease incidence is limited. In the United States of America (USA), persons were found to be at greatest risk of Lyme disease if they lived in the highest or lowest socially vulnerable areas [155]. Two studies found a relationship between Lyme disease incidence and median annual household income, with incidence peaking at around 80,000 USD [156,157]. However, a consistent relationship between the socioeconomic state of an individual and their Lyme disease acquisition risk has yet to emerge. No in-depth research has been published in Europe investigating the socioeconomics of the Lyme disease patient cohort.

The aim of this chapter is firstly to describe the incidence, socio-demographics and geographic distribution of laboratory-confirmed Lyme disease cases captured through routine Lyme disease surveillance in both the RIPL and SGSS datasets. Secondly the two

datasets will be compared, and formal recommendations made about how PHE's laboratory surveillance could be improved.

3.2 Methodology

3.2.1 RIPL Data Analysis

A retrospective analysis was performed using data extracted from the PHE Rare and Imported Pathogens Laboratory's (RIPL) laboratory information management system (LIMS), between 1st January 2013 and 31st December 2016, for laboratory-confirmed Lyme disease cases, the same data as used for PHE's Zoonoses Report [153]. The RIPL LIMS contains information provided on the Lyme disease referral form submitted at the time of sample submission and any additional information provided by clinicians during case follow up and management [158]. The form captures information on the age, gender, location, clinical symptoms and travel history of the patient. Data were cleaned, and duplicates were removed where necessary.

Annual Lyme disease incidence estimates were calculated using the Office for National Statistics (ONS) mid-year population estimates as the denominator population [159]. A Chi-squared test for trend and a Chi-squared test for departure from the trend were used to analyse trends in incidence. Cases were stratified by age and gender. Using binomial tests, the null hypothesis that there was no difference in case numbers between males and females was tested within differing age bands, and overall.

Geographical information was collated based on (1) the regional origin of a diagnostic sample (usually a hospital microbiology department) consisting of eight PHE regions, and Wales as a whole [160], and (2) the postcode area of the patient. These were used to calculate average annual incidence for the study period. In an attempt to account for the unknown distance between a patient's home address and where they were bitten, the disease incidence map for postcode area was smoothed using a k-nearest neighbours (k-NN) approach [161–163]. k is defined as the number of neighbours used for smoothing and is equal to the square root of the total number of discrete geographical areas rounded to the nearest whole odd number (i.e. 105 postcode areas, its square root being 10.2, therefore k=11). Exploratory spatial data analysis (ESDA) [164,165] was used to explore the spatial autocorrelation of the postcode area incidence map. Global and local Moran's I values were calculated, and a LISA (Local Indicators of Spatial Association) significance map constructed to highlight any significant clusters. In both the k-NN smoothing and Moran's I calculations, a queen adjacency matrix was used. A queen contiguity was chosen as, unlike a rook contiguity, it defines neighbours

as being on both the edges and vertices of a polygon, and therefore better captures all true neighbours of a postcode area [166].

Patient postcode was linked to ONS socioeconomic data [159], enabling a description of the socioeconomic characteristics of the population in which a Lyme disease case was resident. If no patient postcode was recorded, these cases were excluded from the analysis. Socioeconomic status is reported through the English Indices of Deprivation (EID) 2015 [167] and the Welsh Index of Multiple Deprivation (WIMD) 2014 [168].

The EID ranks 32,844 geographies ('Lower super output areas') containing between 1000 and 3000 population and groups these in to deciles where 1 represents the areas with the highest levels of deprivation and 10 the lowest. The WIMD is ranked in a similar manner but is then grouped into the following categories; the 10% with the greatest deprivation, moving up through decreasing levels of deprivation in intervals from 10-20%, 20-30%, 30-50%, and the 50% least deprived areas. The EID and WIMD categorize each geographical area with a variety of deprivation domain scores to build a summary index figure (Table 3.2). The ONS classifies these same geographies, in England and Wales, with a rural urban classification [169].

Table 3.2 Summary of the English Indices of Deprivation 2015 (156) and Welsh Index of Multiple Deprivation (157) domains, and their weighting to calculate an Index of Multiple Deprivation

English Indices of Deprivation (EID) Domain	**Description**	**Weighting for construction of Index of Multiple Deprivation (IMD)**
Income Deprivation Domain	Proportion of population experiencing deprivation due to low income	22.5%
Employment Deprivation Domain	Proportion of working age population excluded from the labor market	22.5%
Education, Skills and Training Deprivation Domain	Measures the lack of attainment and skills in the local population	13.5%
Health Deprivation and Disability Domain	Measures the risk of premature death and the impairment of quality of life through poor physical or mental health	13.5%
Crime Domain	Measures the risk of personal and material victimization at local level	9.3%
Barriers to Housing and Services Domain	Measures the physical and financial accessibility of housing and local services (schools, supermarkets, primary care and post offices)	9.3%
Living Environment Deprivation Domain	Measures the quality of the local environment (housing, air quality and road traffic accidents)	9.3%
Index of Multiple Deprivation (IMD)	**Overall measure of deprivation constructed by the weighted sum of the above domains**	
Welsh Index of Multiple Deprivation (WIMD) Domain	**Description**	**Weighting for construction of Welsh Index of Multiple Deprivation (WIMD)**
Income Domain	Proportion of population experiencing deprivation due to low income	23.5%
Employment Domain	Proportion of working age population excluded from the labor market	23.5%
Health Domain	Measures the lack of good health	14.0%
Education Domain	Measures the extent of deprivation relating to education, training and skills	14.0%
Access to Services Domain	Measures deprivation due to a households inability to access services considered necessary for day to day living.	10.0%
Community Safety Domain	Measures deprivation relating to living in a safe community	5.0%
Physical Environment Domain	Measures factors in the local area that may impact on wellbeing or quality of life	5.0%
Housing Domain	Measures deprivation through lack of adequate housing	5.0%
Welsh Index of Multiple Deprivation (WIMD)	**Overall measure of deprivation constructed by the weighted sum of the above domains**	

Postcode area case count data were matched independently to the EID and WIMD, and rural urban classification. As EID and WIMD were on a discrete ordinal scale, Spearman's rank correlation was used to calculate the correlation between the number of cases and deprivation score. The proportion of cases with their home addresses located in either a rural or urban area, were compared to the national rural urban classification from the ONS [169]. This was performed using a Chi-squared test of independence for both English and Welsh data.

3.2.2 SGSS Data Analysis

A retrospective analysis was performed using data extracted from PHE's Second Generation Surveillance System (SGSS) for notifications for *Borrelia* spp as a causal organism, with specimen dates between 1st January 2000 and 31st December 2016. Data was pseudoanonymised prior to extraction. Information captured included; organism name, specimen type, specimen date, local authority of patient's residence, patient's age in years, patient's sex, and source lab name. Information relating to patient's geography was limited to local authority level to maintain patient anonymity.

Annual Lyme disease incidence estimates were calculated, using the Office for National Statistics (ONS) mid-year population estimates as the denominator population [159]. A Chi-squared test for trend and a Chi-squared test for departure from the trend were used to analyse trends in incidence. Cases were stratified by age and gender. Using binomial tests, the null hypothesis that there was no difference in case numbers between males and females was tested within differing age bands, and overall.

Geographical information was collated based on the local authority (or District Council for Northern Ireland) of a patient's home address or residence. Northern Irish data was captured using the pre-2015 district boundaries (n=26). These were used to calculate average annual incidences for the whole study period.

Information regarding source laboratory name and specimen type were described descriptively.

All statistical and spatial analyses were carried out using R language (version 3.2.0) (R Core Team 2015). Results were deemed significant where $p<0.05$.

3.3 Results

3.3.1 RIPL Results

In total 3,986 unique cases (3,893 cases in England and 93 in Wales) meeting a serological diagnosis of Lyme disease were identified in the RIPL LIMS between 1st January 2013 and 31st December 2016. Of these, 98.7% (n=3,935) had complete records for date of submission, gender and age.

The annual incidence of laboratory-confirmed Lyme disease cases in England and Wales rose from 1.62 per 100,000 population in 2013, to 1.95 in 2016. These figures are identical to PHE's official incidence figures as they used the same data source [170]. There was evidence of an overall association between incidence and year (χ^2=43.13, p<0.001). This association took the form of a trend with increasing incidence each year (χ^2=30.17, p<0.001). Departures from the trend were significant (χ^2=43.1-30.1=12.96, p<0.001), as shown by the fall in incidence in 2014. There was marked seasonality, with the peak numbers of cases being diagnosed in the summer months each year (Fig. 3.1).

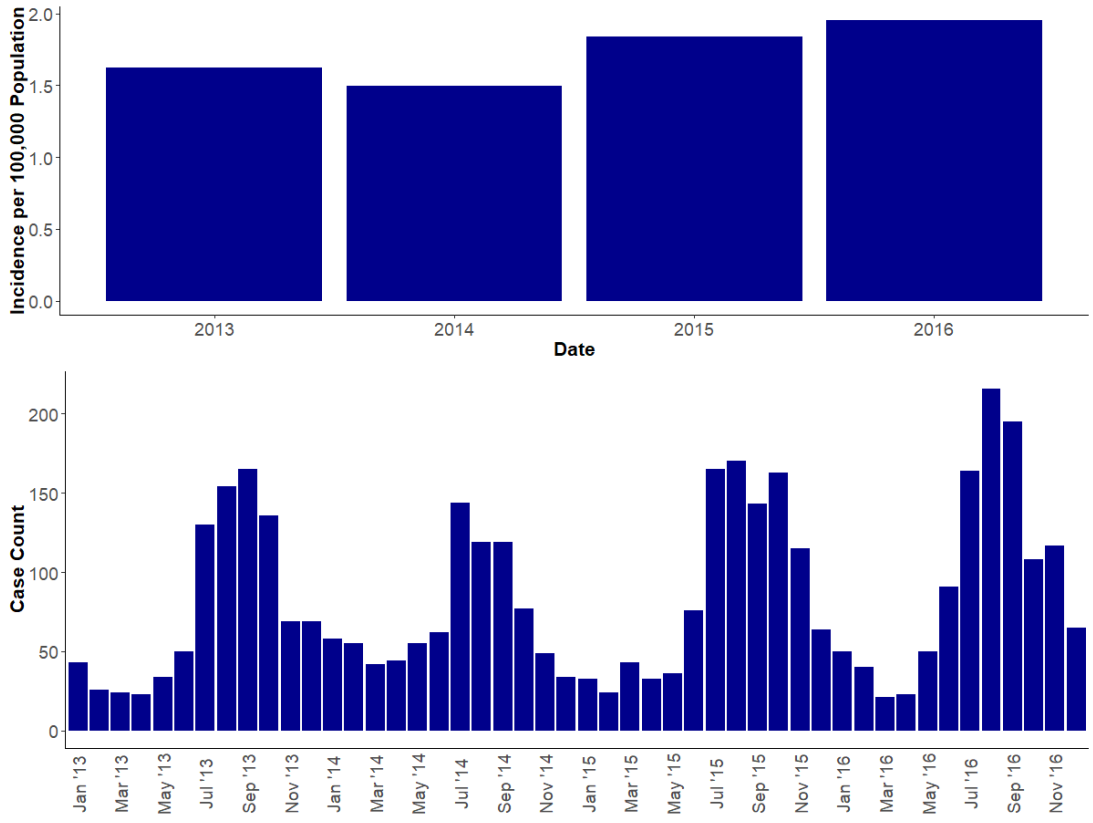


Figure 3.1 The annual incidence of laboratory-confirmed Lyme disease cases in England and Wales (2013 -2016), and the number of cases per month.

Across all ages there were significantly more male (n=2,096) than female (n=1,839) cases (p<0.001), with a bimodal age distribution, with peaks at 6-10 and 61-65 year age bands (Fig.

3.2). Grouping the data in 5-year age bands, there were significantly more men than women in the 6-10 (p=0.03), 11-15 (p=0.03), 36-40 (p=0.01), 41-45 (p=0.02), and 46-50 (p=0.04) age groups.

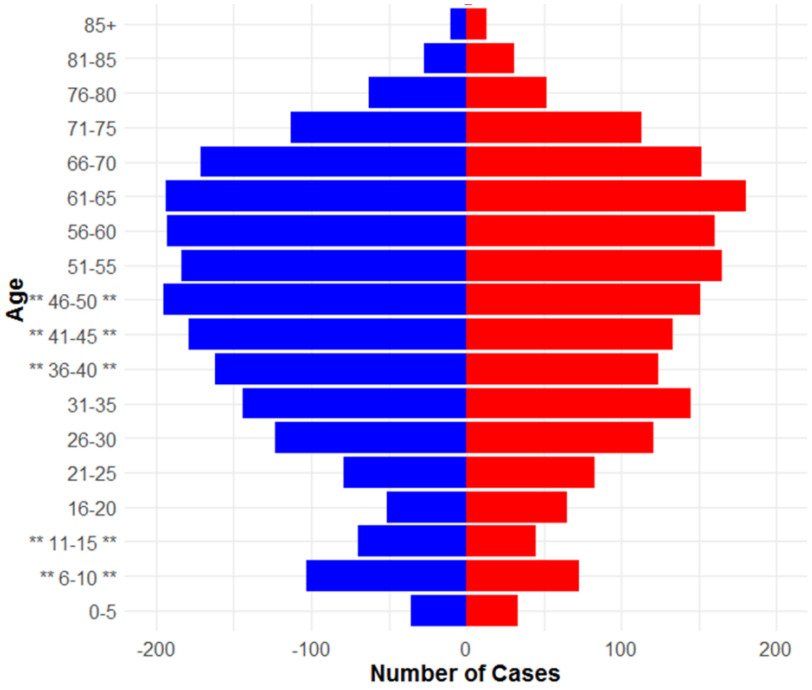


Figure 3.2 Population demographics of laboratory-confirmed Lyme disease cases in England and Wales, 2013-2016. (Asterisks represent age bands with a significant difference between genders. Male = Blue, Female = Red)

Data were available about PHE regions for 99.9% (n=3,985) of the study population (Fig. 3.3), and about patient residence postcode area for 58.2% (n=2,321). The South West PHE region had the highest incidence of Lyme disease in England and Wales; none of the PHE regions, nor Wales, reported zero cases. The postcode areas with the highest average annual incidence of Lyme disease were Southampton (11.65 cases per 100,000 per year), Salisbury (10.75), Bournemouth (5.62), Reading (4.59), Dorchester (4.57), Guildford (4.31), Taunton (2.79), Torquay (2.75), Brighton (1.96), and Bath (1.84) (Fig. 3.4). These areas are all in southern England. Only four postcode areas had no laboratory-confirmed cases in the four year surveillance period (Fig. 3.4), namely Dartford, Eastern Central London, Hull, and Western Central London.

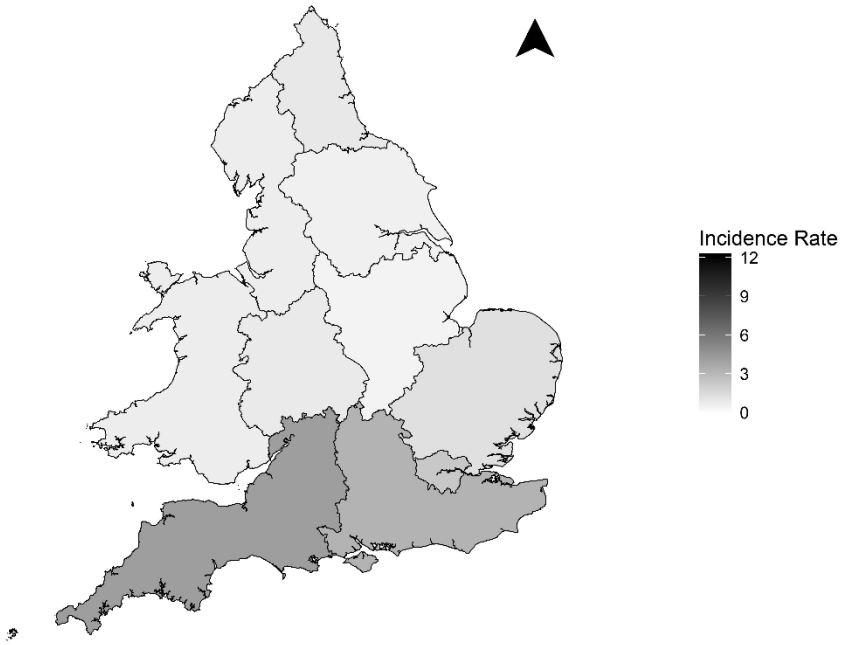


Figure 3.3 The average incidence (cases per 100,000 per year) of laboratory-confirmed Lyme disease in England and Wales (2013-16) in Public Health England regions and Wales (n = 3,985)

Figure 3.4 The average incidence (cases per 100,000 per year) of laboratory-confirmed Lyme disease in England and Wales (2013-16) by patient postcode area (n = 2,321). Highest postcode areas are labelled accordingly; SO-Southampton, SP-Salisbury, BH-Bournemouth, RG-Reading, DT-Dorchester, GU-Guildford, TA-Taunton, TQ-Torquay, BN-Brighton, BA-Bath. Areas with no cases are labelled in red; DA-Dartford, EC-East Central London, HU-Hull, WC-Western Central London.

The smoothed data showed areas of highest incidence to be located in southern-central England (Fig. 3.5). There was significant spatial autocorrelation, the global Moran's I was 0.56

(p=0.01), indicating that postcode areas with similar incidence are clustered together. LISA mapping identified six areas as significant clusters of high incidence (Fig. 3.6); Southampton, Salisbury, Bournemouth, Reading, Dorchester, and Guildford (for all p<0.001).

Figure 3.5 The average incidence (cases per 100,000 per year) of laboratory-confirmed Lyme disease in England and Wales (2013-16) by smoothed patient postcode area.

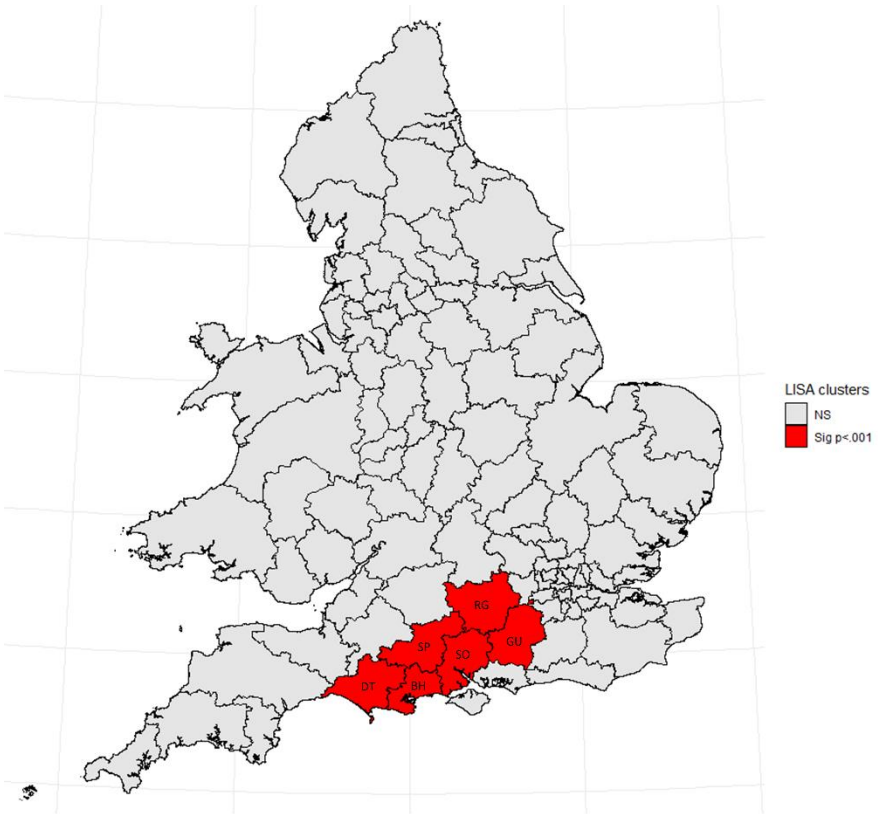


Figure 3.6 LISA map of significant incidence clusters of laboratory-confirmed Lyme disease in England and Wales (2013-16)

Using patient residence postcode data, it was possible to match 55.6% (n=2,165) of English records to the English Indices of Deprivation and 98.2% (n=92) of Welsh records to the Welsh Index of Multiple Deprivation (WIMD). An overall significant positive correlation between the number of cases and Index of Multiple Deprivation (IMD) decile was observed (ρ=0.96, p<0.001), with more Lyme disease cases found in less deprived areas (Fig. 3.7).

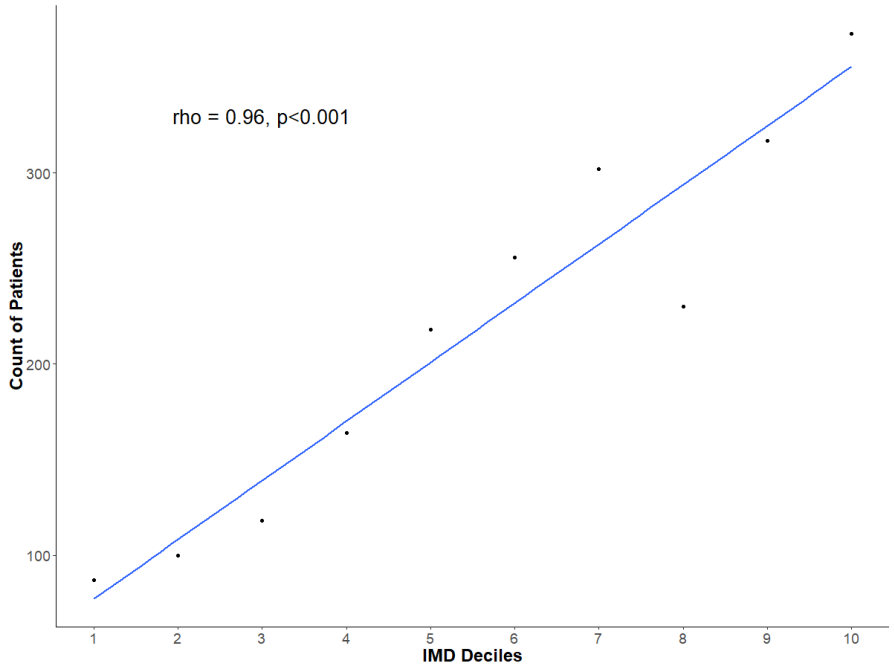


Figure 3.7 Relationship between laboratory-confirmed Lyme disease case numbers (2013-2016) in England and the English Indices of Deprivation 2015.

This significant positive correlation was seen across all domains of deprivation, except the 'Barriers to Housing and Services Domain' where this trend was reversed (ρ=-0.88, p=0.002) and the 'Living Environment Deprivation Domain' where there was no significant correlation (ρ=0.2, p=0.58) (Fig. 3.8). An overall significant positive correlation between the number of cases and WIMD rank was observed (ρ=0.89, p=0.04), with more Lyme disease cases found in the least deprived areas.

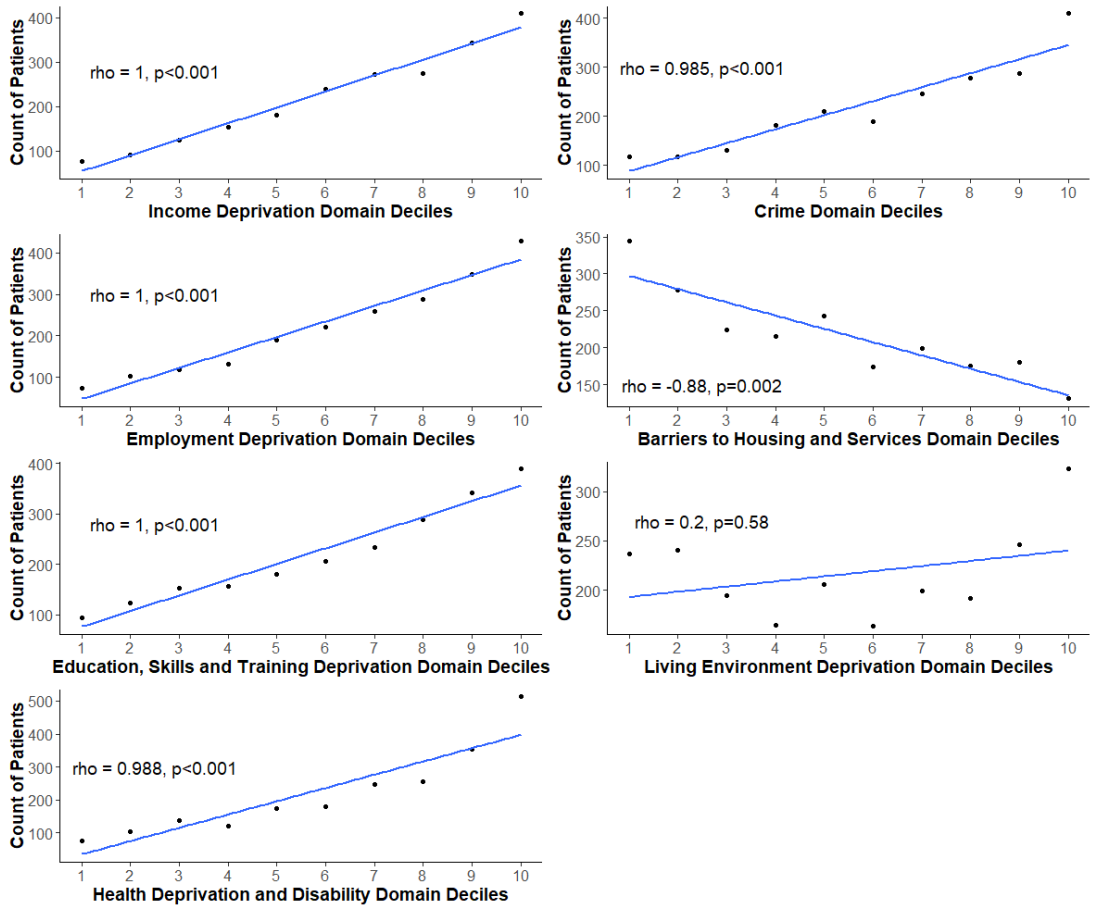


Figure 3.8 Relationship between laboratory-confirmed Lyme disease case numbers (2013-2016) in England and the component measures of the English Indices of Deprivation 2015.

When compared to the national population, the study population was disproportionately more likely to live in a rural area, for both English (p<0.001) and Welsh (p<0.001) sections of the study population (Table 3.3).

Table 3.3 The rural urban classification of laboratory-confirmed cases of Lyme disease in England and Wales (2013-2016) compared to the national census population

Category	Percentage of English Study Population	Percentage of Welsh Study Population	Percentage of 2015 census population
Rural	34.3% (n=743)	47.8% (n=44)	17.9%
Urban	65.7% (n=1,422)	52.2% (n=48)	82.1%

3.3.2 SGSS Results

In total 4,152 unique cases were identified in SGSS which had *Borrelia* spp as the causal organism between 1st January 2000 and the 31st December 2016. Of these, 3,802 (91.6%) of cases had the causal organism specifically named as *Borrelia burgdorferi*. The remaining 8.4% cases were defined simply as *Borrelia* spp and could be part of the relapsing fever group rather than Lyme disease [78]. Analysis was therefore restricted to *Borrelia burgdorferi* cases as these cases could be confirmed as having a Lyme disease causing pathogen. Complete

records were available for 97.7% (n = 3,715) of cases based on the key factors of specimen date, gender and age.

The annual incidence of SGSS reported Lyme disease cases in England and Wales rose from 0.21 per 100,000 population in 2000, to 0.78 in 2016. There was evidence of an overall association between incidence and year (χ^2=903.41, p<0.001). This association took the form of a trend with increasing incidence each year (χ^2=764.16, p<0.001). Departures from the trend were significant (χ^2=139.25, p<0.001) (Fig. 3.9), most notably in 2011 and 2014.

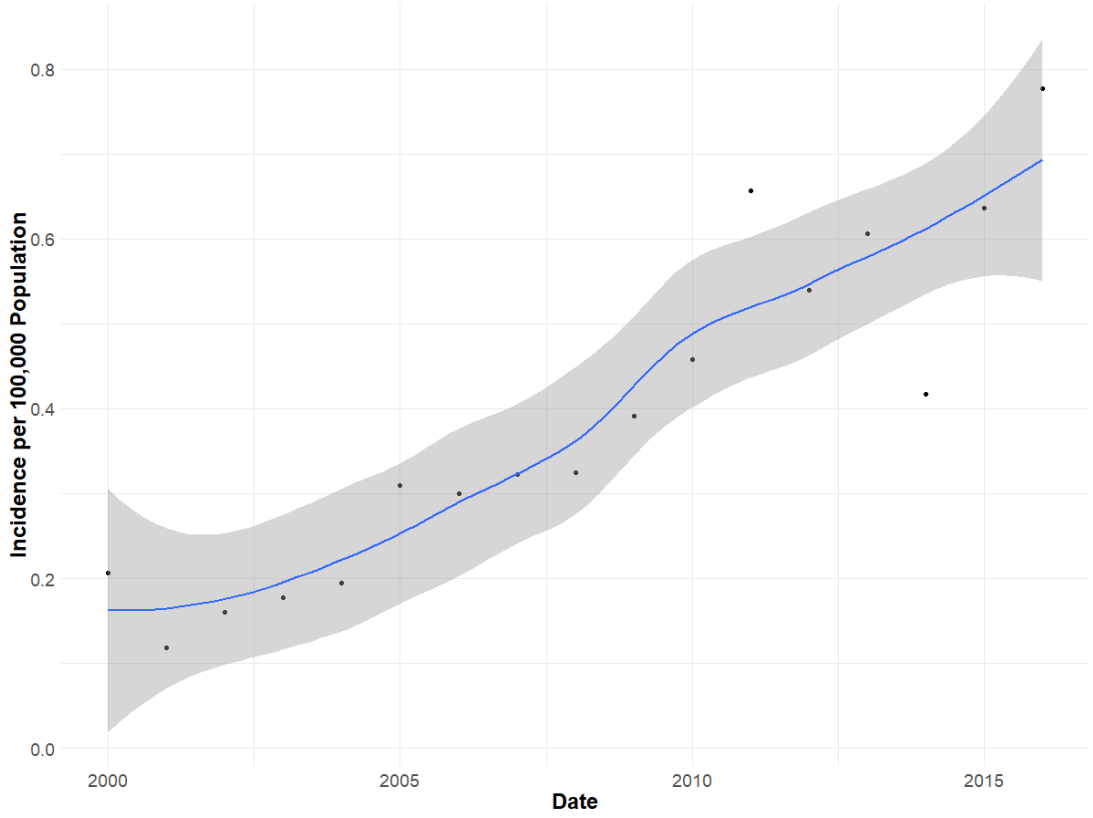


Figure 3.9 The annual incidence of Borrelia burgdorferi patients in England, Wales and Northern Ireland, based on SGSS data.

There was marked seasonality, with the peak numbers of cases being diagnosed in the summer months each year (Fig. 3.10).

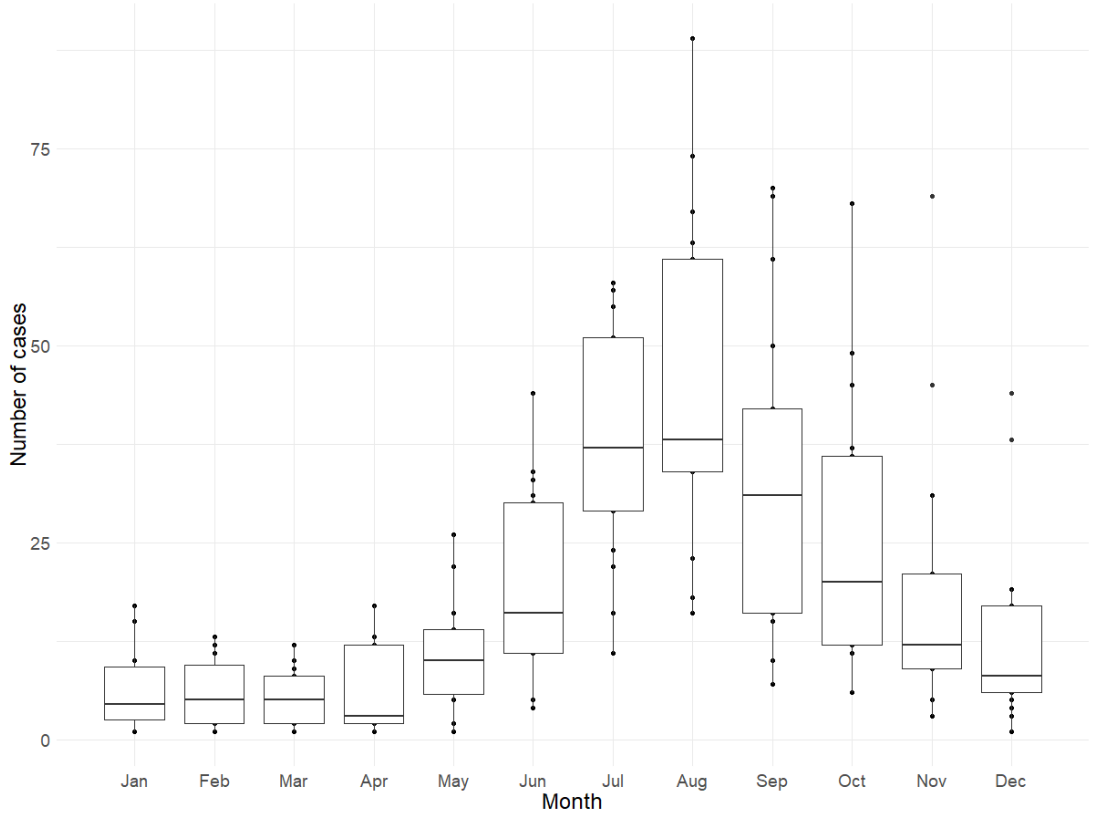


Figure 3.10 The number of Borrelia burgdorferi cases, by month, each year (2000 – 2016) within SGSS.

There was no significant difference between the numbers of men (n=1,876) and women (n=1,839) (p=0.40), an age band comparison analysis was therefore not performed. A bimodal age distribution was seen with peaks at the 6 to 10 and 56 to 60 year age bands (Fig. 3.11).

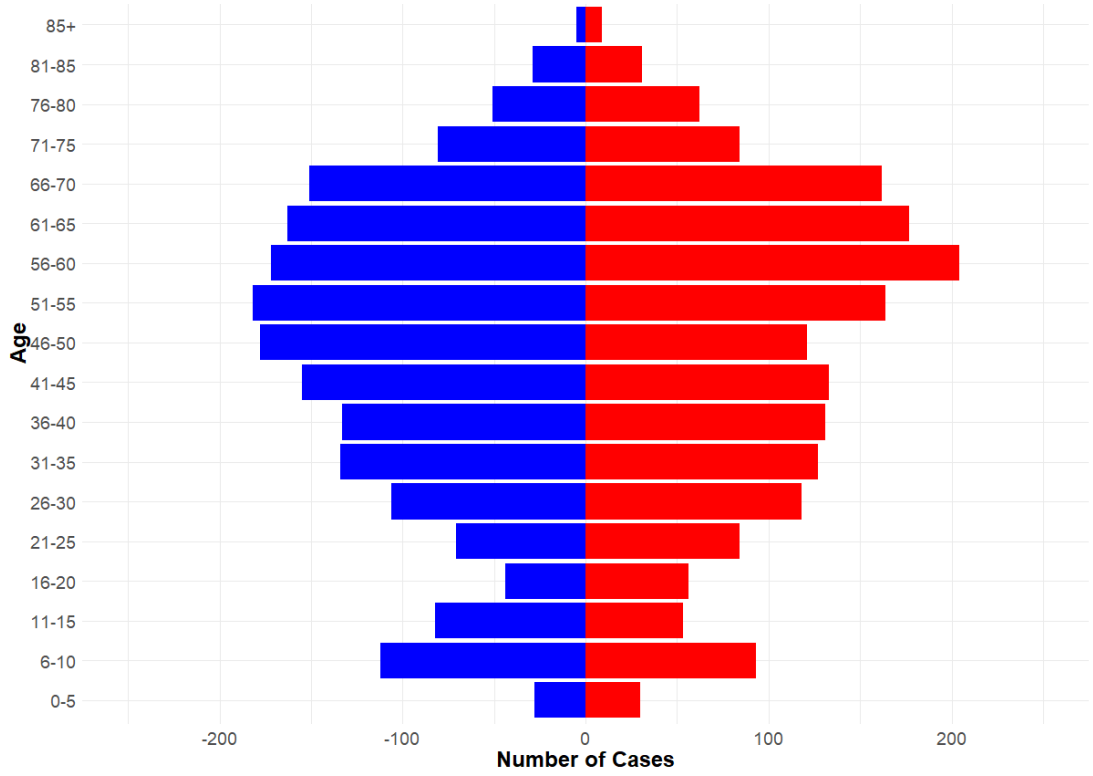


Figure 3.11 Population pyramid for Borrelia burgdorferi patients identified in SGSS (2000 – 2016), (Male = Blue, Female = Red)

All records had local authority information relating to the patient's place of residence (Fig. 3.12). The local authorities with the highest incidence were Exeter with 9.29 cases per 100,000 per year, West Dorset (8.87), West Somerset (7.01), East Devon (6.09), Taunton Deane (4.85), West Devon (4.54), Mid Devon (4.14), East Dorset (4.11), Camden (3.61), and North Devon (3.56). These are all, except Camden, found in south-west England.

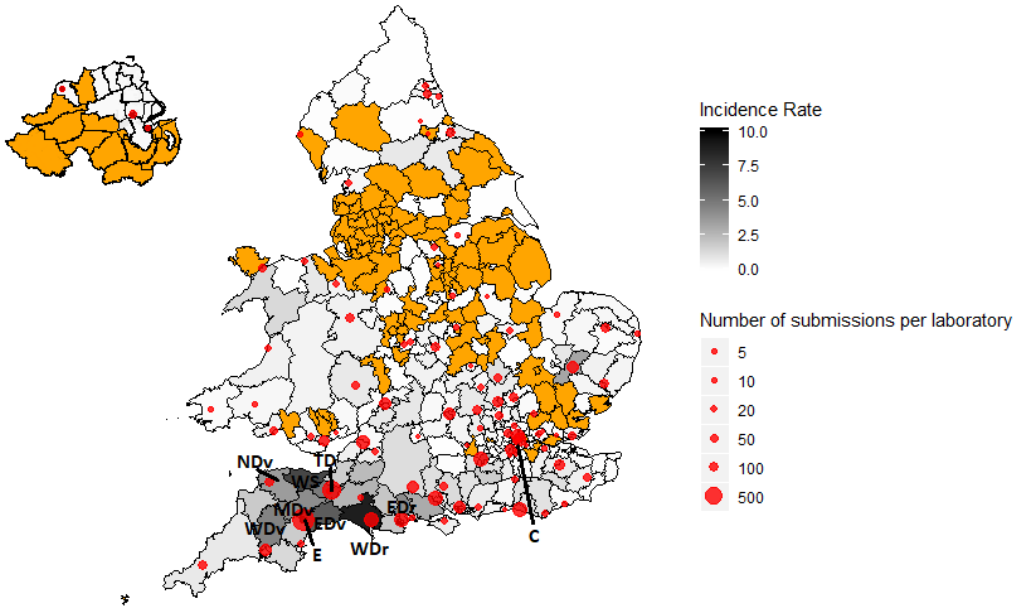


Figure 3.12 The average annual incidence rate (cases/100,000/year) of Borrelia burgdorferi within SGSS, and the number of submissions per laboratory (2000–2016).

(Local authorities with no submissions are in orange. E=Exeter, WDr=West Dorset, WS=West Somerset, EDv=East Devon, TD=Taunton Deane, WDv=West Devon, MDv=Mid Devon, EDr=East Dorset, C=Camden, NDv=North Devon)

Fourteen (53.9%) of Northern Irish districts (n=26), seven (31.8%) of Welsh local authorities (n=22), and 100 (30.7%) of English local authorities (n=326) reported no cases through SGSS. 49.4% (n=1,878) of specimens were blood, 46.2% (n=1,758) were serum, 0.7% (n=26) were cerebrospinal fluid, 3.5% (n=133) unknown, and 0.2% (n=7) were another specimen type.

3.4 Comparison between the datasets

Before a comparison can be made between the incidences and demographics of the Lyme disease populations in the SGSS and RIPL datasets, some important caveats most be highlighted. Firstly, RIPL only captures data from two nations, England and Wales, whereas SGSS additionally captures Northern Ireland. Secondly, they cover different, yet overlapping, time periods. RIPL covers 2013 to 2016, whereas SGSS covers 2000 to 2016. Finally, within the laboratories that submitted information to SGSS, which are a mix of NHS and private organisations, the diagnostic tests that are performed are not being reported. The recommended testing protocol for confirming a case of Lyme disease is a two-tier process, involving an initial screening test followed up with a confirmatory test as per current best guidance [23,41,44]. This is the process used by RIPL. If the precise test performed by the submitting lab is unknown, it is unknown what a positive case refers to, a screening test, the reporting lab's confirmatory test, or a positive result of RIPL's two tier test being reported by the reporting lab. Highlighting this, an internal PHE report assessed the reporting of SGSS

and its overlap with RIPL in 2015 [171]. They found that 54.4% of cases matched cases in RIPL and were positive confirmatory results, 36.0% of cases matched cases in RIPL and were confirmed as negative and therefore represented reporting of positive screening tests by the submitting lab, and 9.6% didn't match with RIPL. This implies that submitting laboratories have very different habits around what they report to SGSS, thus bringing a high level of doubt to the validity and usefulness of the SGSS data regarding Lyme disease. However, it is still worth comparing the two datasets to see if any of the similarities or differences can aid in the improvement of the current reporting and surveillance of Lyme disease.

Overall there were 184 less cases in SGSS than the RIPL dataset. Remembering that the assessment period for SGSS was twelve years longer than RIPL and included cases from Northern Ireland, this is a large overall underestimate of cases. The general bimodal shape of the population pyramid was similar between datasets, both having a pre-pubescent peak at the 6-10 year age band, and SGSS's main peak was at 56-60, whilst RIPL peaked at 61-65. The main difference between the demographics was that RIPL has significantly more men than women, whilst in SGSS there were no significant differences. As previously mentioned, in England and Wales, Lyme disease incidence in men and women has historically been similar [154]. As the validity of SGSS is questionable little can be drawn between the differences in sex; other datasets within this project would be better placed to answer if there is any disparity between the sexes.

When observing the difference between the PHE data (previously reported national figures, including RIPL data from 2012 [153]) and SGSS, the incidence of the PHE data is consistently higher. To compare this formally, a mean incidence rate ratio for the study period was calculated, 3.68 (95% CI 2.16, 5.21). The incidence rate ratio and its confidence intervals show that there is a large degree of variation between the incidences reported by both systems. The difference and variability are likely due to the inconsistent and varied reporting of submitting laboratories. The observed seasonality of SGSS matches that of RIPL, with peaks in summer months, usually peaking in August.

Despite the two datasets reporting different geographical units, postcode area in RIPL and local authority in SGSS, marked differences can still be seen. Discussions will be based solely on England and Wales as there is no Northern Irish data captured by RIPL. In SGSS 30.7% of geographical units have no cases of Lyme disease, compared to 3.7% in RIPL. In SGSS, these are predominantly in the North of England and Midlands. The lack of reporting laboratories in these areas (Fig. 3.8), and the presence of cases in these areas in the RIPL dataset, suggest

that either they are not reporting any cases of their own diagnostic testing as per the Health Protection (Notification) regulations [77], or they are submitting samples to RIPL and assuming that RIPL will notify any resultant positive cases. The areas in SGSS that do have cases show a hot spot of higher incidence in southern England, but they lie further west than the hotspot seen in the RIPL data. It is likely that these areas do have a high incidence of cases, but there may be a degree of over-reporting compared to RIPL. It is possible that they are reporting the number of positives based on a screening test rather than a confirmatory test as RIPL does. Without an audit of the notification and diagnostic practices of all diagnostic laboratories it would be difficult to establish how reliable these incidence figures are.

To explore the geographical relationships between the datasets further, the RIPL postcode data was transformed to local authority data. Average annual incidence of each local authority for each dataset was plotted against one another, with linear regression performed (Fig. 3.13).

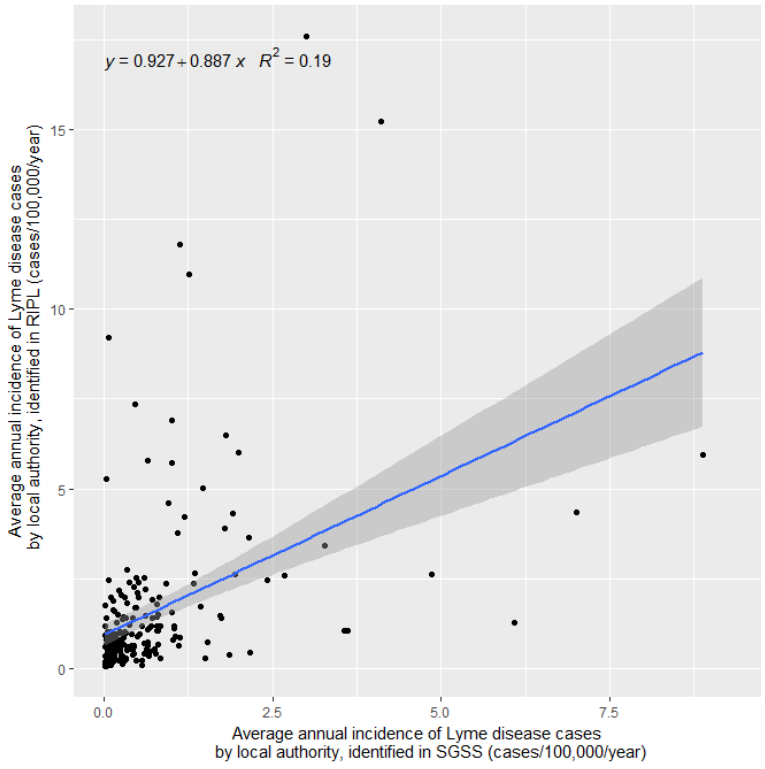


Figure 3.13 A comparison of average annual incidence of laboratory Lyme disease cases, at local authority resolution, between the RIPL and SGSS datasets

A significant association between the two datasets was seen ($p<0.05$), however the r^2 value was low ($r^2=0.19$). Geographical concordance of significant clusters of high and low incidence was explored by constructing a bivariate LISA (Local Indicators of Spatial Association) plot

[164,165] (Fig. 3.14).

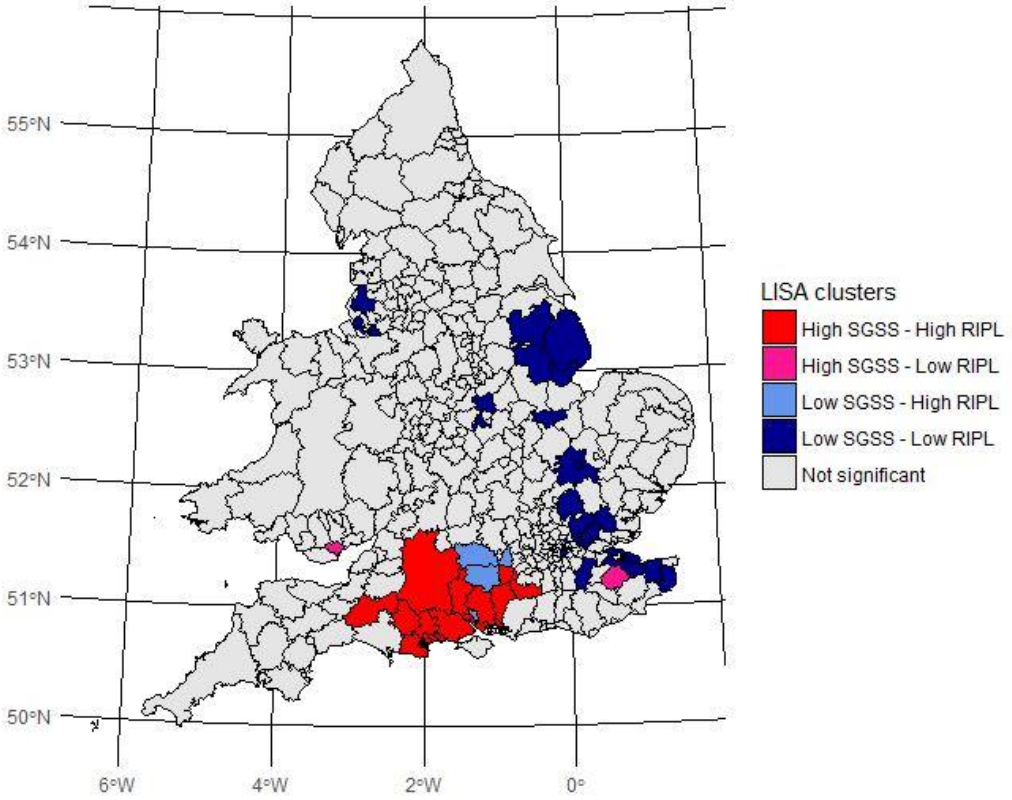


Figure 3.14 Bivariate LISA plot of Lyme disease laboratory cases incidence in the SGSS and RIPL datasets.

A global Moran's I score was calculated and tested for significance. This resulted in I=0.28, and $p<0.05$. This suggests that the spatial distribution of incidence were similar, with high and low values in the dataset being more spatially clustered than would be expected if underlying spatial processes were random. The plot shows significant concordance of high incidence clusters in southern England, with scattering of concordant low incidence areas across the country. The areas of statistically significant discordance are likely to do with differences in regional reporting habits in SGSS, as previously discussed. The large degree of non-significance shows that incidence for both datasets does not differ from a randomly generated distribution of incidence.

One area of note with a high incidence, in the SGSS data, was Camden, a local authority in central London. The equivalent area in RIPL had a low incidence. On further examination of the data, almost all the samples were submitted from The Doctors Laboratory, a private commercial diagnostic laboratory [172]. They reported 132 cases from across England only in 2013 and 2015; 98 of these were from Camden. This laboratory performs both screening ELISAs and confirmatory immunoblot tests. If these are the same tests performed by RIPL,

and they are reporting confirmed positives to SGSS, then national incidence figures will be missing these patients, as PHE only reports cases from RIPL. If this is true, then it needs to be understood why and where patients in central London are acquiring Lyme disease. These are private (non-NHS) testing services; it could be concluded that patients submitting samples here are from a more affluent background. Lyme disease has been found to be present in ticks in London parks [173], these cases could represent this transmission. These cases may have also acquired infection on national or international travel. These cases need to be followed up, as they may represent transmission locations and cases that are unreported in the national literature.

The majority of tissues sampled for diagnostics were appropriate following NICE diagnostic guidelines [23]. However, the non-recommended specimen types included were; skin biopsies, faeces, semen, sputum and urine. They are not recommended as tests relating to these tissue types have a very low sensitivity and specificity [23]. These samples come from a mix of NHS and private laboratories, and so the potential for the cynical use of unverified diagnostic tests for profit is limited. It is more likely that continued professional development is needed in some microbiological units to educate on the most appropriate procedures for Lyme disease, the publication of NICE guidelines should assist with this.

3.5 Discussion of the RIPL dataset

Between 2013 and 2016 there was a significant, but small, increase in annual incidence of cases of confirmed Lyme disease, with a seasonality that matched previous publications and has been well documented [174]. The observed seasonality closely matches *I.ricinus* tick population dynamics in the UK, which annually peak around June and July [24,142]. Concerns have been raised about how the expansion of tick habitats due to changes in land use and management, and climate change, may be increasing the risk of Lyme disease infection [26,175]. Although the incidence of confirmed cases increased over the study period, there was significant deviation from the trend, most notably in 2014. The reasons behind this variable, but increasing, incidence of Lyme disease are likely to be multifactorial and may include raised public and practitioner awareness, variable weather patterns causing alterations in tick abundance and/or carriage of *B. burgdorferi s.l.,* and changes in human activity and behaviour.

The RIPL data displayed a bimodal age distribution, with peaks at 6-10 and 61-65 years, and an overall predominance of males. This bimodal distribution has been reported in other European countries [176–178], and matches previous UK studies [154,179]. However, the

predominance of males in the current study population does not concur with other European studies, where women are over-represented [176–178]. In the USA, Lyme disease is more prevalent in males compared to females less than 60 years old, and equal or higher in women above 60 than among men [21]. In contrast, more men were hospitalised in France due to Lyme disease and more women were diagnosed by general practitioners [180]. Historically, in England and Wales, Lyme disease incidence in men and women has been similar [154,179]. The male predominance in the RIPL data may be due to the difference in health seeking behaviour between genders, with women more likely to seek healthcare at early stages of illness [109]. By presenting at later stages of Lyme disease, when pathognomonic signs may have waned, male cases may require laboratory confirmation more frequently. Further work is needed to establish the causes behind these gender differences and whether they are related to environmental or behavioral risk factors, such as occupation, leisure activities, or differences in health seeking behaviours.

There was geographical variation in Lyme disease incidence across patient residence postcode area in England and Wales, based on 58.2% of laboratory-confirmed cases. The global Moran's I statistic showed that there was significant positive spatial autocorrelation, and clusters of high incidence were found in southern England. This area includes the New Forest National Park, the South Downs National Park, Salisbury Plain, Cranborne Chase Area of Outstanding Natural Beauty (AONB), Dorset AONB and Purbeck Heritage Coast. These are all popular destinations for outdoor activities and are in southern England where the Lyme disease vector *I.ricinus* is most prevalent [24,137,175]. The exposure risk from ticks is likely to be higher in these areas than other parts of the country. It is interesting that previously observed Lyme disease hotspots, such as Thetford Forest [154], were not evident in the current study. This may be due to changing tick population dynamics and/or the prevalence of *B. burgdorferi s.l.* infection in host-seeking vectors, changing human behaviour, or the larger number of patients within the RIPL study population. It is also possible that awareness of Lyme disease is higher in these areas, and cases are successfully identified and managed in primary care without the need for serological diagnosis. Throughout the rest of England and Wales the incidence of confirmed Lyme disease cases remains relatively low (69.2% of resident postcode areas have an incidence of less than 1.0 per 100,000 population per year) compared to the majority of western Europe [66]. These data suggest that although *I.ricinus* ticks are widespread across England and Wales [24], the proportion that carry *B. burgdorferi s.l.* is relatively low, and a higher prevalence may only exist in the tick populations in the localities highlighted. Several studies would appear to support this hypothesis [36,181,182],

but further work is needed to compare the incidence of human cases, abundance of ticks and prevalence of *B. burgdorferi s.l.* in ticks in the same geographic area. The areas with high incidence are predominantly rural and this is reflected in the results where the study population were disproportionately more rural compared to the national population. Information about case locality represented by PHE region is reflective of the case's referring hospital microbiology department rather than the cases' residence, or location of exposure. In some instances, mainly in rural areas, this hospital may be a significant distance from the abode of the patient. This figure therefore is more reflective of the burden of Lyme disease on local microbiology departments.

Information provided at postcode area level relates to the patient's home address, and not necessarily to where the patient was bitten by a tick. Some patients are likely to have been bitten outside their resident postcode area. The further the exposure from home, the larger this spatial error will be. To date, no work has been done to quantify this error in the UK. The smoothed map (Fig. 3.5) attempts to account for this and shows an area of high incidence in southern-central England, centred around Southampton, Salisbury, and Weymouth and extends further west than the raw incidence data. This map highlights theoretical Lyme disease risk areas more accurately, as it accounts for the bite distance spatial error, and should be the map used for targeting public health strategies. The observed strong geographical clustering of positive cases (Fig. 3.6), suggests that patient residence postcode does correlate to some extent with disease risk.

This is the first time that a cohort of laboratory-confirmed Lyme disease cases across England and Wales has been described in terms of the socioeconomic status of their residential postcode area. The results suggest that patients in England diagnosed with Lyme disease are more likely to live in areas which are more affluent, have high levels of employment and education, have a higher quality of life, are less exposed to crime, but have issues with access to housing and local services. This is in contrast to the classic income gradient of health [183–185], where the lower an individual's socioeconomic position the worse their health, but supports previous socioeconomic analyses of Lyme disease in the USA [156,157]. This study has not investigated why areas with higher socioeconomic status appear to correlate with a higher incidence of Lyme disease cases but it may reflect the type of leisure activities undertaken, available leisure time, access and attitudes to the countryside by this section of society [186]. Further research is needed to better define the population of cases diagnosed with Lyme disease and why there is an association with socioeconomic status.

The only negative association with Lyme disease in England was observed for the barriers to housing and services domain and is likely due to the rural nature of the areas with the highest incidence. Rural areas score poorly as the housing tends to be expensive in relation to income and houses are a greater distance from services such as hospitals, schools and post offices. It could be reflective of this population only accessing health care, and so needing serological diagnosis, once symptoms have progressed beyond the early stages of disease. The living environment deprivation domain is a mix of housing quality, air pollution and road traffic accidents, and it is unsurprising that no association with Lyme disease incidence was observed.

In Wales, there was a significant positive correlation between case counts and the WIMD domain scores. There were an increasing number of patients living in more affluent areas. The reasons for this difference are likely to be similar to the English study population.

The main limitation of this study is the use of patient residence postcode area as a proxy both for the place where Lyme disease was acquired and the socioeconomic status of Lyme disease cases. It is unknown how representative the socioeconomic characteristics of a postcode are of individual cases. Clear socioeconomic and demographic trends and associations have been identified; however, these factors cannot be disentangled using the current datasets and so the degree of bias inherent in them is unknown. Future studies should be designed, where a multivariable model can be created to identify any interaction or confounding effects of the variables under examination.

Current guidance for Lyme disease state that an erythema migrans rash is pathognomonic and further laboratory diagnostics are not required [23]. An unknown proportion of cases will be clinically diagnosed and managed in early illness by primary care clinicians and will not make it in to this dataset. Laboratory-confirmed figures will therefore underestimate the true incidence of Lyme disease seen in the general population. Without surveillance of primary care presentations, it will be hard to establish a more accurate incidence figure.

The majority of geographical data presented is reliant on case postcode data. Due to data attrition only 56.6% of cases in the dataset contained this data. Data attrition may have occurred in three ways; poor completion of the laboratory referral forms (something well documented for health professionals [187]), the non-notifiable status of clinical Lyme disease and the lack of statutory obligation to provide information about suspect cases, and the indirect route by which clinical samples are submitted for testing. Lyme disease testing is

usually requested in primary care and samples are routed through hospital laboratories before reaching RIPL. There is the potential that some cases are also missed due to some laboratories (both private and public) performing their own diagnostic testing without sending samples to RIPL, as the reference laboratory, for confirmation. Testing rates may also vary in different geographies dependent upon Lyme disease awareness of health care professionals.

3.6 Conclusions and recommendations

This chapter has shown that laboratory diagnosed Lyme disease cases in England and Wales have a bimodal age distribution and male predisposition. Geographical clustering of cases was seen and new insights into the socio-demographic nature of laboratory-confirmed Lyme disease patients were described. The incidence maps highlight areas where Lyme disease may place the highest burden on primary and secondary care, and therefore can be utilized by medical and public health professionals, and the general public, to raise awareness in these areas.

SGSS has been shown to be flawed as a surveillance tool for Lyme disease. The incidence rates calculated are consistently different to those published by PHE and RIPL, there are large gaps in geographic coverage, it is unknown what test type is used to define a positive result, and inappropriate tissues have been sampled. Due to these problems, it would be recommended that RIPL is used as the primary surveillance tool for laboratory-confirmed cases of Lyme disease in England and Wales.

Nonetheless, there are ways that SGSS could be improved so that meaningful Lyme disease surveillance data could be extracted from it. These include:

- A compulsory audit of all (NHS and private) microbiological laboratories in England, Wales and Northern Ireland, to establish what Lyme diagnostic tests are being performed and what is being reported to SGSS. If any appropriate confirmation tests are being performed, their results could be included in national surveillance figures, something which isn't currently occurring. Another option would be to legislate that all NHS confirmatory diagnostic tests for Lyme disease must be sent to RIPL. Thus, RIPL would report all NHS positive cases, and SGSS would collect all private laboratory positive cases.
- Compulsory inclusion of identification method on the SGSS form. This would rule out any false positives received by the reporting of solely positive screening tests.

- Refresher courses for laboratories on the appropriate tests to use and the Health Protection (Notification) Regulations for appendix 2 organisms [77]. The importance of compliance must be stressed, to ensure complete coverage. This should be primarily targeted at laboratories that are reporting and testing inappropriately.
- Establish whether a sample sent to RIPL and confirmed positive, needs to be reported to SGSS by the submitting laboratory.

RIPL remains the "gold standard" surveillance system for recording laboratory-confirmed cases in England and Wales. There are understandable reasons that explain lower reporting of cases through SGSS. However, it is still likely to be an underestimate of national laboratory-confirmed incidence, as the SGSS data suggests that some laboratories (both NHS and private) are performing their own confirmatory tests. Understandably the incidence reported matches that of national surveillance publications and shows an increasing trend. It is the deeper analytics of patient demographics that offer new insight in to Lyme disease in England and Wales. There is strong clustering of cases, with significant proportion of cases being in rural areas, and most intriguingly the strong association between the level of deprivation and number of cases.

The extent to which this work contributes to the overall aims of this thesis regarding Lyme disease surveillance can be summarised as follows:

- I**ncidence**: Increase ($p<0.001$) from 1.62 per 100,000 in 2013 to 1.95 per 100,000 in 2016. Cases peaked in the summer.
- **Sociodemographics**: Bimodal age distribution, with significantly more men than women. There was a significant socioeconomic trend, with the number of cases decreasing as societal deprivation increased. Cases were significantly more likely to occur in rural areas, compared to the national population.
- **Geographical hotspots**: These were identified in southern England.
- **Patient presentation and management**: No information about patient presentation and management was collected.
- **Additional information:** By comparing RIPL and SGSS, it can be concluded that there is still an underreporting of laboratory-confirmed cases, as not all microbiology units send samples to RIPL, and no private laboratories do.

Chapter 4 Surveillance of Lyme disease in hospital datasets

In Chapter 3 the demographics and geographic distribution of current national laboratory surveillance figures were presented and discussed. These potential trends may be seen in other datasets analysed throughout this thesis. As discussed in Chapter 2, laboratory data represents the traditional peak of a surveillance pyramid (Fig. 2.1). This chapter will discuss the second tier of the pyramid; secondary care. If this data were to follow the hypothetical norm of a surveillance pyramid one would expect a larger population to be represented, with a higher incidence and with demographics more characteristic of the diseased population than the top tier. This chapter will describe the hospital Lyme disease patient cohort, and the quality of the available datasets for analytics reviewed.

The following content is currently in peer review as;

Tulloch, JSP., Decraene, V., Radford, AD., Warner, JC., Christley, RM., Vivancos, R. Characteristics of Lyme borreliosis patients – retrospective analysis of hospital episode data in England and Wales; 1998 – 2015.

4.1 Background

Lyme disease is an important emerging tick-borne disease caused by members of the spirochaetal complex *Borrelia burgdorferi* sensu lato. The population-weighted incidence across Western Europe has been estimated to be 22.04/100,000 person-years [66]. In England and Wales the national incidence of laboratory-confirmed cases has risen from 0.38 per 100,000 population in 1997 [174] to 1.95 per 100,000 population in 2016 [153]. Lyme disease is associated with a range of clinical presentations which may vary as infection progresses, though it commonly presents as erythema migrans with associated flu-like symptoms [23]. Other presentations include: borrelial lymphocytoma, Lyme neuroborreliosis, carditis, arthritis and acrodermatitis chronica atrophicans (ACA) [21]. This has resulted in broad and varied case definitions [23,39,40,43]. However, all case definitions agree that erythema migrans alone, without any laboratory confirmation, is sufficient for case confirmation. Considering this, current surveillance for England and Wales, which is based on laboratory diagnosis [174], is likely to underestimate the true incidence of disease. This resulted in the recent NICE (National Institute for Health and Care Excellence) guidelines explicitly stating that ‘there is a lack of robust epidemiological data on Lyme disease in the UK’ [23].

Patients may present with Lyme disease in either a primary care or hospital setting, with an unknown proportion receiving confirmatory laboratory diagnosis. The relative proportion of patients presenting to either setting is currently unknown, as is the patient pathway between primary, secondary and tertiary care. Within England and Wales, studies that describe patients in a hospital setting have been either limited to one hospital [188,189], specialist referral centres [190–192], or one clinical presentation [179].

Since 1989 Hospital Episode Statistics (HES) have recorded every 'episode' of admitted patient care (APC) delivered in National Health Service (NHS) hospitals in England [87]. Outpatient attendance (OA) and accident and emergency departments (A&E) datasets were added in 2003 and 2007. Patient Episode Database for Wales (PEDW) is a central administrative database that collects admissions data from NHS hospitals in Wales [92,93]. The primary use of these data is the calculation of health care costs and therefore mainly administrative data is collated. There is now an increasing body of medical research using the HES and PEDW databases; nevertheless, a recent systematic review highlighted that only seventeen out of 148 HES publications were related to the epidemiology of a specific disease [88].

The aim of this study was to perform a retrospective analysis of HES and PEDW records to describe the incidence and demographics of Lyme disease patients in a hospital setting, and to describe their patient pathways through the NHS.

4.2 Methods

A retrospective search of both HES (including all datasets of APC, OA and A&E) and PEDW databases was performed to identify patients coded with Lyme disease. A case was defined as a patient with a Lyme disease diagnostic code drawn from the International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) (Table 4.1)[23,39,40,43,86].

Table 4.1 Lyme disease ICD-10 codes used to query hospital administrative data

ICD-10 Code	Description
A69.2	Lyme disease
M01.2	Arthritis in Lyme disease
L90.4	Acrodermatitis chronica atrophicans

A list of variables for each dataset within HES and PEDW was constructed. These variables could be split into three categories; patient demographics, patient geography, and patient management (Table 4.2).

Table 4.2 Variables queried of Lyme disease coded patients in hospital administrative data

Hospital Episode Statistic variable codes	Patient Episode Database for Wales variable codes	Description
HESID	Patient ID	Unique pseudoanonymised patient identifier
ADMIAGE	Admitted Age	Age on day of admission
ADMIDATE	Date first admitted	Date of admission
ADMISOURCE	Admission Method	Source of admission
AEARRIVALMODE		Accident and emergency source
AEATTENDDISP		Accident and emergency discharge destination
APPTAGE		Age on day of appointment
APPTDATE		Appointment date
ARRIVALAGE		Age on arrival to accident and emergency
ARRIVALDATE		Date on arrival to accident and emergency
ATTENDED		Did or did not attend outpatient appointment
DEPDUR		Time spent in accident and emergency until departure
DIAG_CODE		Diagnose code searched in all diagnosis code fields
DISDEST		Discharge destination
EPIDUR		Duration of episode
ETHNOS		Ethnicity
IMD04		Index of Multiple Deprivation
	Deprivation Index	Welsh Index of Multiple Deprivation
LSOA11	LSOA_Code	Lower super output area – 2011 census
REFSOURCE		Source of referral for outpatients
RURURB_IND	Urban Indicator	Rural-urban indicator
SEX	Sex	Sex
TRETSPEF		Main treatment speciality

Data were extracted for patients presenting between 1 January 1998 and 31 December 2015 who had a Lyme disease code in any of the diagnostic fields. Data was cleaned by identifying

missing values and deduplication of records. Date of first appearance of a patient within any of the databases, based on pseudo-anonymised patient identifiers and admission date, was used for analysis. Using these index records, the incidence of Lyme disease coded patients was described for each dataset; mid-year population estimates provided by the Office for National Statistics (ONS) were used as the denominator population data [159]. Annual incidences were analysed using linear regression.

Information on patient sex was stratified by age and compared using a binomial test. Ethnicity was compared to national figures available from the ONS using a Chi-squared test [159]. The average annual incidence was calculated at the geographical area of local authority. The rural-urban indicators of the study populations were compared to the national population using a Chi-squared test.

Associations were assessed using linear regression for the Index of Multiple Deprivation (IMD) of English patients, whereas the Welsh Index of Multiple Deprivation (WIMD) of Welsh patients was assessed using a Chi-squared test for trend. Linear regression could not be performed on WIMD as the defined WIMD groups were of uneven proportion, unlike the IMD which is organised in equally sized deciles. Both were compared to the national populations using a Chi-squared test of independence.

Information relating to patient management primarily was analysed descriptively. To determine if any 'day of the week' bias existed in the data, the number of cases per day was compared to the expected number of cases per day, using a Chi-squared test. This was performed for each dataset and by the admission method recorded in the APC dataset, with the null hypothesis being that there were an equal number of cases every day of the week.

Pseudoanonymised patient identifiers were used to describe the patient pathway. Statistical analyses were carried out using R (version 3.2.0) (R Core Team 2015), and associations were deemed significant where a p value was less than 0.05.

4.3 Results

After de-duplication, 2,361 patients were identified with Lyme disease codes between 1998 and 2015. Within English records (HES) 2,259 unique patients were identified, 2,045 of these were found in APC alone, 180 in outpatients, 13 in A&E, 18 were found in APC and outpatients, and three were found in APC and A&E. Within Welsh records (PEDW), 102 patients were identified. Even though they could not be linked with the HES databases, these were likely to be unique patients, as none of them shared age, sex and lower super output

area (LSOA) of home address combinations with any HES patients. The combined results of both datasets were therefore described, unless otherwise specified.

The annual incidence of Lyme disease coded patients rose significantly from 0.08 cases per 100,000 population in 1998 to 0.53 in 2015 (r^2=0.93, p<0.01; Fig. 4.1).

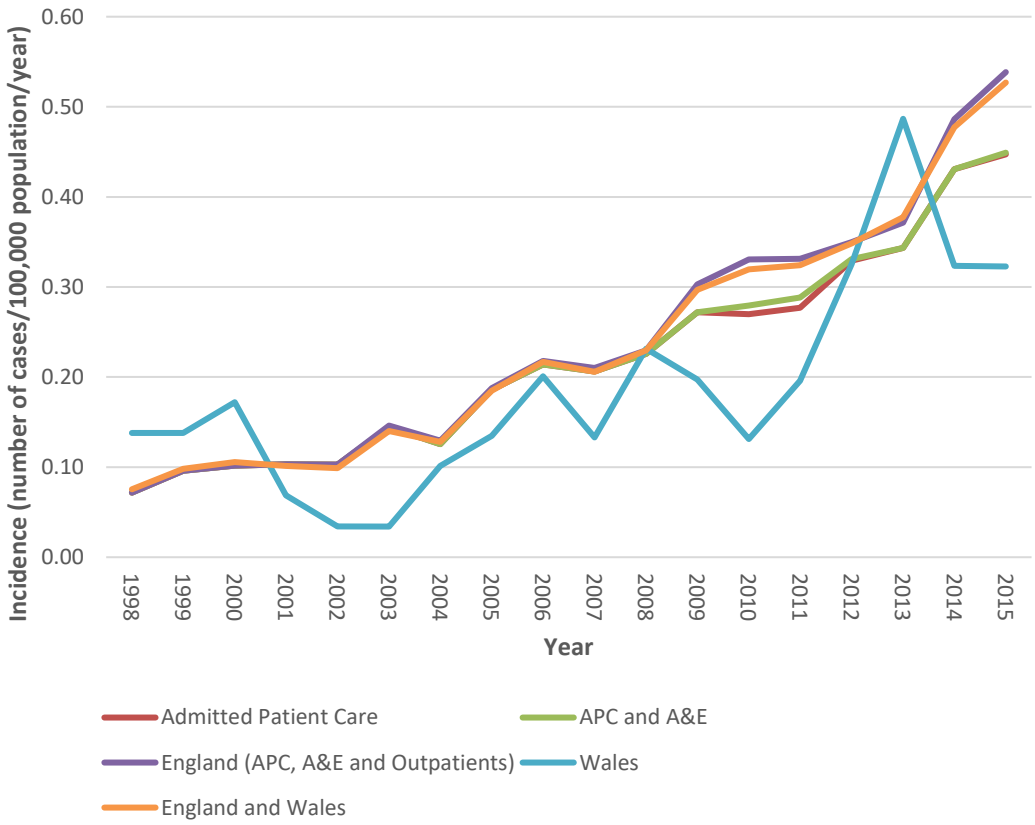


Figure 4.1 Incidence of Lyme disease coded patients within hospital administrative records in England and Wales (1998-2015)

This significant correlation was seen both in English (r^2=0.93, p<0.01) and Welsh (r^2=0.55, p<0.01) populations. There was marked seasonality, with peak number of cases recorded in August (Fig. 4.2).

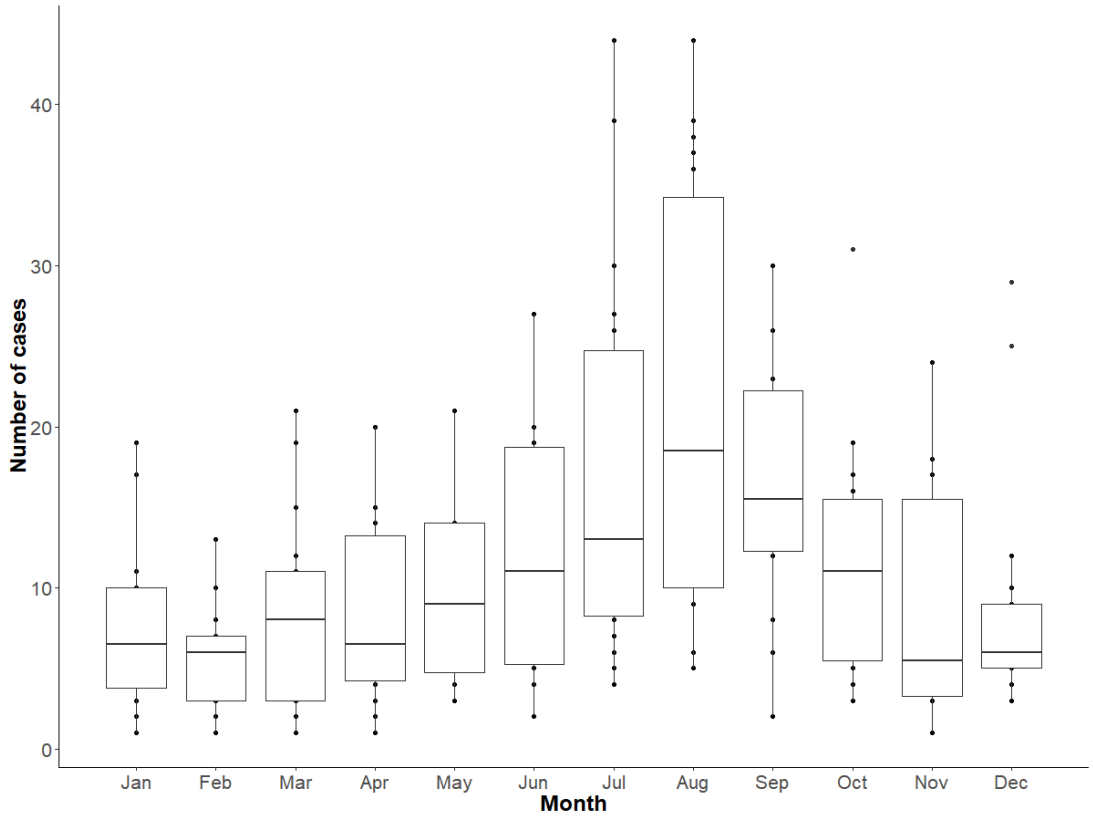


Figure 4.2 Lyme disease patient monthly count, within hospital administrative records in England and Wales (1998-2015)

4.31 Demographic characteristics

In England and Wales, 70.9% (n=1,673) of records contained information on the patients' age and sex. When stratified by country, English records contained 69.5% (n=1,571) of this information, and Welsh 100% (n=102). There were significantly ($p<0.01$) more female patients than male in England and Wales 60.1% (n=1,005), displaying a bimodal age distribution, with peaks at 6-10 and 61-65 year age bands (Fig. 4.3).

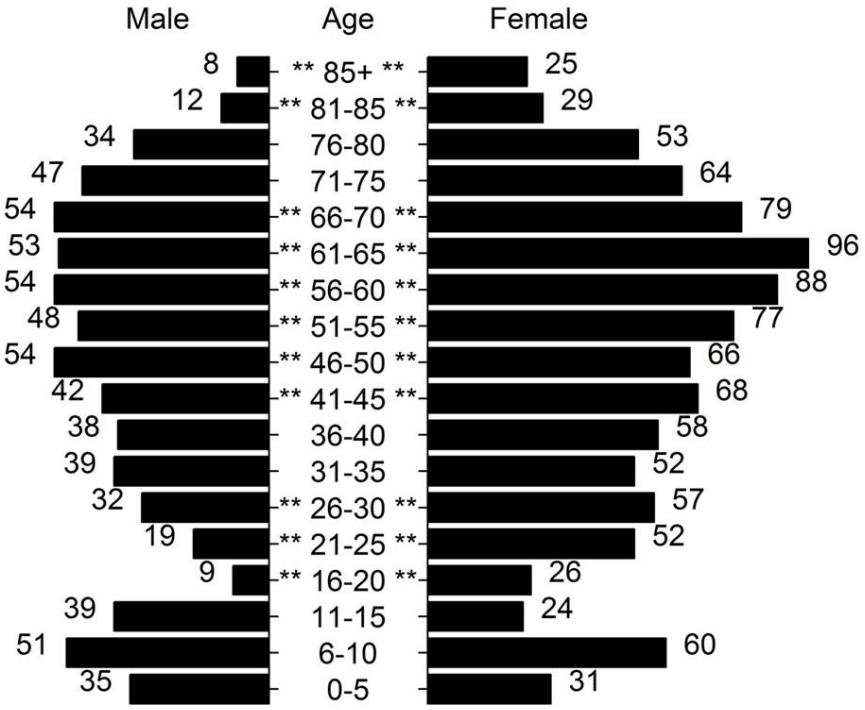


Figure 4.3 Population demographics of Lyme disease patients within hospital administrative records in England and Wales (1998-2015). Asterisks represent a significant difference (p<0.05) between sexes.

This sex ratio held true in England (60.5%, p<0 .01), in Wales there were more female patients 52.9% (n=54), however this was not significant (p=0.62). Ethnicity information was available for 79.5% (n=1,877) of records in England and Wales. Of these records, 96.1% (n=1,803) of patients were recorded as identifying with being white. Using a Chi-squared test to assess white ethnicity vs other ethnicities, a significantly (p=0.01) greater proportion of this population was white compared to the 2011 Census population [193].

4.32 Geographical distribution

Over two thousand (2,078; 88.0%) records contained geographical information (Fig. 4.4).

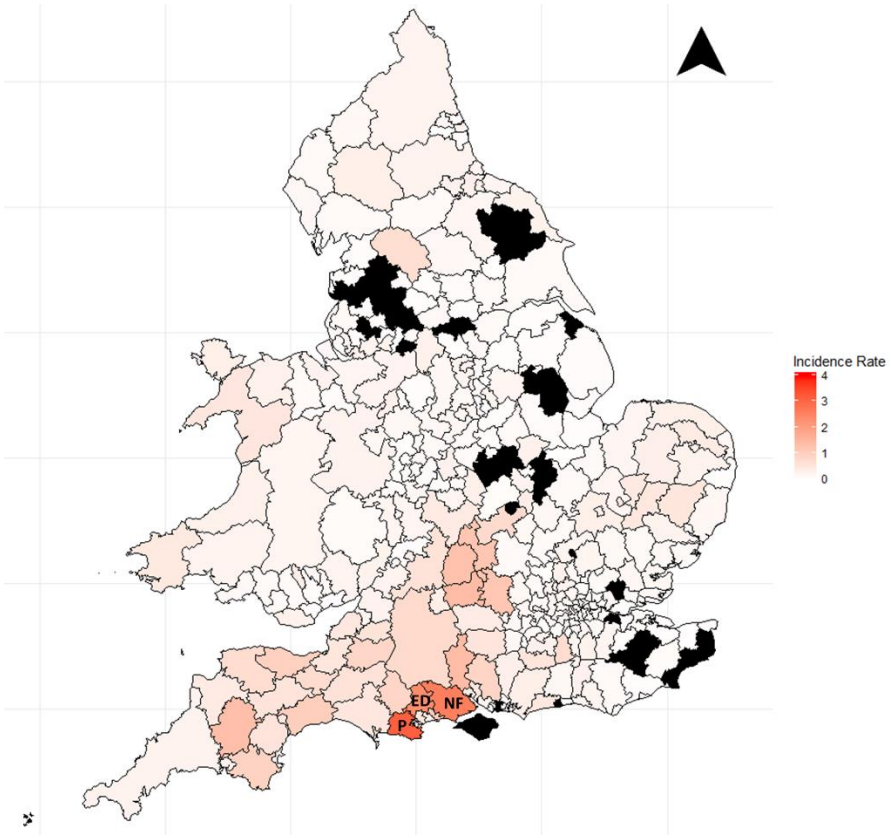


Figure 4.4 The average incidence rate of Lyme disease in English and Welsh local authorities (n=348), (1998-2015). These data were based on hospital administrative records, and incidence measured as number of cases per 100,000 per year. Black areas recorded no cases over the study period. P = Purbeck, ED = East Dorset, NF = New Forest.

The areas with the highest incidence were located in the south west of England. The local authorities with the highest incidence were Purbeck with 3.13 cases per 100,000 per year, New Forest (2.58), and East Dorset (2.32), with the incidence in neighbouring areas in central southern England also with high rates. Thirty-four (9.8%) local authorities recorded no hospital cases assessed for Lyme disease.

Analysis of rural-urban indicators showed a significant difference between the study population (n = 2,292) and the national population, where Lyme disease patients were more likely to live in rural (37.4%) rather than urban areas, compared to the national population (17.9% live in rural areas) ($p<0.01$).

4.33 Sociodemographic characteristics

Information on IMD deciles was available for 96.7% of English patients (n=2,186). There was a significant difference ($p<0.01$) between this population and the national English population, with a significant linear trend showing that patients were found in increasing numbers in less deprived areas ($r^2=0.87$, $p<0.01$). Information on WIMD was available for 90.1% (n=92) of Welsh patients; using Chi-squared tests, there was a significant difference

(p<0.01) between this population and the national Welsh population, and there was a significant linear trend, with increasing number of patients found in the least deprived areas (p<0.01).

4.34 Patient Management

There were significant differences between the daily cases in APC (p<0.01), OA (p<0.01), and Welsh admissions (p=0.01), compared to the expected number of cases per day of the week. For these three datasets, there were fewer cases at the weekend, and the APC dataset had a high number of cases on a Monday (Fig. 4.5).

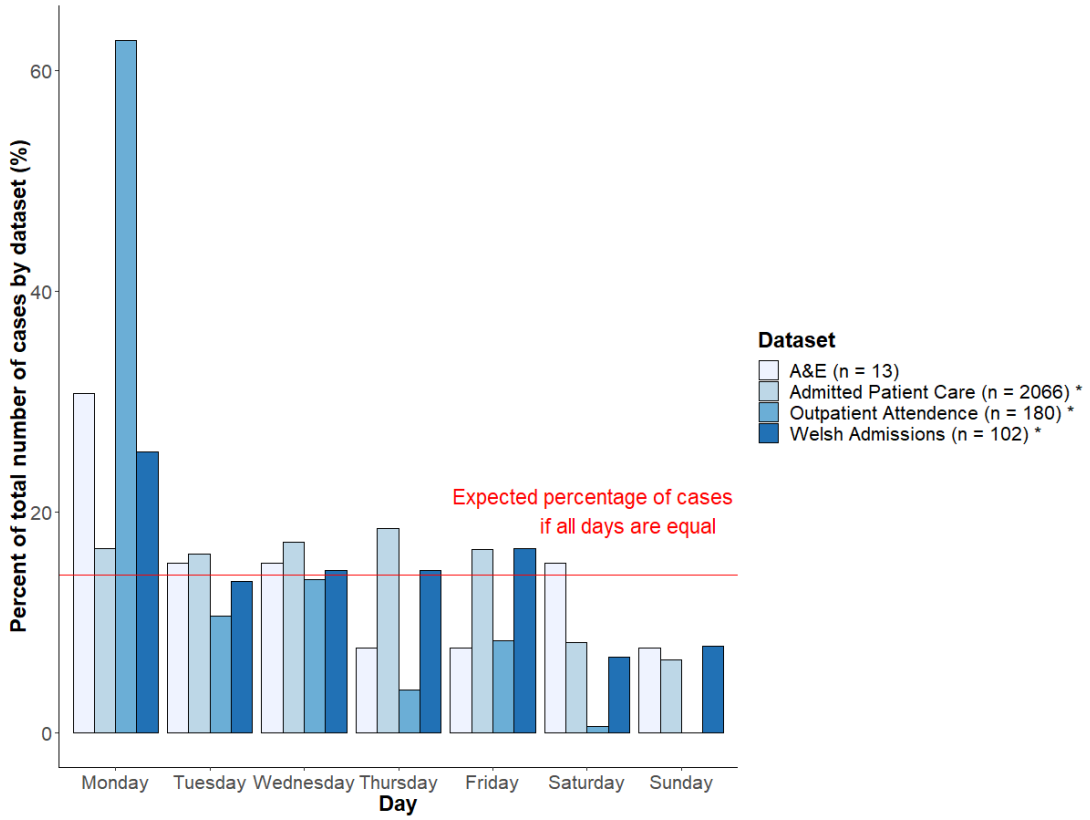


Figure 4.5 Proportional daily Lyme disease case attendance, in English and Welsh hospital administrative records (1998-2015). Asterisks represent a significant difference (p<0.05) compared to the expected proportion of daily cases.

There was no significant difference between daily case numbers for the A&E dataset (p=0.72). Within the APC dataset, there were significant differences between the daily cases admitted via the elective (p<0.01), GP (p<0.01), and other (p<0.01) routes, compared to the expected number of cases per day of the week. There were fewer cases admitted via these routes at the weekend. There was no significant difference between daily case numbers for patients admitted through A&E (p=0.67) (Fig. 4.6).

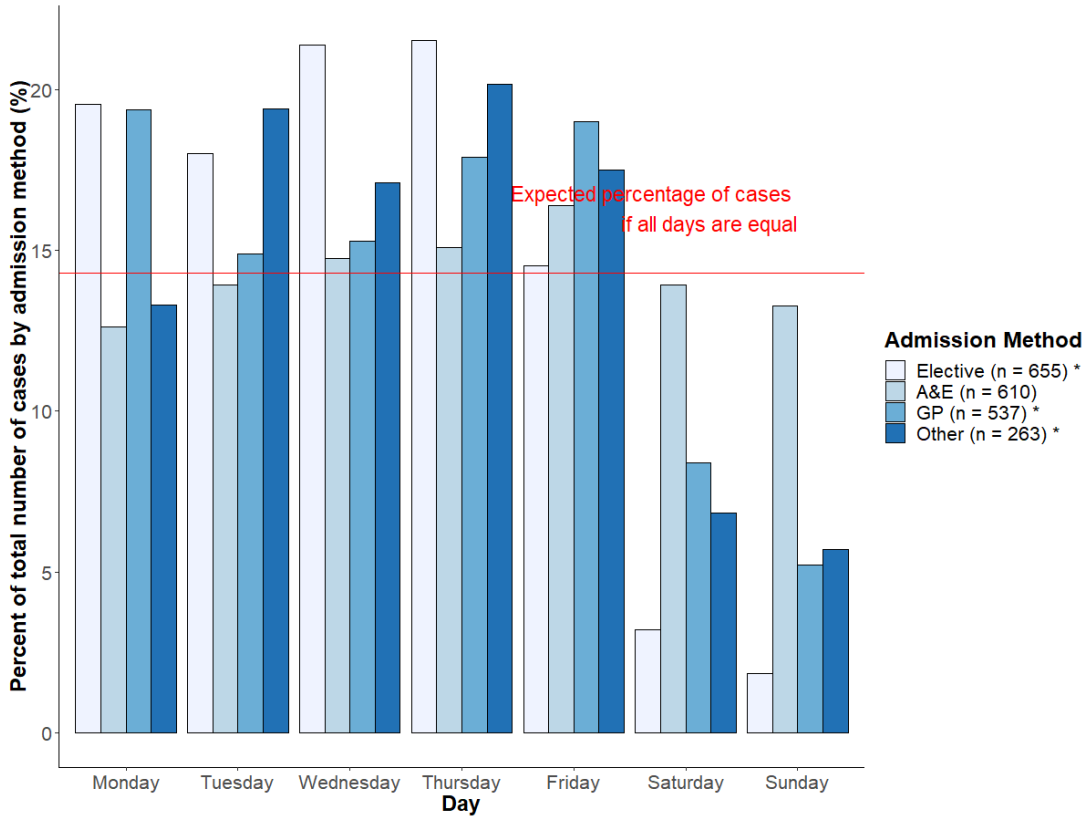


Figure 4.6 Proportional daily Lyme disease case admission routes, in English and Welsh hospitals (1998-2015). Asterisks represent a significant difference ($p<0.05$) compared to the expected proportion of daily cases

In table 4.3, coding patterns, department of treatment, bed days, number of appointments, and length of time in A&E for HES data, are shown.

Table 4.3 Patient management statistics for Lyme disease coded patients in Hospital Episode Statistics (1998-2015)

	Admitted Patient Care (APC)	**Outpatients**	**Accident and Emergency (A&E)**
ICD-10 Codes			
Lyme disease	91.5% (n=1,891)	27.8% (n=55)	100% (n=16)
Acrodermatitis chronica atrophicans	8.0% (n=166)	71.4% (n=142)	0
Lyme Arthritis	0.1% (n=2)	0.5% (n=1)	0
Lyme and LA	0.3% (n=7)	0	0
Number of Departments of Treatment Recorded	63 (2,065 patients)	20 (198 Patients)	N/A
Top 5 Departments of Treatment	General medicine 28.9% Paediatrics 14.7% Neurology 10.8% Gynaecology 4.8% Infectious disease 4.5%	Dermatology 70.7% Rheumatology 5.6% Neurology 5.1% Infectious disease 4.0% General medicine 3.0%	N/A
Mean number of episodes per patient	1.72 episodes (range: 1-50)	N/A	N/A
Mean number of bed days with patients with one episode (n=1,638, 79.3% of APC patients)	4.47 days (range: 0-137) 733 (35.5%) with one episode and no bed days. 258 (12.5%) with one episode and one bed day.	N/A	N/A
Mean number of total bed days for patients with more than one episode (n=427, 20.7% of APC patients)	11.2 days (range: 0-315)	N/A	N/A
Mean number of outpatient appointments (n=308, 24 cancelled)	N/A	1.5 (range: 1-25)	NA
Mean time in A&E (minutes)	N/A	N/A	140 (32-237)

Lyme disease was the predominant code in admissions (91.5%) and A&E (100%) data, where as it was ACA (71.4%) in outpatients. Data on patient management for PEDW data was limited to patient admission method; 67.6% (n=69) of Welsh patients were admitted through the A&E department, the remainder were electively admitted.

Patient pathways were described using the source of the patient and their discharge method (Fig. 4.7).

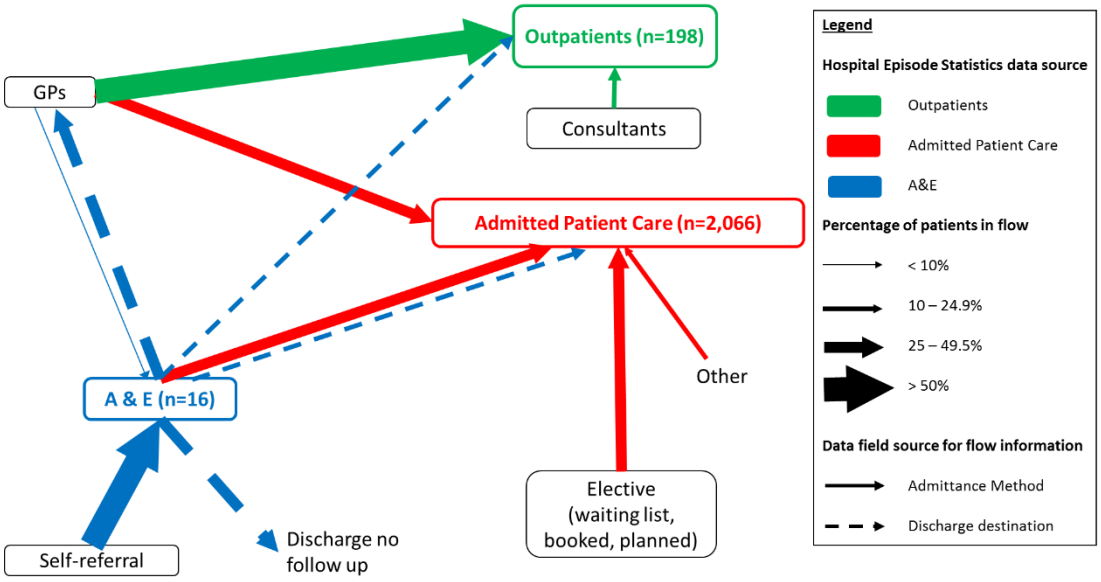


Figure 4.7 The pathway of Lyme disease coded patients through the NHS, based on hospital administrative records

There was no discharge information for OA, and information for APC was excluded as discharge destination codes did not explicitly describe whether patients were to receive primary care on discharge or whether patients were referred to an outpatient or inpatient clinic.

4.4 Discussion

This study provides an in-depth description of Lyme disease patients seen in English and Welsh hospitals, and addresses some of the NICE guidelines calls for new epidemiological data [23]. Incidence rose over the study period, showing a similar trend, but at lower levels, to officially published figures based on laboratory-confirmed cases [153,174]. This discrepancy is to be expected, as national laboratories will receive samples from both hospital and primary care patients, and will therefore have a higher incidence. Not all cases would need to be referred to a hospital clinician from primary care, as the majority of cases are likely to present with an uncomplicated erythema migrans rash [23]. The cause for the increase in incidence is unknown, but may be the result of, among other causes: increased awareness by the public and/or hospital clinicians, increase in referrals by clinicians in primary care, or a true increase in incidence within England and Wales. Further research is needed to understand the drivers for this increase in incidence. Compared to other European countries the incidence described is lower. In France the annual hospitalisation rate due to

Lyme diseases is 1.55 cases per 100,000 [180], with an estimated average national incidence of 42 cases per 100,000 population. Whereas in Germany the inpatient incidence was 9 cases per 100,000 population, but with large regional variation [194]. The reasons for this are mixed, and are likely due to: differences in *Ixodes spp* prevalence and *Borrelia spp* carriage rates, different levels of exposure to ticks by the general population, and differences in how patients access healthcare.

The seasonality observed here supports the known risk factors and epidemiology of Lyme disease. Tick populations in the UK have been shown to peak in June or July each year [24,28,142]. One would therefore expect to see tick bite incidence and exposure to Lyme disease to peak similarly. Clinical signs will appear anywhere from several days to a few weeks after a tick bite [40]. Previous work in England and Wales showed a peak of serologically confirmed cases in August and September, with an assumed peak of symptoms earlier in the summer [154]. This work would support this conclusion. This mirrors other Northern European countries, such as Finland and Germany, where clinically diagnosed cases peak throughout July and August [67,176].

The age structure of this population compares closely with a recent study performed in England and Wales [179]. It shows the classic bimodal age distribution seen with Lyme disease, with an initial peak incidence in pre and peri-pubescent children, followed by a second larger peak from late middle age. The reasons for this age structure haven't been formally assessed, however there is agreement that it likely reflects an increased exposure to tick habitats due to leisure behaviour rather than occupational exposure [67]. These data display a predominance of female cases, unlike the study referenced above. The reasons for this are hard to explain, but could be related to differences in health seeking behaviour [109].

Ninety-six percent of patients identified as being white, compared to 86% in the 2011 national census [159]. There is no clear reason why ethnicity has any impact on a person's susceptibility to Lyme disease. Instead, this apparent association is most likely due to sociocultural and behavioural reasons. Patients were found, in increasing numbers, living in less deprived areas. It must be noted that all ethnic minority groups were more likely to live in areas of higher deprivation than the white population [195], and this could explain the higher proportion of white patients within this population. Lyme disease patients were more likely than the national population to live in rural areas. The characterisation of Lyme disease patients as white and from suburban or rural areas with low deprivation may be explained by a complex combination of risk factors related to access to habitats which support ticks

(either through work or recreation), and access to health care [196]. Further research is required to understand the link between ethnicity, deprivation, area of residence and presentation to hospitals with Lyme disease.

There is clear geographical variation in incidence between local authorities. The highest incidence is in southern-central and western England, which has traditionally been seen as a Lyme disease hotspot [154]. Areas with no cases are unlikely to be due to an absence of disease but may reflect differences in case management or hospital coding practices. The remainder of England and Wales is a patchwork of low incidence with no obvious hotspots of disease. Interestingly, there are no clear foci of infection observed in either the Thetford Forest, the Lake District or the North Yorkshire Moors as identified previously by Public Health England (PHE) [174]. In these areas the awareness, diagnosis and management of Lyme disease may differ from other areas, perhaps with primary care clinicians treating cases in the community and with fewer subsequent cases referred to hospitals. The geographical data collected by HES and PEDW is based upon the patient's home address and no information is recorded on recent travel history or where a tick bite may have occurred, and so there may be an element of bias in the results.

Bed day analysis showed three distinct populations; those with one episode who weren't admitted (35.5% of patients) or stayed for one night (12.5%), those with multiple episodes and a low number of bed days and those with one or many episodes that had a large number of bed days (Table 4.3). The first group is likely to represent patients with uncomplicated cases of Lyme disease. The second group often had consecutive daily episodes totaling fourteen to twenty-one days, which could be consistent with daily intravenous doses of antibiotics as recommend by the British Infection Association and National Institute for Health and Care Excellence (NICE) guidelines [23,39]. The final group appear to represent complicated cases of Lyme disease that require prolonged stays in hospital. It was not in the scope of this project to see whether any clinical presentations predisposed patients to these three groups, but further investigations are recommended.

Analysing the patient flow through the datasets has enabled better understanding of the care pathway for Lyme disease infected patients. Thirty percent of Lyme disease admissions in England, and 67.6% in Wales, originate from the A&E department. To place this into context, in 2011 69% of all NHS England admissions originate from A&E [197]. The same report saw a decline in admissions through primary care referral and an increase through A&E between 2001 and 2011. It would be unlikely that the numbers of patients admitted in

this study have more acute/severe presentations of disease that require immediate hospital attendance, however this cannot be ruled out. A combination of two factors possibly result in this finding; the lack of knowledge of the recommended care pathways for symptoms associated with Lyme disease (such as flu-like illness and rashes), and the difficulty in getting a prompt appointment in primary care [190,197–200]. Peak non-urgent attendance at NHS emergency departments has been recorded at weekends [201], which may be due to the lack of access to primary care at the weekend [190,198–200,202]. However, the data shows that the number of cases appearing in A&E is relatively evenly distributed throughout week, suggesting that the lack of knowledge of where to seek help with Lyme disease symptoms may be the predominant cause of the above findings. Further work is needed to explore why so many patients would seek treatment at a hospital when, for the majority of cases, management could occur at primary care level. By linking with primary care electronic health records, one may be able to see whether they had sought help first in primary care before arriving at A&E.

The major limitations of this study revolve around the use and validity of ICD-10 codes. A case of Lyme disease can be defined without laboratory confirmation, so there is no way to independently validate the accuracy of diagnostic coding in this context [23,39]. Previous work has shown that coding practices in hospitals are not infallible, but are steadily improving; quality issues were primarily focused on patient management variables, rather than demographics and geography [203]. Without such an audit, any potential inconsistencies in coding behaviour cannot be fully understood or quantified. Subjectively, admissions data in HES and PEDW were the most robust. As such, further work on the Lyme disease patient hospital population should primarily focus on admissions data.

Sixty-three treatment departments were recorded, some of which have no discernible link to Lyme disease. This may represent simple coding errors or that the code has been added for completeness when the primary reason for admission was unrelated to Lyme disease. The outpatient dataset was significantly overrepresented by two hospitals; both had the main treating department as dermatology and resulted in a high number of ACA codes. This is further seen by the large number of outpatients seen on a Monday. These cases were all from one hospital, and likely represent one dermatology clinician's outpatient clinic. This suggests that outpatient departments across England and Wales were not coding consistently and episodes may be being lost. The A&E dataset contained very low numbers of patients, in stark contrast to the large number being admitted through A&E as recorded

in the APC dataset. The main reasons for these low numbers is not through lack of attendance but how coding is encouraged. Within A&E, coding is not required to be as specific as the admissions data, and is just needed to code a generalised condition, sub-analysis of more serious conditions and anatomical area involved [204]. This results in Lyme disease potentially falling into multiple categories depending on symptoms, such as "Infectious disease", "Local infection", "Dermatological conditions" and "Facio-maxillary conditions". This has been seen in previous work on arthropod bites, where all cases were recorded as "Bites/Stings" and routinely didn't specify the causal arthropod [205].

PEDW only collects admission data and so some of the issues discussed above for the English dataset were negated. Unfortunately, linkage between the PEDW and HES datasets was not possible; though, for reasons described above, these patients were likely to be unique. Without linkage there still is the potential of duplication of patients within the records and therefore there is a small degree of uncertainty attached to these results.

4.5 Conclusions

This chapter has described the demographics of hospital patients who are coded with Lyme disease, across England and Wales. The demography of this population poses some interesting questions, especially around female predominance, the relative lack of ethnic diversity and the trend towards habitation in areas of low deprivation. This chapter provides a platform to inform future work on Lyme disease patients within hospital settings. Analysis of secondary care data can inform and help target health promotion messages, and as this is an ongoing dataset, interventions relating to Lyme disease could be formally assessed.

The extent to which this work contributes to the overall aims of this thesis regarding Lyme disease surveillance can be summarised as follows:

- I**ncidence**: Increase ($p<0.001$) from 0.08 per 100,000 in 1998 to 0.53 per 100,000 in 2015. Cases peaked in the summer.
- **Sociodemographics**: Bimodal age distribution, with significantly more women than men. They were more likely to identify with being white than the national population There was a significant socioeconomic trend, with the number of cases decreasing as societal deprivation increased. Cases presenting in secondary care were significantly more likely to occur in rural areas, compared to the national population.
- **Geographical hotspots**: These were identified in southern England, particularly in the south-west.

- **Patient presentation and management**: No information about patient presentation was collected. Cases tended to be admitted during weekdays. A single admission averaged 4.5 days in stay length. Admissions to hospitals are evenly distributed between accident and emergency departments, GP referrals, and elective lists.
- **Additional information:** The APC dataset provided the most robust resource to use for analysis. The A&E, and outpatients data sets had significant flaws. A recommendation from this work would be for future research to focus solely on the APC dataset.

Chapter 5 Surveillance of Lyme disease in a primary care dataset

The results from the hospital datasets (Chapter 4) are suggestive that the theoretical surveillance pyramid model may not accurately reflect the Lyme disease population, as the hospital dataset represents a smaller population of patients than that found in laboratory surveillance data (Chapter 3). The implications of this will be discussed more thoroughly in Chapters 9 and 10. The next tier in the surveillance pyramid (Fig. 2.1) represents Lyme disease patients that present to primary care. This chapter will discuss the analysis of a primary care electronic health records datasets to describe the incidence and sociodemographics of patients coded with a Lyme disease related Read code.

The following content is currently in peer review as;

Tulloch, JSP., Christley, RM., Radford, AD., Warner, JC., Beadsworth, MBJ., Beeching, NJ., Vivancos, R. The incidence of Lyme disease cases in a UK primary care cohort, 1998-2016.

5.1 Introduction

Lyme disease, caused by some members of the spirochaetal genospecies complex *Borrelia burgdorferi* sensu lato, has been the topic of much debate and created many headlines in the United Kingdom (UK) [12,54,206]. It is transmitted by the bite of an infected *Ixodes* spp of tick, and is the most common zoonotic disease transmitted by ticks in the Northern Hemisphere [21]. It has a variety of clinical presentations, most usually including erythema migrans, flu-like symptoms, and joint and muscle pain, or more uncommonly neurological and cardiac presentations [18,21,23,152]. Current recommendations are to treat patients presenting with an erythema migrans rash with antibiotics. Laboratory diagnostic tests are recommended when erythema migrans is absent and if there is clinical suspicion and a strong supportive history of Lyme disease [23]. However, as the (National Institute for Health and Care Excellence) NICE guidelines state, 'there is a lack of robust epidemiological data on Lyme disease in the UK' [23]. This lack of knowledge includes incidence data in different health care settings, basic patient demographic information, and an understanding of current case management strategies by health care professionals.

As notification of clinical cases is not required, national incidence figures in the UK are based on reports of laboratory confirmed cases from the reference laboratories of Public Health England and Health Protection Scotland [82,153]. In 2016, the national incidence reported was 1.95 cases per 100 000 population in England and Wales, and 3.15 cases per 100 000 in

Scotland. Over the last decade, cases in England appear to be rising, whilst the incidence in Scotland is reported to be stable [76,82,153]. A 2016 review compared reported incidence across Western Europe and calculated a population-weighted average incidence rate of 22.05 cases per 100 000 person-years [66]. In the United States of America, a study of the incidence of clinician-diagnosed Lyme disease calculated an annual incidence of 106.6 cases per 100 000 persons [64]. These differences in incidence are likely due to a combination of differing surveillance methods and differences in true incidence. Without a comprehensive surveillance system and an internationally standardised case definition, comparisons between nations prove challenging.

Within a health care system, primary care manages the greatest number of Lyme disease patients [21,23,64,176,177,180,207]. No work has examined UK Lyme disease patients that accesses primary care. Without understanding the potential burden for general practitioners (GPs) and the demographics of these patients, it is difficult to shape policy, deliver targeted education to the general public and clinicians, perform financial assessments, or to understand case management strategies. The incidence of Lyme disease identified within primary care in the UK remains unknown, and no population-based UK study has been previously performed. There are two methods of recording primary care data; Read codes representing presenting symptoms or diseases, and free-text narrative. Read codes are a coded thesaurus of clinical terms that are used in primary care electronic health records in the UK and New Zealand [101]. A narrative analysis of health record free text, on a national scale, would prove ethically challenging due to difficulties in data anonymisation. On the contrary, primary care databases coded via Read codes are pseudo-anonymized and capture a large sample of the UK population. The aim of this ecological study was to describe the incidence and demographics of Lyme disease as recorded in primary care between 1998 and 2016 in the UK using Read code analysis.

5.2 Methods

5.2.1 Data Source

Population-based primary care data from The Health Improvement Network (THIN) were used to identify patients with Lyme disease, suspected Lyme disease or Lyme disease related conditions. The design of this study was approved by the THIN Scientific Review Committee (16THIN103).

THIN represents 11.1 million patients with around 4.0 million active patients, collected from over 700 general practices. An active patient is defined as one being registered to a general

practice currently supplying data to THIN, who is not dead and has not left the practice since the last data collection point. THIN has representative coverage of 6.1% of the UK population, and is representative in terms of demographics, major condition prevalence and adjusted death rates [118]. All patients and general practices are pseudoanonymised and demographic information is available at patient level for: age, sex, ethnicity, and nation of residence. The representativeness of ethnicity data within THIN has been questioned, as the level of missingness at case-level is high. Between 2000 and 2013, 60% of THIN patient records had missing ethnicity information [208]. Ethnicity data are based upon patient-provided information categorised into the following 2011 census groups; 'White', 'Mixed', 'Asian', 'Black', and 'Other' [193,209].

The remaining sociodemographic variables under assessment were Townsend scores (an indicator of material deprivation) [210], and rural urban classification. Within THIN these data are not related directly to the case but are based upon the case's resident postcode, and then linked to 2001 census data [159]. These data are therefore not a direct measure of the case's sociodemographics, but rather a proxy, and reflect better the area that cases reside in. Townsend scores were converted, by THIN, from exact scores in to quintiles of equal size. The quintile of 1 includes patients living in the lowest 20% of Townsend scores (i.e. the least deprived areas), whereas the quintile of 5 includes the highest 20% and the most deprived areas.

5.2.2 Participants and statistical analysis of the data

In primary care the presenting symptoms of a patient are coded with Read codes. Currently we do not know which set of symptoms clinicians identify to code the patient as 'Lyme disease'. Our case definition was therefore restricted to Read codes specific solely to Lyme disease and suspect Lyme disease (Table 5.1). The 'Suspected Lyme disease' and 'Suspected erythema migrans' codes were only introduced as Read codes in 2014 [211]. Conditions with multiple aetiology, such as Bell's palsy, were not included. This strict definition was chosen to minimise the number of false positives identified. Choosing strict case definitions will likely underestimate the number of cases and sensitivity may be lost, as cases of mixed non-specific clinical signs could be missed. These codes were used to identify patients accessing primary care between 1st January 1998 and 31st December 2016. No other exclusions were placed on the patients. The index episode was taken as the first occurrence of any one of the Read codes identified in a patient's record. All calculations and demographic information were derived from this date.

Descriptive statistics were used to summarise the epidemiology of Lyme disease by demographic factors. Overall incidence rates, and incidence rates and ratios by sex, country, and Read code per 100 000 person-years were calculated. Incidence rates were constructed using patient-level data from THIN. This was defined as the number of cases divided by the sum person-time contributed by all patients in THIN between 1998 and 2016.

Age-standardised incidence rates per year were calculated for the whole dataset, and were stratified by, sex, country and Read code. Denominators were calculated as the mid-year population of the THIN database. These annual incidence rates were assessed for trend using linear regression. Crude monthly incidence rates per year were calculated for the whole dataset and stratified by nation.

Due to the poor recording of ethnicity within THIN, the complete electronic health record of each identified case was read to visually confirm ethnicity status, rather than constructing a Read code search. It was not in the remit of this project, or feasible, to perform this for the remainder of the THIN population. As such, incidence could not be calculated, and proportions of case ethnicity classification were compared with the national population. Ethnicity was compared to the national population using a Chi-squared goodness of fit test.

The crude incidence rates of rural and urban place of residence were calculated. An incidence rate ratio was calculated to compare whether statistically significant differences in incidence existed between urban and rural areas. The crude incidence rate of each Townsend quintile was calculated. A Chi-squared test for trend and a Chi-squared test for departure from the trend were used to analyse trends in the incidence of cases across the Townsend quintiles.

The study was designed as an ecological study and it was felt that the creation of a multivariable model, based on demographic variables, would not be statistically appropriate. This was due to the inability to calculate incidence for ethnicity, and that the incidence for rural urban classification and Townsend quintiles were based upon proxy geographical units rather than on individual case data, thus leading to issues of ecological fallacy. These variables were therefore assessed descriptively.

All statistical analyses were carried out using R language (version 3.2.0) (R Core Team 2015), and results were deemed significant where $p<0.05$.

5.3 Results

In total 3,725 unique patients were identified with a Read code for Lyme, suspected Lyme disease, or related conditions Read code between 1st January 1998 and 31st December 2016 (Table 5.1).

Table 5.1 Read codes used to identify Lyme disease patients in The Health Improvement Network (THIN) dataset (1998 - 2016)

Description	THIN Read Code	Number of patients
Lyme disease	A871000	2,386
Erythema migrans	AA41.00	992
Suspected Lyme disease	1JN1.00	233
Suspected erythema migrans	1JN2.00	50
Acrodermatitis atrophicans chronica	M21y000	30
Lyme arthritis	N010A00	21
Lyme neuroborreliosis	A871100	8
Borrelial lymphocytoma	A871300	5
Lyme carditis	A871200	0
Total		3,725

The most frequently used Read codes ('Lyme disease' and 'Erythema migrans') represented 89.1% (n=3,318) of all Read codes identified with Lyme disease. The suspected Lyme disease codes only represented 7.6% (n=283) of all codes.

The overall incidence rate was 4.42 per 100,000 person-years (95% CI 4.23-4.63). Incidence rates and incidence rate ratios between sexes, countries and Read codes for patients with Lyme disease in the UK are presented in Table 5.2.

Table 5.2 The incidence of Lyme disease cases by demographic factors and Read codes in The Health Improvement Network (THIN); 1998 – 2016

Demographic Category	n	Person-years	Incidence rate (per 100 000 person-years) (95% CI)	Demographic Comparison	Incidence rate ratios (95% CI)	p value
Sex						
Female	1,907	43,109,106	4.42 (4.23-4.63)	Female versus male	1.04 (0.97-1.10)	0.29
Male	1,818	42,534,510	4.27 (4.08-4.47)			
Country						
England	2,055	60,016,197	3.42 (3.29-3.56)	England vs NI	2.59 (2.00-3.43)	<0.01
Northern Ireland (NI)	50	3,781,630	1.32 (0.99-1.74)	England vs Scotland	0.32 (0.30-0.34)	<0.01
Scotland	1,455	13,508,169	10.77 (10.23-11.34)	England vs Wales	1.73 (1.48-2.03)	<0.01
Wales	165	8,337,619	1.98 (1.69-2.23)			
Read Code						
Lyme disease	2,386	85,643,615	2.79 (2.68-2.90)			
Erythema migrans	992	85,643,615	1.16 (1.09-1.23)			
Acrodermatitis atrophicans chronica	30	85,643,615	0.04 (0.02-0.05)			
Lyme arthritis	21	85,643,615	0.02 (0.02-0.04)			
Lyme neuroborreliosis	8	85,643,615	0.009 (0.004-0.02)			
Borrelial lymphocytoma	5	85,643,615	0.006 (0.002-0.01)			
All Lyme codes	3,442	85,643,615	4.02 (3.89-4.12)			
Suspected Lyme disease	233	85,643,615	0.27 (0.24-0.31)			
Suspected erythema migrans	50	85,643,615	0.06 (0.04-0.08)			
All suspected Lyme codes	283	85,643,615	0.33 (0.30-0.37)	All Lyme codes vs all suspected Lyme codes	12.16 (10.77-13.73)	<0.01
Rural Urban Status – Only available for English and Welsh cases						
Rural	623	11,129,541	5.60 (5.17-6.05)	Rural vs Urban	1.92 (1.75-2.11)	<0.01
Urban	1,450	49,793,661	2.91 (2.77-3.07)			
No status recorded (including all of Scotland and Northern Ireland)	1,652	N/A				
Townsend Quintiles – a measure of material deprivation						
1 (least deprived)	916	18,632,692	4.92 (4.61-5.24)			
2	1,081	16,365,633	6.61 (6.22-7.01)			
3	787	16,076,059	4.90 (4.56-5.25)			
4	452	14,223,891	3.18 (2.90-3.48)			
5 (most deprived)	221	9,986,715	2.22 (1.94-2.52)			
No score recorded	287	N/A				

There was an increase in incidence for both sexes across the study period, however there was no significant difference in incidence rate between sexes. There was a significant difference between countries, with Scotland having the highest incidence of 10.77 per 100 000 person-years (95% CI 10.23-11.34). This was more than three times higher than in England; the country with the next highest incidence of 3.42 per 100 000 person-years (95% CI 3.29-3.56). There was a significant increase in the age-standardised incidence of Lyme disease coded patients in primary care in the UK between 1998 and 2016 (Fig. 5.1-3). This rise was seen in all nations except Wales.

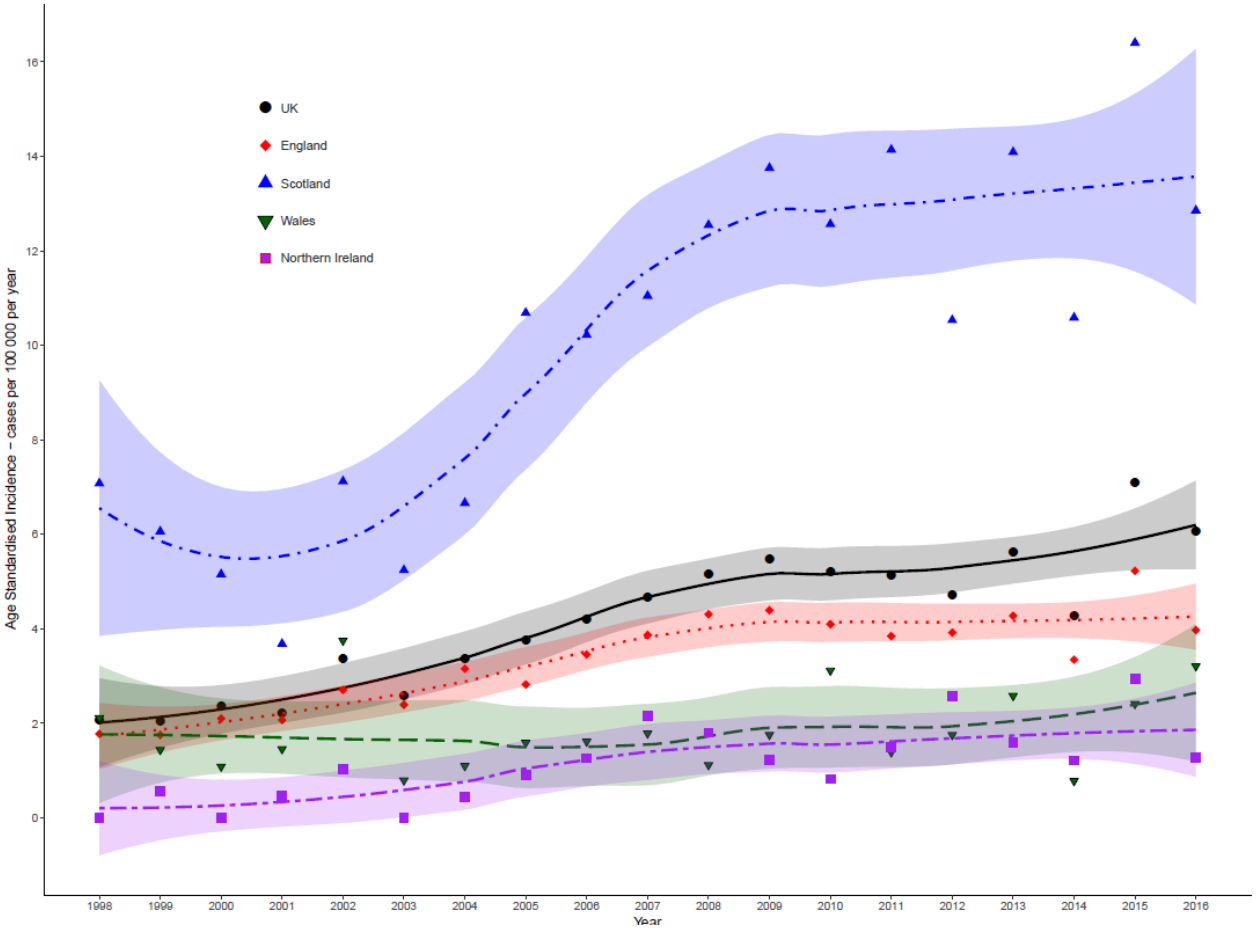


Figure 5.1Age-standardised incidence rates of Lyme disease coded patients in The Health Improvement Network (THIN) during 1998-2016, with smoothed lines of best fit (95% CI represented by shaded areas). Presented by UK and constituent nations

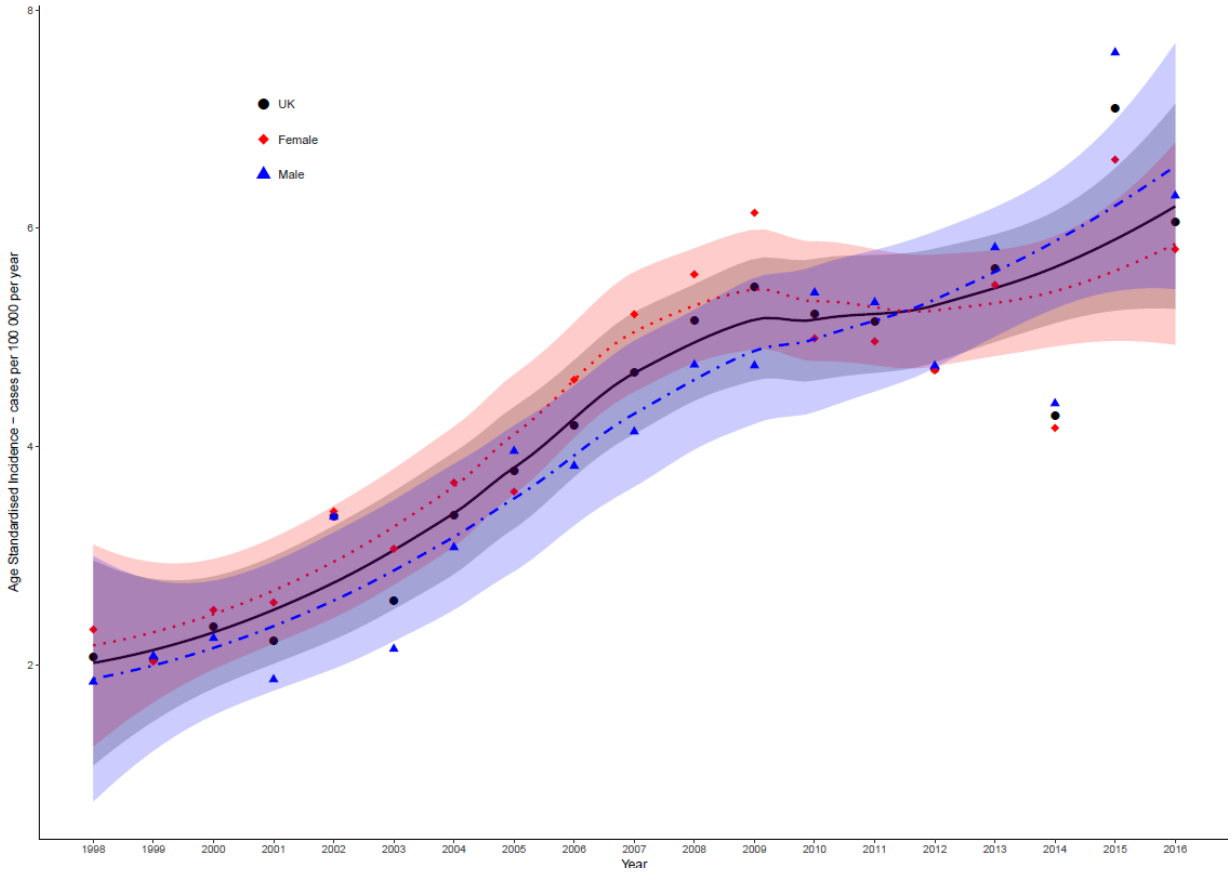


Figure 5.2 Age-standardised incidence rates of Lyme disease coded patients in The Health Improvement Network (THIN) during 1998-2016, with smoothed lines of best fit (95% CI represented by shaded areas). Presented by sex.

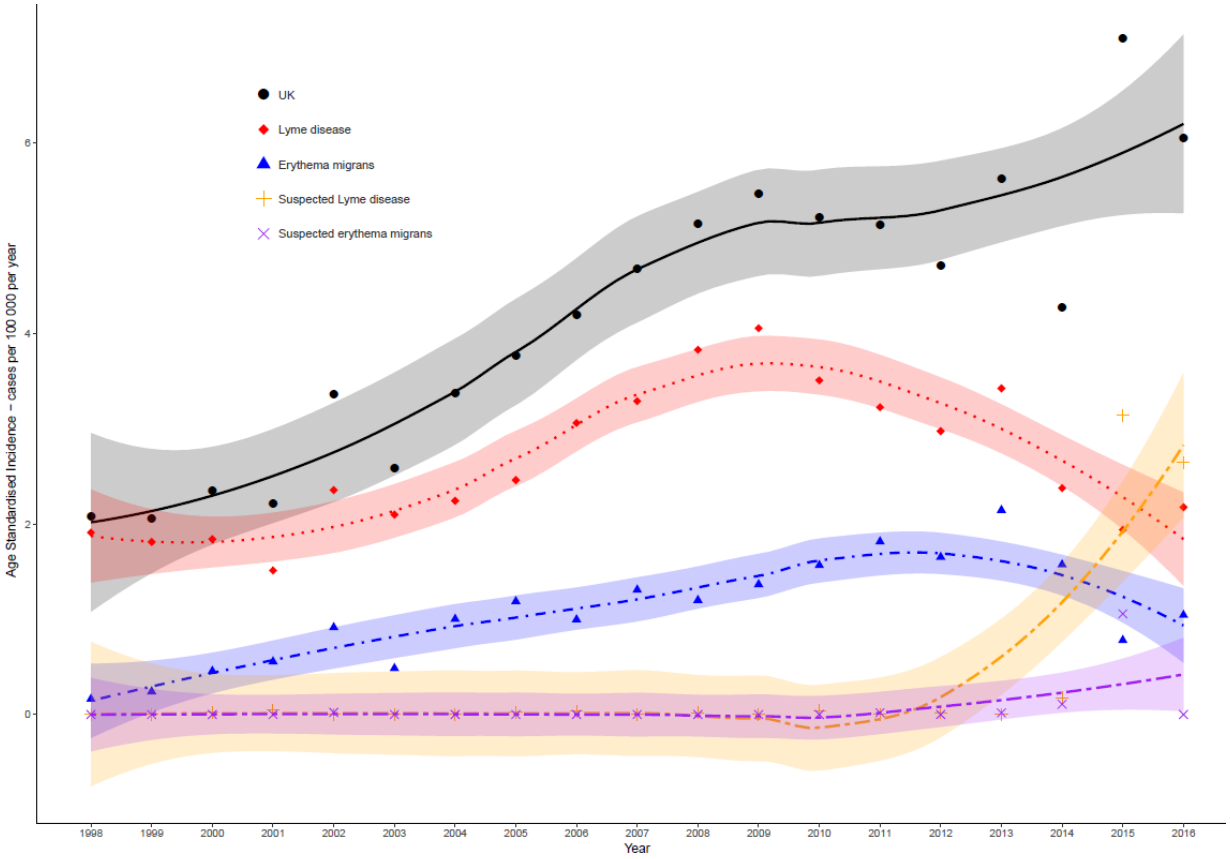


Figure 5.3 Age-standardised incidence rates of Lyme disease coded patients in The Health Improvement Network (THIN) during 1998-2016, with smoothed lines of best fit (95% CI represented by shaded areas). Presented by the four most prevalent Lyme disease Read codes

Across the UK, cases displayed a seasonal pattern, with the highest incidence in the summer and peaking in July and August (Fig 5.4).

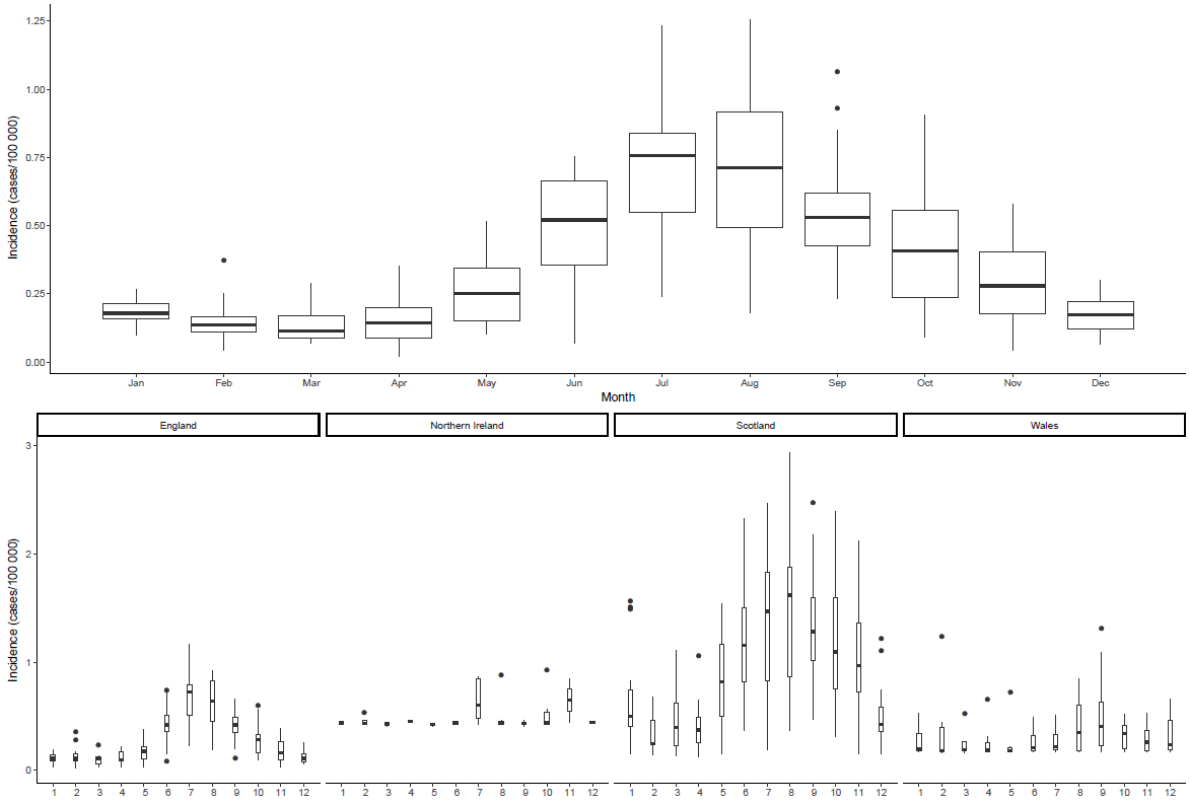


Figure 5.4 Box and whisker plots of monthly incidence rates of Lyme disease coded patients in The Health Improvement Network (THIN) during 1998-2016. A – UK, B – Constituent nations

This seasonality was seen in England and Scotland with incidence peaking in July and August respectively. In Northern Ireland and Wales no obvious trends were seen.

There was a high degree of missing data for ethnicity, with only 35.1% (n=1,306) of cases providing information. Of these, 73.5% (n=960), had an ethnicity description that matched the ethnicity categories defined in the UK 2011 census [209]; the remaining 346 all identified with being 'British/Mixed British'. There was a significant difference in ethnic diversity ($p<0.01$), with a higher percentage of the Lyme disease coded THIN patients (96%) identifying with being white compared to the national population (87%).

Information relating to Townsend score was available for 93.6% (n=3,487) of cases. Quintile five, that with the highest levels of material deprivation, had the lowest incidence rate, 2.22 cases per 100,000 person-years (95% CI 1.94-2.52) (Table 5.2). The highest incidence rate was reported in the second quintile, 6.61 cases per 100,000 person-years (95% CI 6.22-7.01). There was evidence of an overall association between incidence rate and Townsend quintile ($\chi^2=336.99$, $p<0.01$). This association took the form of a trend of decreasing incidence with each increase in quintile number ($\chi^2=197.38$, $p<0.01$). Departures from the trend were significant ($\chi^2=139.61$, $p<0.01$).

Rural urban classifications were only available for English and Welsh cases, as Scottish and Northern Ireland authorities do not record this measure. Of the 2,220 of English and Welsh cases, 93.4% (n=2,073) had useable data. The incidence rate of cases living in rural areas, 5.60 cases per 100,000 person-years (95%CI 5.17-6.05), was significantly higher than those cases living in urban areas, 2.91 cases per 100,000 person-years (95% CI 2.77-3.07).

5.4 Discussion

This study describes the incidence and demographics of Lyme disease coded patients using primary care data in the UK, fulfilling one of the key research needs identified by the NICE guidelines [23]. There has been a significant increase in the annual incidence of Lyme disease coded patients in UK primary care between 1998 and 2016. Incidence rates varied between nations, with Scotland experiencing the highest incidence of disease. There was a higher incidence of Lyme disease coded THIN patients living in rural areas and within areas of lower deprivation.

European studies using similar data from primary care sentinel practices have described a large range in incidence [176,212,213], from 42 cases per 100,000 per year in France [180], to 148 per 100,000 per year in Norway [177]. The UK had a much lower incidence rate across the study period; 4.23 cases per 100,000 person-year. The UK had its peak age-standardised incidence in 2015, 7.10 per 100,000 population (Fig. 5.1). The annual incidence significantly varied between nations; Scotland peaked in 2015 with an incidence of 16.40, England in 2015 with 5.23, Northern Ireland in 2015 with 2.94, and Wales in 2016 with 3.21. Even in Scotland, the incidence of Lyme disease is lower than in most areas of continental Europe. The reasons for this are likely to be multiple and need to be further explored. They may include: a lower prevalence of *Ixodes spp* of ticks, a lower prevalence of *Borrelia spp* carriage by ticks (4.2% in southern England [214], 0-8.2% in northern England [182], and 10.2% in Scotland [37], compared to 13.6% across Europe [215]), and different levels of exposure of the general populace to ticks, possibly due to differences in occupational and/or recreational dispositions. One possible explanation is lower awareness about Lyme disease in the general population and primary care in the UK, compared to the rest of Europe. This would result in fewer presentations to primary care, the potential for mis-diagnosis and a resultant underreporting of cases.

The incidence figures are notably higher than those reported in current surveillance figures based on laboratory confirmed cases. The laboratory confirmed incidence of Lyme disease in England and Wales in 2016 was 1.95 cases per 100,000 (95%CI 1.84-2.06) [76,82,153],

whilst that identified in THIN was 3.77 (95% CI 3.62-3.94). The laboratory confirmed incidence of Lyme disease in Scotland in 2016 was 3.15 cases per 100,000 (95% CI 2.70-3.65) [76,82,153], in THIN it was 12.86 (95% CI 11.93-13.84 The laboratory confirmed incidence of Lyme disease in Northern Ireland in 2016 was 0.21 cases per 100,000 (95% CI 0.07-0.52) [76,82,153], in THIN it was 1.29 (95% CI 0.85-1.89). The large non-overlapping differences suggest that the incidence described in primary care data, for each country of the UK, was significantly larger than that described by official laboratory confirmed cases. This was to be expected as not all cases of Lyme disease (in particular those with erythema migrans) require confirmatory diagnostic laboratory tests.

These differences between nations are notable and likely to be multifactorial. Scottish GPs may be more confident in diagnosing a case of Lyme disease, due to the higher prevalence of Lyme disease compared to England and Wales [153], and so manage more patients within primary care without submitting samples for serological testing. Conversely English and Welsh GPs could be more reluctant to diagnose and treat Lyme disease cases and may refer cases to secondary care sooner than their Scottish equivalents. There may be differences in patient access to primary care or differences in health-seeking behaviour between the different nations, dependent on differing clinical presentations. Further analysis of the THIN database may provide information about case referrals, and differences in case presentation and management. However, the exploration of differences in GP recording or patient behaviour would best be conducted through qualitative research.

The rise in annual incidence of Lyme disease, and the differences in incidence with the laboratory datasets, could be a result of a real increase in disease, an increasing awareness of the disease in the general public, a change in general practitioners' behaviour resulting in the submission of fewer diagnostic samples, or a combination of the above. Further work is needed to understand how general practitioners diagnose and manage Lyme disease cases. Wales is the only country that does not have a significant increase in cases, which may be due to, at least in part, a low number of cases (n=165) and registered THIN practices in Wales. The peak number of cases we observed in summer months is consistent with other studies [75,153,154]. This peak occurs slightly earlier in England than in Scotland. This is likely due to latitudinal, climatic and ecological differences between the two nations impacting on the emergence and peak numbers of nymphal ticks [216]. The low case numbers in Wales and Northern Ireland (n=50) likely explain the lack of an obvious seasonal trend.

The Lyme disease patient demographic have shown predominance in either sex in various settings in the UK [76,153,154,179]. In comparison to other national primary care datasets, Switzerland and France have no statistical difference between sexes, but numerically have more women [180,212]. Finland and Norway have significantly more women [176,177]. The results from THIN indicate no difference in the incidence between sexes (Table 5.2, Fig. 5.2), however, local differences may exist relating to differences in tick exposure or presentation to health services [109].

A few papers have suggested that areas with higher Lyme disease incidence are likely to be less deprived [156,157]. The results presented here support this evidence and build on previous hypotheses that use and access to the countryside is a driver of Lyme disease risk. In England 45% of all outdoor visits were to the countryside, 68% of these were within two miles from their starting point (usually a home address), and that people were less likely to visit if they were from a BAME (Black, Asian, minority ethnic) background, or from a 'DE' social group (i.e. semi-skilled and unskilled occupations, unemployed and lowest grade occupations) [217]. In Scotland, 50% of outdoor visits were taken in the countryside, the average distance travelled from home being 4.8 miles, and that people were less likely to visit if they lived in the 15% most deprived areas, and were of 'DE' social grade; no difference in regards to ethnicity was identified [218]. All ethnic minority groups are more likely to live in areas of higher deprivation compared to the white population, and there is a lack of ethnic diversity in wealthy areas [195]. Taking this into consideration, we believe that members of the general population who live in areas of low deprivation, predominantly rural locations [167,168], are more likely to identify with a white ethnicity, and due to their residential location have greater and closer access to the countryside. This increased potential access to the countryside enables increased risk of a tick bite and therefore subsequent risk of developing Lyme disease. The lack of representation of non-white ethnicity patients may also be due to inadequate healthcare access, lack of Lyme disease awareness, or simply that EM rashes are harder to identify on non-white skin colour [219,220]. The latter assumption would not hold true with other clinical presentation, and it is recommended that ethnicity should be explored in relation to clinical presentation prevalence.

There is obviously a complex interplay between ethnicity, socio-economic status and place of residence of a case, probably related with either outdoor employment or leisure activities. The current descriptive analysis only provides baseline exploratory information on these variables. The ethnicity data has a high degree of missingness, 74%, more so than prior

analyses, 60% [208]. Its representativeness must therefore be questioned; our data only provides a general indicator of the true situation. Our characterisation may be simplified, but the structure of the dataset negates multivariable analysis of these potential risk factors associated with Lyme disease. Future work should be designed to take this into account and allow for the identification of any confounding and interaction between these variables.

With this large scale work we provide UK specific baseline data that is greatly needed for further epidemiological research around Lyme disease, and have fulfilled one of the NICE guidelines identified research needs [23]. We have highlighted new insights into the demographics of Lyme disease patients in primary care. THIN has been shown to be representative of the UK population and, as such, the results are likely to be representative of the Lyme disease cases seen in primary care. However, there are some limitations that must be highlighted. The majority of research investigating conditions within a primary care database also try to validate the Read codes investigated. This is usually via the result of a diagnostic test or a questionnaire of general practitioners [106,221]. Validation of Read codes relating to Lyme disease therefore proves a challenge, as if there is an uncomplicated clinical presentation the clinician is recommended to prescribe antibiotics without performing subsequent diagnostic tests [23]. Therefore, matching a Read code case with a positive test result may be a fruitless exercise. Instead, validation through a GP questionnaire would be recommended. In the majority of cases there will be no confirmatory diagnostics, so GPs would have to confirm a case by remembering the exact consultation, as the information collected by THIN does not substantially differ from what is in the practices' clinical records. Hence, there would be scope for considerable error. Methodology for validating conditions with broad clinical presentations needs to be explored, but this was beyond the scope of this study.

One of the largest limitations of this study is the absence of knowledge about GP coding practices and changes in their coding behaviour. Further work is required to better understand coding practices and how they may vary. The Read codes used by clinicians were consistent until 2010, with the majority being 'Lyme disease' and 'Erythema migrans', at which point the use of these terms started to decline (Fig. 5.3). A year after the introduction of the 'suspected' case codes in 2014, the 'suspected' codes were already more prevalent in use than 'Lyme disease' and 'Erythema migrans'. The reasons for the changes in GP coding behaviour, potentially indicated by changes in code incidence, are unknown; the change may be due to the increasingly politicised landscape of Lyme disease and the debate around

'chronic Lyme disease' [54,222]. We need to know what symptoms are identified to code a patient with 'Lyme disease'; this could be only an erythema migrans rash or another presentation described by NICE [23]. Qualitative research around general practitioners' recognition and coding behaviour regarding Lyme disease would help answer these questions. Only 25.8% (n=960) of the study population had information that could be analysed around ethnicity. We assumed that the trends seen in this subset of patients is representative of the THIN population as a whole; further work is needed to verify this.

Finally, the geographical resolution of THIN only allows us to carry out analysis to the level of the constituent nations of the UK, so analysis of the spatial distribution of incidence with this dataset is not possible. Previous research in the UK has shown clear clustering of cases both from laboratory confirmed cases [23,76,153,154], and hospital admissions [179]. The largest number of identified cases will be in primary care, because not all cases require diagnostics or hospital admissions. Therefore, without greater resolution, we cannot see whether the observed hotspots of disease in laboratory surveillance systems are reflected in primary care activity.

The NICE guidelines highlighted the 'lack of robust epidemiological data' on Lyme disease in the UK, and called for research in this area [23]. This research provides a description of the demographics and incidence of a representative UK primary care population. This work will help ensure the appropriate public health prioritisation of Lyme disease in the UK, however, many basic epidemiological questions remain unanswered. These mainly revolve around person-tick interaction and include; the *Borrelia* spp seroprevalence of the UK population, the total exposure of tick bites to the UK population, and the risk of contracting Lyme disease after a tick bite in the UK. These final two points could be explored using the THIN dataset.

Our data provides the primary care practitioners with basic sociodemographic information about the type of patient who is more likely to present with Lyme disease. This information can be used to raise awareness of increasing Lyme disease presentations in primary case and their seasonality in the UK. This information is critical to their diagnostic clinical decision making and ensures that their clinical suspicion of Lyme disease is increased in suitable situations. This research, alongside the NICE guidelines [23], will raise Lyme disease awareness amongst primary care clinicians and thus ensure that Lyme disease is aptly placed on their differential diagnosis list. Patients are therefore less likely to be misdiagnosed and will be managed more appropriately. This will only enhance Lyme disease patient care. Comparing Lyme disease presentations in primary care with incidence in laboratory

surveillance systems can highlight areas where differences exist regarding awareness, reporting, and management of Lyme disease. These differences would require further investigation. Future research using Lyme disease coded patients within THIN, will investigate concurrent symptoms, and treatment and referral choices as part of case management plans. This study provides a platform to describe patient management in the UK primary care setting and enables ongoing epidemiological analysis of Lyme disease.

5.5 Conclusions

This chapter has described the demographics of primary care patients who are coded with Lyme disease, across the United Kingdom. The demography of this population poses some interesting questions about the relative lack of ethnic diversity and the trend towards habitation in areas of low deprivation, and in rural areas. This chapter provides a platform to inform future work on Lyme disease patients within primary care. Analysis of primary care data can inform and help target health promotion messages, and as this is an ongoing dataset, interventions relating to Lyme disease could be formally assessed.

The extent to which this work contributes to the overall aims of this thesis regarding Lyme disease surveillance can be summarised as follows:

- I**ncidence**: An overall incidence rate of 4.42 cases per 100,000 person-years (95% CI 4.23-4.67). There was an increase ($p<0.05$) from 2.07 cases per 100,000 in 1998 to 6.05 per 100,000 in 2016. Cases peaked in the summer.
- **Sociodemographics**: A slight bimodal age distribution, with no differences between sexes. They were more likely to identify with being white than the national population. There was a significant socioeconomic trend, with the incidence decreasing as societal deprivation increased. The incidence of cases in rural areas was almost double (1.92 per 100,000 (95% CI 1.75-2.11)) that of cases residing in urban areas.
- **Geographical hotspots**: No fine geographical resolution data was available. Scotland had the highest incidence (10.77 cases per 100,000), followed by England (3.42), Wales (1.98), and Northern Ireland (1.32).
- **Patient presentation and management**: No information about patient presentation and management was analysed. Such data were available in THIN and were extensive. This meant that it was beyond the scope of this thesis. However, it would be critical for the data to be analysed so that the current management of Lyme disease patients in primary care could be understood.

- **Additional information:** Read codes have been identified that can be used to identify Lyme disease patients. These codes could be used in other primary care datasets for further research purposes or for the surveillance of Lyme disease in primary care.

Chapter 6 The validation of Lyme disease Read codes through blinded interviews with primary care clinicians

6.1 Introduction

In Chapter 5 the wealth of information that can be extracted from primary care electronic health records was discussed. However, when a disease does not require a confirmatory diagnostic test, how do we know the reliability of Read code defined cases? (Read codes are a standardised coded thesaurus of standard clinical terminology used in primary care electronic health records in the United Kingdom.) Essentially, what is the positive predictive value (PPV) of the Read codes that were used to define a case of Lyme disease?

Many research projects have tried to validate their Read codes and calculate a PPV for them. Two systematic reviews have been performed on this topic [221,223]. To date, no evidence could be found of research into the negative predictive value of Read codes. This would be challenging research, as it would be exceptionally difficult to identify true false negatives in datasets of this scale, based solely on Read codes. Until a robust methodology is designed to calculate negative predictive values for Read codes, interpretation of Read code based incidence studies must bear this in mind. The two systematic reviews reported four ways of calculating a PPV and validating Read codes: comparing incidence figures to a national database: requesting the health records of identified patients to confirm the diagnosis themselves, using questionnaires to general practitioners (GPs) to confirm cases, or a mixture of the above.

An example of the comparison validation methodology is a study which observed the national rates of suicide and self-harm [107]. This compared the incidence described in a primary care database to official figures (Office for National Statistics - ONS). It was concluded that the primary care figures were unreliable, as there was significant under and over-reporting of cases. Another study explored the quality of recording for a QOF ("quality and outcomes framework" introduced in 2004 as part of the contract for general practices in England) dementia record in primary care. The QOF financially rewards primary care practices for the 'provision of quality care' to 'help standardize improvements in the delivery of primary medical services' [224]. Certain clinical conditions, and their associated Read codes, are deemed "QOF-able" if they are of high importance to the health and well-being of the nation. Dementia is one of these QOF-able areas, and practices are paid for the recording of all dementia cases. One would therefore expect a high level of reporting accuracy. Russel et al [225] initially found 1,007 cases on the QOF dementia records, but on

further examination of the practice Read codes, they found an extra 132 patients. This significant difference suggested that the QOF records underreported primary care cases by 12% [225]. A further study described the incidence of non-fatal acute myocardial infarctions in primary care records, hospital records, and a national disease registry [226]; 63.9% of cases were found in at least two sources, with the crude incidence being 25% lower if based solely on primary care records rather than all three data sources.

If it is possible to link cases across multiple databases, one has a much greater ability to get closer to the true incidence of disease. This methodology relies on a reliable national database to work effectively. This proves a challenge with Lyme disease, as most patients are only likely to present in primary care and not appear in another database. Hence, this methodology for validation had to be ruled out for the current study.

The remaining methodologies of sending GPs questionnaires or requesting health records are the commonest approaches, and have been performed for at least forty different diseases [223]. Often studies use a combination of both. There is, however, a huge range of PPV, ranging from 16.5% to 100%. On the questionnaires the GPs are usually asked how they confirmed a case. This can be through hospital confirmation letters, clinical notes, hospital records, electronic health records or laboratory reports[106,221,223,227,228]. Independent validation of the health records can be performed by the research team and has been described as the most robust method, but this will come with a cost, averaging around £70 per clinical note [221]. It has been described that questionnaires alone have a PPV of 91.7%, record request alone, 82.7%, and a combination as 90% [221]. Due to the high expense and ethical challenge of requesting clinical notes, most studies solely use a questionnaire [221]. Within these, the most common methods of case validation is firstly a consultant's letter, followed by clinical notes, and electronic health records [106,221,223,227,228]. Interestingly, one study found high PPVs for cataracts (92%) and glaucoma (84%) based on questionnaires; however, there were huge discrepancies (20-33% accuracy) between reported timings of diagnosis in the primary care database and the questionnaire response [228]. This clearly shows the disparity in how certain elements of clinical notes are recorded with differing amounts of rigour and accuracy.

All the studies captured within these systematic reviews had narrow case definitions, and the cases often involved referral or contact with hospital consultants. One suspects that, due to this, GPs are comfortable assigning a specific Read code to the patient. Lyme disease differs to all of these previous studies in that the case definition is very broad, and the

majority of patients are likely to be managed in primary care without contact from either a hospital or laboratory [23]. If validation is performed through questionnaires, it is likely that confirmation will occur via clinical notes. Therefore, the best method for validation would be requesting the clinical records of each patient identified through our analysis. To do this properly was too costly (estimated at over £210,000) and would exceed the time constraints of this PhD.

Only one primary care database study utilising Read codes shared similarities with Lyme disease, in that it had a vague case definition, and diagnosis does not always require laboratory diagnostics. This was a 2013 study investigating the prevalence of irritable bowel syndrome (IBS) [229]. IBS is a functional bowel disorder whose presentation includes abdominal pain, discomfort with defaecation, and changes in bowel habit and defaecation, amongst other symptoms. This study was unable to validate their codes but found a much lower prevalence in GP coded records compared to the literature. It was also noted that most patients coded with IBS had no other codes relating to gastrointestinal symptoms. The authors noted a few interesting reasons why there is a disparity in prevalence, firstly that;

> '*Read Codes used to code a consultation are at the discretion of the individual clinician, which means that there can be considerable variation in their use to describe the same set of symptoms in practice (particularly for conditions not incentivised in the Quality Outcomes Framework (QOF))*'

They describe how IBS is not a priority for GPs, and that GPs appear reluctant to code patients for IBS and thus recording is patchy [230]. The same authors stated that;

> *'GPs reported that they did not initially add a Read code for IBS to the patient record, but delayed until they were more confident in the diagnosis.* [230]*'*

The authors hypothesised that IBS patients did not feel that primary care can offer help, and that patients felt that doctors were unsympathetic, ignorant about IBS and believed that IBS is psychosomatic. Overall the authors suggested that the alienation felt by some IBS patients meant that they were not accessing health care, resulting in underreporting of cases. They suggest that this shares

> '*similarities with many other medically unexplained symptoms which are typically difficult to diagnose in clinical practice*.'

Even though Lyme disease's case definition is broad [23], it is subjectively more clearly clinically defined than that of IBS. However, one can hypothesise that the potential reasons for underreporting is likely to be similar, with GPs not being incentivised to consistently code, GPs not coding Lyme until a definitive diagnosis is reached, and some patients reluctant to see GPs. The question, therefore, is not how we validate patients coded for Lyme disease, but 'How do GPs code patients with Lyme disease presentations?' If we can understand this, we may have a better understanding of the level of underreporting of Lyme disease cases. It was felt that this question could best be addressed by using a questionnaire.

The decision-making behaviour of GPs regarding the selection of Read codes is poorly understood [231]. It has been acknowledged that the primary care consultation is a complex social and professional interaction and that the recording of electronic and written clinical notes is an important part of this [232]. However, barriers to coding consultations have been acknowledged and these include; the perception that coded data is unimportant, codes having a predominant biomedical focus and are difficult to use in complex clinical presentations, the difficultly in creating a high quality clinical record as well as leading a patient-centered consultation, and that targets and financial incentives drive and distort coding behaviour [231]. It has been reported that clinicians do not see a direct patient benefit to coding their records [231–233]. In addition, if an electronic health record receives a 'definitive diagnosis', via a Read code, it can be harmful to the patient and damage the patient-clinician relationship [233,234]. Through questioning GPs on a disease with a non-specific presentation there is an opportunity to understand some of the motivators and logic behind Read code selection.

Many other studies have used a questionnaire design focused on Lyme disease. None of these have focused on primary care coding, but instead have focused on clinician knowledge and management of Lyme disease cases. In France, overall 50% of clinicians believe that serology is needed to confirm a case; however, if they have training, 72% believe than an erythema migrans (EM) rash is diagnostic [46,235]. A Belgian study found that 17.5% of GPs would offer Lyme disease serology for a tick bite and 9% would prescribe antibiotics. If they identified an EM rash 54% would order serology and 91% would prescribe antibiotics [236]. Dutch GPs had highly varied approaches about when they ordered Lyme disease serology even though national guidelines exist [237]. In the USA, clinician beliefs are very varied on how they manage cases in terms of antibiotic choices, and length of treatment for both tick bites and erythema migrans rashes [238]. One study found that 31% of GPs give antibiotic

prophylaxis for tick bites. When presented with EM with a tick bite history 70.6% prescribe antibiotics and perform a serological test, 18.7% only prescribe antibiotics and 4.3% perform only a serological test [239]. Two separate studies in Quebec found inconsistencies in how GPs manage tick bites and EM rashes [240,241]. A more controversial paper that has provoked extensive commentary [242], found that 2% of GPs believe in and treat 'Chronic Lyme disease', 48.1% are unsure of its existence, and 49.8% don't believe it exists [243].

The two most relevant papers, with very similar methodologies, examined GP beliefs and practices in an endemic area (New Hampshire) [244] and a non-endemic (British Columbia) area [245]. In New Hampshire, 93.9% of GPs knew the causal pathogen, and 52.4% knew EM was diagnostic. When presented with an EM rash, 85.5% would prescribe antibiotics, 13.1% would order Lyme disease serology, 0.6% would do nothing and 0.6% would refer. When presented with a tick bite, 12.9% would prescribe antibiotics, 28.8% would order serology, 57.7% would do nothing, and 1.1% would refer. With a Lyme arthritis presentation, 13.7% would prescribe antibiotics, 48.2% would investigate without serology initially, 4% would do nothing, 7% would order serology, and 29.6% would refer. Overall 89.7% of GPs believe that patients that requested a Lyme disease diagnostic test have another cause of their symptoms other than Lyme disease [244].

In contrast, in the non-endemic area of British Columbia, 98.2% of GPs knew the causal pathogen, and 26% knew EM was diagnostic. When presented with an EM rash, 58.3% would prescribe antibiotics, 36% would order Lyme disease serology, 0.2% would do nothing and 3.4% would refer. When presented with a tick bite, 7.5% would prescribe antibiotics, 40.2% would order serology, 51.3% would do nothing, and 1% would refer. With a Lyme arthritis presentation, 11.1% would prescribe antibiotics, 40.8% would investigate without serology initially, 1.5% would do nothing, 5.7% would order serology, and 40.8% would refer. In consults were Lyme disease was mentioned, patients would initially raise it rather than GPs 57.8% of times. Overall 78.7% of GPs believe that patients that requested a Lyme disease diagnostic test have another cause of their symptoms other than Lyme disease. Thirty one percent of GPs prescribed antibiotics due to Lyme disease related patient concerns, even though they didn't believe the patient had Lyme disease [245].

These studies show the varied approaches of Lyme disease case management in primary care. The inconsistencies are not only present in non-endemic areas but also in endemic areas, and very few GPs appear to be following their nation's guidelines. As this study was designed to capture how GPs code for Lyme disease, it was felt appropriate to also ask

questions about case management. The above studies were used to inform the design of case management questions in our research; the results of which will help inform us about the current management of Lyme disease patients within a primary care setting. This will highlight any need for the future education of GPs in Lyme disease, so that their patients receive the best and appropriate care.

6.2 Aims

To investigate how well GPs recognise the various clinical presentations of Lyme disease, and what Read codes they would use to record those presentations within patient electronic health records. To understand what ideas and motives help shape primary care clinicians coding behaviour.

6.3 Methods

Through discussions with the Royal Liverpool and Broadgreen University Hospitals NHS Trust (RLBUHT) research team, it was decided that the project was to be split in to two phases. These were (a) semi-structured interviews of GPs, as a pilot study, followed by (b) a national blinded online questionnaire.

A pilot study was run for two main reasons. Firstly, no questionnaire-based studies were found where GPs were blinded to the researchers' motives, and it was unknown how GPs would react to being blinded. Secondly, no previous study had attempted to understand coding habits based on case studies. There was therefore a need to ensure the questionnaire worked and that the correct questions were being asked. The feedback from the GPs in the pilot study would enable the design of a better questionnaire and provide ideas on the most successful distribution route for a national online questionnaire.

6.3.1 Semi-structured interviews of GPs

In all the studies mentioned in section 6.1, the GPs knew they were being asked about Lyme disease. This may cause bias through GPs answering with what they perceive to be the correct answer, rather than what they clinically do. In addition to this concern, if the GPs knew what the aims of the study were they may not answer the coding questions honestly, due to Lyme disease being at the forefront of their clinical decision making process. This would impact the validity of any results generated. The GPs were therefore blinded to the fact the study was about Lyme disease; aiming to make their decision making process as close to an unbiased clinical consultation as possible.

Interviews commenced with the presentation of a 'Participant Information Sheet' (Appendix 1), in which the study was entitled 'Understanding the decision making process of general

practitioners (GPs) presented with non-specific conditions'. Thus, answers generated from the questionnaire should reflect GPs coding practices and case management strategy closer to what they would do in a clinical setting. It was felt that through this design, information garnered from this study could help inform future research studies, of not only Lyme disease, but other conditions with a similarly vague presentation.

Eleven clinical cases were constructed based on current clinical case definitions [23,39] and through discussions with infectious disease consultants based at the RLBUHT (Table 6.1). The cases were loosely based on real cases. Most had vague symptomology that could have easily been attributed to a variety of causes other than Lyme disease.

Table 6.1 A summary of clinical cases presented to GPs, for full details see Appendix III

Case	Further details
1 – Classic erythema migrans (EM) rash	Initial presentation of EM. A photo used by multiple organisations, so GPs likely to be familiar with. Then patient revealed to have been bitten by a tick.
2 – Borrelial lymphocytoma	Borrelial lymphocytoma of the ear lobe.
3 – Acrodermatitis chronica atrophicans (ACA)	With associated peripheral neuropathy
4 – Bell's palsy	Later revealed to have insect bite on scalp.
5 – Recurrent synovitis of knees	N/A
6 – Multiple EM rashes	Patient had been walking in Dartmoor
7 – Heart rhythm abnormalities	N/A
8 – Fatigue, post-exertional malaise, anxiety, headaches, and memory issues	Later patient reveals an international lab report saying she has Lyme disease, and she demands long term antibiotics.
9 – Fatigue, arthralgia, poor ability to concentrate, myalgia, mood swings	N/A
10 – Non-engorged tick attached to scalp	N/A
11 – Poor fine motor movements, rash two months previously at scout camp. This had been treated with erythromycin.	N/A

For each case, a series of questions were constructed regarding the GPs' differential diagnoses, how they would code their patient (using Read codes) on the patient's electronic health record, and decisions around diagnostics, prescriptions and referrals. To reduce cognitive bias (recollection of previous cases in the questionnaire could impact their answers to the current case), the cases were presented in a random order, determined by a random number generator.

At the start of the interview, the basic demographic information of each GP was collected. The GPs were told that all patients presented in the cases were otherwise healthy with no other known health problems, prior to the symptoms and history presented to them. The cases were then presented, with answers and comments transcribed. On completion of the cases, the GPs were handed the 'Exit Information Sheet' (Appendix IV), which revealed the purpose of the study. Questions were then asked about the blinding process and whether they found it acceptable. Questions were also asked around the questionnaire structure and how it would work online. Guidance was sought for the best online distribution channels for phase 2, to maximize participation rate. Finally, information was collected regarding their basic knowledge about Lyme disease. The GPs were then provided with resources to learn more about Lyme disease. The questionnaire pack used in the interviews is included in Appendices I to IV.

The interviews were recorded on a dictaphone and transcribed, this enabled the analysis of any discoursive elements of the participants answers to the questionnaire. A thematic approach was taken to explore coding behaviour of the GPs [246]. The transcribed manuscripts were read multiple times to allow thorough immersion of the data. The transcripts were then coded using an inductive approach, this allows for thematic codes to emerge from the data rather than from pre-planned codes. The resultant codes were reread and combined into a set of themes. The themes were then checked with the coded extract and the main corpus. To ensure reliability and repeatability of the results, these themes were then refined with members of the research team until a consensus was reached. The themes identified were then organised and quotations selected to highlight the main themes. Interpretation of the themes occurred to theorize the explanatory causes behind them.

The interviews took place at the GPs' practices and were designed to last between 30 and 45 minutes. On arrival they signed a consent form and were given the opportunity to opt out at any point during the process. They were also given the option to end the interview if it was taking too long, but still allow their results to be used for research. They could also opt out if they felt uncomfortable with the blinding process.

The aim was to identify ten GPs in the Merseyside area, chosen on a purely convenience basis, who would be willing to participate in the study. However, when initial recruitment began in July 2017, there were zero responses. This meant engaging with the Clinical Research Network (CRN) to aid recruitment, and with a resulting larger recruitment area. Recruitment was not complete until August 2018. Recruitment had to stop in November

2017; it would not resume until May 2018. The pilot project was granted REC (NHS Research Ethics Committee) and HRA (Health Research Authority) approval under the IRAS (Integrated Research Application System) project ID: 208815.

The severe delays and challenges faced getting the pilot phase started (Figure 6.1), resulted in Phase 2 not being undertaken as part of this PhD, due to restrictive time constraints. These issues will be further debated in the discussion of this chapter.

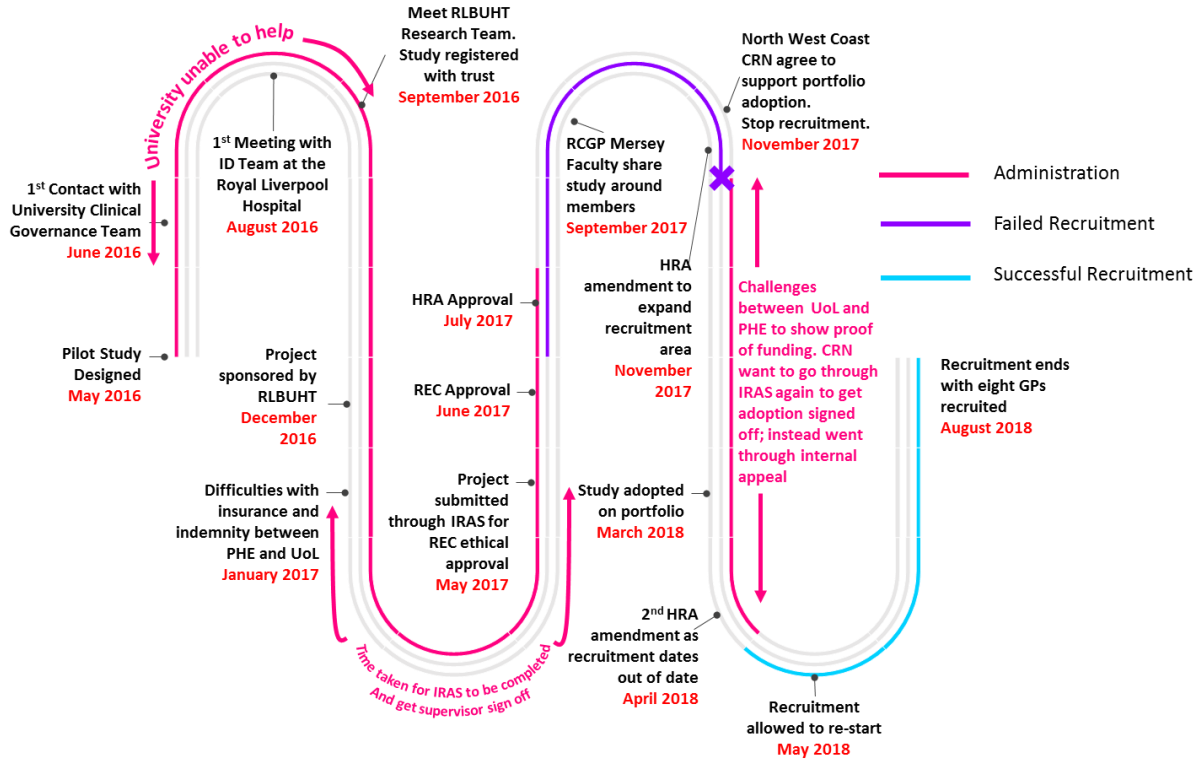


Figure 6.1 Timeline of GP validation study: from study protocol design to end of recruitment. (ID – Infectious disease, RLBUHT – Royal Liverpool and Broadgreen University Hospital Trust, PHE – Public Health England, UoL - University of Liverpool, IRAS - Integrated Research Application System, REC – NHS Research Ethics Committee, HRA – Health Research Authority, RCGP – Royal College of General Practitioners, CRN – Clinical Research Network)

6.4 Results

Ten GPs in the Clinical Research Network (CRN) North-West Coast region were approached to take part in the interviews. All ten responded that they were willing to take part. However, one GP failed to attend multiple appointments, and another never set a date for the interview. The remaining eight GPs had similar enough responses that we felt that data saturation was reached for the purposes of this project [247,248], and further recruitment was not needed.

Five (62.5%) of the GPs were male. The GPs had a mean age of 46 (range: 38-53). The mean time in clinical practice was 16 years (range: 10-26). Three (37.5%) were a GP with a specialist interest. These interests were, diabetes, ear, nose and throat (ENT), and one GP whose were substance misuse, and sexual and reproductive health.

The study area that CRN North West Coast covers includes the regions, South Cumbria, Lancashire, Merseyside and Cheshire (Figure 6.2).

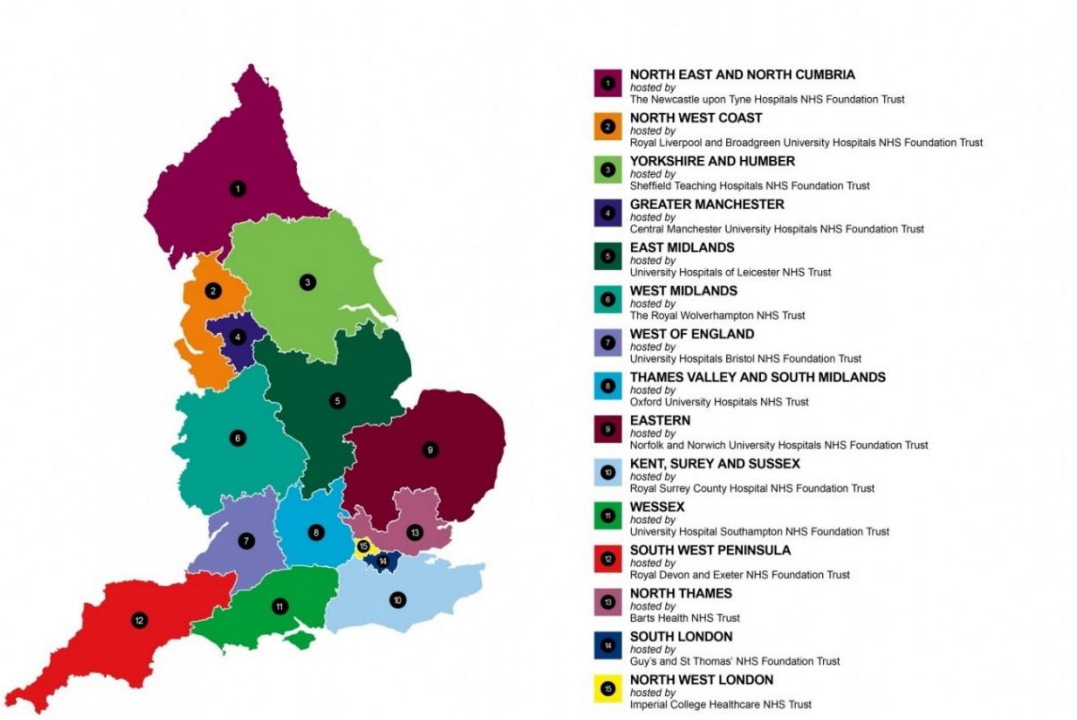


Figure 6.2 The clinical research networks (Image from the UK Clinical Research Collaboration)

None of the GPs were based in South Cumbria, five were based in Lancashire, two in Merseyside and one in Cheshire (Figure 6.3). These areas all have a low incidence of laboratory-confirmed cases of Lyme disease and hospital admissions (Chapters 3-4).

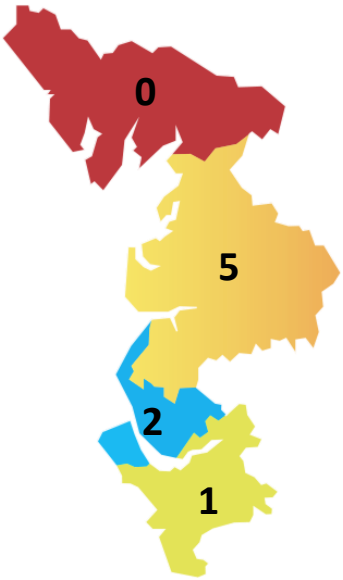


Figure 6.3 The study area (CRN North West Coast), with the number of GPs interviewed in each subdivision. (Image from the Clinical Research Network North West Coast)

The participant GPs took, on average, 43 minutes (range 20-60) to complete the interview. None opted out at any point of the process. Four of the GPs were able to complete all 11 cases, three completed 10 cases, and one GP was only able to complete two questions. This variability was due to time constraints in their professional schedule. The responses to the cases are summarised in table 6.2, for more detailed responses see Appendices V to XV.

Table 6.2 Summary of key results of a blinded case-study questionnaire about coding and diagnosis regarding Lyme disease. If the case had split sections, the cells in the table are similarly split.

Case, and number of GPs presented with it	Top three differential diagnosis, where n>2, and any mention of a Lyme disease diagnoses		Top three Read Codes, where n>2, and any Lyme disease Read Codes		Diagnostics (Y/N)	Prescription (Y/N)	Referral (Y/N)
1: Erythema migrans (EM) rash (n=7) See Appendix V	**Lyme disease** Bite *(10 differentials were mentioned in total)*	**71.4%** 28.6%	Rash **Lyme disease** **Erythema migrans** **Suspected Lyme disease** *(5 Read codes were mentioned in total)*	42.6% **28.6%** **14.3%** **14.3%**	71.4%	71.4%	57.4%
	Lyme disease *(3 differentials were mentioned in total)*	**85.7%**					
2: Borrelial lymphocytoma (n=7) See Appendix VI	No differential was mentioned more than once. *(13 differentials were mentioned in total)*		Ear swelling *(5 Read codes were mentioned in total, and one GP wouldn't code the case)*	28.6%	28.6%	42.9%	57.1%
3: Acrodermatitis chronica atrophicans (n=8) See Appendix VII	Alcohol related issue Vitamin B12 deficiency *(19 differentials were mentioned in total)*	50% 37.5%	Peripheral neuropathy Sensory loss *(4 Read codes were mentioned in total, and one GP wouldn't code the case)*	37.5% 25%	100%	25%	50%
4: Bell's palsy (n=7) See Appendix VIII	Bell's palsy Brain tumour Space occupying lesion *(9 differentials were mentioned in total)*	100% 42.9% 28.6%	Bell's palsy Nerve palsy *(5 Read codes were mentioned in total)*	57.1% 28.6%	71.4%	14.3%	85.6%
	Bell's palsy Brain tumour **Lyme disease** *(7 differentials were mentioned in total)*	100% 28.6% **14.3%**					

5: Recurrent synovitis of knees (n=7) See Appendix IX	Osteoarthritis Rheumatoid arthritis Gout *(9 differentials were mentioned in total)*	71.4% 57.1% 42.9%	Knee pain Synovitis *(4 Read codes were mentioned in total)*	42.9% 28.6%	100%	100%	42.9%
6: Multiple EM rashes (n=7) See Appendix X	Insect bite Fungal infection **Lyme disease** *(11 differentials were mentioned in total)*	85.7% 42.9% **42.9%**	Non-specific rash **Erythema migrans** *(6 Read codes were mentioned in total, and one GP wouldn't code the case)*	28.6% **14.3%**	42.9%	85.7%	0%
7: Heart rhythm abnormalities (n=7) See Appendix XI	Arrhythmia Palpitations *(11 differentials were mentioned in total)*	57.1% 42.9%	Palpitations *(2 Read codes were mentioned in total, and one GP wouldn't code the case)*	71.4%	85.7%	28.6%	28.6
8: Fatigue, post-exertional malaise, anxiety, headaches, and memory issues (n=5) See Appendix XII	Thyroid problems Chronic fatigue syndrome Depression *(8 differentials were mentioned in total)*	80% 60% 60%	Tired all the time *(3 Read codes were mentioned in total, and one GP wouldn't code the case)*	60%	100%	20%	60%
	Thyroid problems CFS Depression **Lyme disease** *(7 differentials were mentioned in total)*	60% 40% 40% **20%**					
9: Fatigue, arthralgia, poor ability to concentrate, myalgia, mood swings (n=7) See Appendix XIII	Depression Anaemia Thyroid problems *(13 differentials were mentioned in total)*	71.4% 42.9% 42.9%	Fatigue Polyarthralgia *(4 Read codes were mentioned in total, and one GP wouldn't code the case)*	42.9% 28.6%	100%	71.4%	14.3%

10: Non-engorged tick attached to scalp (n=6) See Appendix XIV	Cutaneous horn Foreign body Skin lesion **Tick** *(10 differentials were mentioned in total)*	33.3% 33.3% 33.3% **33.3%**	Skin lesion **Tick bite** *(4 Read codes were mentioned in total)*	33.3% **33.3%**	66.7%	16.7%	0%
11: Poor fine motor movements, rash two months previously (n=8) See Appendix XV	Depression Neurological condition **Lyme disease** *(18 differentials were mentioned in total)*	50% 50% **12.5%**	Feeling low Low mood Neurological symptoms *(4 codes were mentioned in total, and one GP wouldn't code the case)*	25% 25% 25%	87.5%	0%	37.5%

All the GPs had heard of Lyme disease. Two (25%) correctly gave the causal pathogen as *Borrelia burgdorferi*, five (62.5%) said it was caused by tick bites, and one (12.5%) did not know what the cause was. One (12.5%) knew that there was a national reference laboratory at the Rare and Imported Pathogens Laboratory (RIPL), Porton Down, one (12.5%) thought that it was at the Liverpool School of Tropical Medicine, and one thought it was at the Royal Preston Hospital. The remaining five (62.5%) did not know that there was a reference laboratory. Five (62.5%) GPs felt that current testing available on the NHS was reliable, two (25%) did not know, and one (12.5%) thought that it was not reliable. Six (75%) GPs thought that chronic Lyme disease exists, and all eight GPs recognised that Chronic Fatigue Syndrome exists. Four (50%) of the GPs would seek advice about Lyme disease from an infectious disease consultant, three (37.5%) from a consultant microbiologist, and one (12.5%) from a Lyme disease specialist.

6.4.3 Thematic analysis results

Thematic analysis of the interview transcripts yielded five main themes relating to coding behaviour; personal experience of the GP, evidence-based medicine, diagnostic uncertainty, professional integrity and defensive practice, and patient and primary care politics. The personal experience of the GP underpins the other four themes and is the central core that runs through each of the other themes. Due to this, examples of this core theme will be illustrated with quotations nested amongst the other themes' quotations.

Evidence-based medicine and diagnostic uncertainty

Primary care clinicians are taught to diagnose and manage patients in an evidence-based manner [249]. This can mean that without any strong diagnostic or clinical evidence, the GP may have diagnostic uncertainty and choose a Read code that is relatively non-specific.

> *'I think first, 'What is the primary complaint?', and unless there is something definitive, I tend to code with the primary complaint or not code at all. So for this patient, what is the primary complaint? Numbness, dropping things, or is it his blue hands?'* [GP5, case 3]

Here we can see that GP 5 is unsure of the diagnosis and does not have the prior personal experience or strong evidence base to choose ACA as the definitive diagnosis. They therefore suggest that they will adopt a highly generic code for this patient.

GP 3 showed similar behaviour. They wanted a stronger evidence base, and lacked personal experience, this resulted in diagnostic uncertainty. They would choose not to use any Read code at all for the patient.

> *'If I can't diagnose, I will pick the main symptom to code. I will always do this unless I'm almost a hundred percent positive of the diagnosis. Sometimes, if I'm really not sure, I will write everything in free text and not code anything.'* [GP3, case 2]

GP 6 shared the same feelings very succinctly;

> '*If I don't know what it is, then I won't code it. Otherwise it gives the appearance of certainty, where there isn't.'* [GP6, case 6]

The need to build a strong evidence base was strengthened by multiple GPs enquiring whether the patient in case 1 had a significant travel history, and that some would perform internet searches or ask their colleagues about unusual rashes. In all instances the GPs asked for a better description, about the firmness and texture of the tick (case 10).

This shows that GPs need a certain level of supportive evidence, which will vary for each condition and for each GP, to have certainty and confidence in their diagnosis. This can be clearly seen in the results of case 1 (Appendix V) where the majority of GPs thought the case was Lyme disease, but then coded the patient with a rash. This suggests that without diagnostic certainty a non-specific presenting symptom will be chosen as a Read code rather than one with a definitive diagnosis.

This is exemplified by several GPs who noted that they had heightened diagnostic uncertainty regarding Lyme disease, and without sufficient confirmatory evidence, would not code it as such. This may be due to a lack of personal experience,

> *'Lyme disease is a possibility here. But I wouldn't leap to it without a history of a tick bite.'* [GP5, case 11]

If they had had more supportive evidence or personal experience, they may have been more confident in their diagnosis and then coded Lyme disease. In many of the cases, Lyme disease was never discussed. It is probable that either the GPs' knowledge of Lyme disease's various presentations is poor, or that GPs do not think of Lyme disease as a differential diagnosis until they have a greater evidence base, or that Lyme disease only enters their differential list through a diagnosis of exclusion.

However, at times personal experience can become its own form of evidence base, and when this occurs diagnostic certainty improves and the Read code selected becomes highly relevant.

> *'This is a tick; I've been bitten many times before.'* [GP 4, case 10]

GP 8 had recently been involved in the recent management of a Lyme disease case;

> *'I had a patient diagnosed in the last couple of months; a child with non-specific knee pain. We initially suspected an infected knee joint. He's now been successfully treated and has been fine since.'* [GP8]

This resulted in this GP having much greater personal awareness of Lyme disease presentations; this GP confidently placed Lyme disease on their differential lists and chose relevant Read codes.

Professional integrity and defensive practice

Professional integrity drives certain coding behaviour. Here we see a GP showing frustration around how other members of their practice don't respect practice policy nor understand the need for appropriate coding;

> *'We're told off if we don't code all patients. Others in the practice use the code 'chat with patient', it drives me mad!'* [GP7]

This was seen more acutely where a GP felt hindered in coding Lyme disease due to both a lack of diagnostic evidence, and the desire to protect themselves and their colleagues from litigation;

> *'I would never write Lyme disease on a patient's record until I had a positive lab diagnosis. I'm wary because of potential litigation, and I don't want to cause problems for future doctors treating that patient.'* [GP 2]

Patient and primary care politics

The current political landscape around Lyme disease and the lack of trust in certain evidence bases results in a diagnostic uncertainty,

> *'I won't code Lyme disease until they'd seen a NHS specialist. I'd be very suspicious if it* [laboratory results] *was a 'high street' or 'internet' lab, so I would arrange serology to be sent to a local lab.'* [GP 4, case 8]

> *'I would be very suspicious of results from a non-NHS supported lab. She also has no supportive history for Lyme disease.'* [GP3, case 8]

One of the GPs helped place this landscape in context by discussing their own experiences with a cohort of patients in their practice;

> *'We've had a few patients, around five, who have formed a little chronic Lyme disease group. All white middle aged, very wealthy women. They have spent thousands at the Breakspear* [250] [a private clinic and laboratory offering unvalidated testing and treatment for Lyme disease] *and claim to have Lyme disease, but present more with ME* [myalgic encephalomyelitis] *or not coping with life. They don't believe that Porton do their tests properly, and they are now majorly concerned about getting Anaplasma too.'* [GP5]

This political landscape is the framework in which all coding behaviours are contained. It will vary for every disease and medical condition, and with the awareness that patients can ask for direct access to their health records. For example, depression is a mental health condition that is QOF-able [224] and is on PHE's [251] and the UK governments policy agenda [252]. There is heightened awareness of this condition and its QOF-able nature, which could lead GPs to have confidence in diagnosing and coding their patients with a 'depression' read code. In contrast, cancer is also QOF-able and is a series of conditions with huge awareness amongst the general public. However, it is likely that a GP would not code a suspected patient with cancer, unless they had confirmation from a referral department. The consequences of misdiagnosis could be severe, and a GP would not want to place 'cancer' on a patient's health record without any pathological evidence first as it could lead to litigation and unwarranted medical and legal investigations.

Lyme disease sits within a very confused political landscape. On the one hand NICE guidelines [23] have recently been published and the national incidence is low [70]. GPs should therefore have confidence in making a diagnosis and should not feel the need for laboratory diagnostics for the majority of cases, who are likely to present with an EM rash. On the other hand, there are now multiple patient groups and charities raising awareness about Lyme disease, but also discussing 'Chronic Lyme disease', a condition not currently accepted by any medical organisation [21,23,39,43,53–55]. 'Chronic Lyme disease' is the narrative that is often portrayed in the media rather than that of research or medical organisations [12,13,253]. These contrasting narratives make it a challenge for both the GPs and patients

to sift through scientific fact and fiction [254]. This was made evident by two of the GPs interviewed

> *'Loads of people [patients] are asking about it* [Lyme disease] *in the practice, through I've only seen four or five over my career'* [GP 5]

> *There are so few* [Lyme disease] *specialists across the country* [in fact, there are none]. *A friend of mine has Lyme, so I know the difficulties.'* [GP 4]

These issues have now been potentially confounded by the Medical Defense Union offering specific advice about Lyme disease [255], something which is rather unusual for a disease that is, in general, readily treatable and has a low incidence. This complex political landscape could result in GPs being hesitant to diagnose and code a patient with Lyme disease without confirmation from a laboratory or referral unit.

6.4.2 Study design feedback

Three (37.5%) GPs realised that the case series were about Lyme disease. GP 2 said they suspected this on their third case; case 11 (fine motor movement issues and rash). This was self-confirmed when the following case was case 1 (the classic EM rash). GP 3 suspected it on their final case; case 1 (the classic EM rash), and GP 6 suspected it on their eighth case; case 6 (multiple EM rashes).

All the GPs did not mind being blinded and felt that unless several stereotypical Lyme disease cases were presented sequentially, they would likely remain blinded in a future questionnaire. To improve the blinding they suggested shortening the questionnaire to five random cases, with additional dummy cases and to finish the case series with the most obvious case (case 1).

All GPs requested a more thorough case history to be included for each case. General feedback that was mentioned by many GPs, was to add the following questions after being unblinded;

> *'Have you seen Lyme disease in your practice, if so how many?'*

> *'How many chronic Lyme disease patients have you seen?'*

They also requested that the number of clinical questions were reduced and to ask more questions about coding, for example:

> 'W*hy did you choose that code over another?'*

'Why did you have that diagnosis on your differential list but not select it as the Read code?'

To aid recruitment the GPs suggested that it should be advertised as a continuing professional development (CPD) learning exercise, with official Royal College of General Practice (RCGP) accreditation. A couple said that they would not have done it online as they prefer doing this type of research in person. And another GP said that the only way to attract GPs to the study was via compensation for their time.

Regarding the distribution of the questionnaire online at a national level, there was no overwhelming consensus. Organisations that were suggested as distributions channels were; the RCGP (2 GPs mentioned), the University of Liverpool (2), the local Clinical Commissioning Group (CCG) (2), the local primary care network (2), a local NHS leader (1), the Clinical Research Network (1), and any organisation with an NHS badge or affiliation (1).

6.5 Discussion

This study has provided insight not only in to the coding behaviour of GPs in relation to Lyme disease, but coding behaviour in general. The results and feedback from the questionnaire will help shape the design of a more detailed Read code validation study for Lyme disease.

In Chapter 5, 'Lyme disease' and 'Erythema migrans' represented 89.1% of all Lyme disease related Read codes, and suspected codes represented 7.6%. In this validation pilot study, GPs only used the codes, 'Lyme disease', 'Erythema migrans' and 'suspected Lyme disease'. This appears to be reflective of the codes used in Chapter 5. The remaining 3.3% of codes identified in Chapter 5 were not mentioned by the GPs in this study; this likely relates to the small sample size of GPs interviewed, and the rarity of the general usage of these codes in primary care. Only two cases were coded with Lyme disease codes. To formulate an idea of the sensitivity and specificity of these codes is difficult, as so few GPs utilised the Lyme disease codes, and the study size is small. One could speculate that the specificity is likely to be high as Lyme disease codes were only used in the 'textbook' clinical presentations, and there are likely to be very few false positive codes, given the case descriptions. It is therefore possible that the positive predictive values of the Lyme disease Read codes are high. The sensitivity of the codes is likely to be low, with a lot of false negatives with cases not being coded at all; reasons for this are discussed above. If these conclusions are true it means that our incidences calculated in Chapter 5 are likely to be an underestimate of the clinical incidence of Lyme disease. Phase 2 of the project needs to progress, in order to take this theory beyond conjecture.

The thematic analysis suggests that when a GP is presented with a case of vague symptomology or there is diagnostic uncertainty, they will not code with a highly specific and definitive code such as 'Lyme disease'. They prefer to code the patient's primary complaint, the main symptom, or not to code at all. They will only code with a definitive diagnosis code if the evidence base or personal experience is strong enough to justify it. Harkness et al's [229,230] investigation of GPs' perspectives on IBS, appears to support the identified themes. The GPs in that study were constrained by the politics of giving a patient an IBS label, and they wanted to have confidence in their diagnosis (through an evidence base, and personal experience) before coding with IBS. For most cases they would code with symptoms of IBS. Both their data and this work suggest that for these diseases a surveillance system based upon specific Read codes is likely to underreport cases. This may also be relevant to other diseases with either a non-specific presentation or with certain political constraints.

Some of the themes identified match those identified in previous research about coding behaviour [231–234]. These include diagnostic uncertainty and a GP's personal experience. In these studies, the GPs state that they were fearful of the potential negative impact of giving a patient a definitive diagnosis through a Read code. This is something that we had speculated in relation to Lyme disease but was not explicitly stated by the GPs in our research. Previous research has also identified that the GPs personal experience and 'emotional motives' can drive decision-making behaviour [233]. This can be based simply on the recognition of a similar case or on the previous negative experiences of an incorrect or missed diagnosis of a patient. These works and our study highlight that human emotion and experience may be a larger driving force, in some instances, of decision-making behaviour than the medical evidence base alone.

Themes that had been identified in other studies but not in this chapter include; the lack of direct patient benefit, targets and financial incentives driving coding, and the negative impact coding can have whilst trying to lead a patient-centred consultation [231]. These may not have been discussed by the GPs in this chapter as no questions were directly asked about coding behaviour. It may also be due to the small biased population of GPs that were interviewed and that they may not be representative of the UK GP population. However, a targets and incentive theme could mirror or reside in the professional integrity theme identified in this research. This chapter builds on previous evidence and supports the evidence that diagnostic uncertainty and personal experience are critical elements of a GPs coding behaviour. The other three themes of evidence-based medicine, professional

integrity and defensive practice, and patient and primary care politics, need to be explored to assess whether they can be generalised across the UK primary care clinician population.

Syndromic surveillance, using a variety of Read codes to build a case definition, may offer an alternative approach for primary care disease surveillance. However, problems identifying cases could still remain if the presentation is relatively non-specific. Take, for example, case 6 in which the following non-Lyme disease Read codes were mentioned; rash, insect bite, non-specific dermatitis and skin lesion. In isolation, or in combination, these codes are still non-specific and the potential differential list is so long that it would be very difficult to conclude that this patient actually had EM rashes. The budgetary requirements to request the medical free text notes of all patients that presented like this, to identify a definitive diagnosis, would be very large. Challenges still remain regarding surveillance of conditions with a non-specific presentation.

The thematic analysis highlighted that the GPs personal experience core theme driving coding behaviour. The data from the exit interview questions showed that this group of GPs had limited clinical exposure to cases of Lyme disease. This should not result in poor clinical and case management knowledge. Some acknowledged their lack of personal experience and did mention that they would enhance their evidence base (through internet searches and discussing cases with colleagues). If this had been done in reality, maybe diagnostic certainty would have improved, and more cases would have been coded with Lyme disease. This lack of knowledge needs to be explored at a national scale to evaluate the potential for regional biases. A recommendation resulting from this work is the provision of a national programme of CPD (continued professional development), promoting NICE (National institute for health and care excellence) guidelines and aiming to arm GPs with the appropriate knowledge to manage cases of Lyme disease.

GP 5 was the most vocal about their views on Lyme disease. They had the strongest views on the demographics of chronic Lyme disease patient groups and discussed their symptoms in what may be perceived to be a condescending or dismissive manner. They tried to clarify why they do not believe that this sub-set of patients do not have chronic Lyme (they believed the symptoms were psychosomatic) and what they perceive to be the case definition of chronic Lyme disease. GP 5's opinions were likely constructed through the themes described, but with the greatest weight attached to personal experience and the politics of patients and primary care. It is essential to remember that both the BIA (British Infection Association) and NICE guidelines do not recognise chronic Lyme disease as a clinical entity [23,39]. These

views were given anonymously and in confidence, so they may not be reflective of the terms and tone used when consulting and managing these patients directly. However, even taking this in to consideration, it mirrors the opinion and experience of some patients that a proportion of GPs do not believe them and that they are quickly dismissed, something also seen in the IBS study [230]. This has been nicely summarized previously,

> '*To make matters worse, some* [Lyme disease patients] *have grown frustrated or cynical with the medical profession because of ineffective treatments, unsatisfying explanations and fruitless testing. A commonly expressed perception is that physicians become impatient or dismissive once it becomes apparent that a patient's symptoms are medically inexplicable. In other words, a dominant feeling is that the suffering of these patients is not effectively heard or validated.* [54]'

In contrast to North American studies [244,245], the responses to the diagnostics and treatment option questions were often very generic. However, the referral location was often very specific. As the interview was supposed to replicate the online experience proposed for phase 2 of the study, the interviewer could not ask for a more specific response in relation to what blood test would be used or which drug would be prescribed. The reasons for the generic responses could be due to diagnostic uncertainty and a lack of personal experience. The fact that they knew where to refer suggests that they have an idea of the disease pathology even if they do not know the exact cause. Hence, GPs' responses were not necessarily indicative of them not knowing the recommended care pathway for Lyme disease patients, but rather show a reluctance to commit to a specific treatment plan without diagnostic certainty and an evidence base supportive of the cause of an unfamiliar or vague clinical presentation. These questions did not provide much insight in to how GPs manage Lyme cases, and therefore should not be included in phase 2 of the project. However, they could form the basis of a separate non-blinded study about how GPs manage cases of Lyme disease, similar in design to studies mentioned in the introduction.

The feedback on study design will prove useful to the design of the questionnaire for phase 2 of this project. Unfortunately, there was no consistent distribution channel that the GPs preferred. This indicates that to achieve the highest penetration and response rate of an online questionnaire, a multi-channel and multi-organisational approach is needed. The overwhelming response of the GPs was that they preferred doing the questionnaire face-to-face and may not have even responded or would have engaged less if they had been sent it online. This poses a fundamental research problem, which is well beyond the scope of this

project; how do researchers increase engagement rates with GPs? Their own suggestions involved financial or CPD incentives. The creation of a successful national questionnaire may need large financial and time commitments, especially if face-to-face interviews are required. For Phase 2 of this project to go ahead very careful planning is needed to ensure that resources are used as efficiently as possible.

6.5.1 Limitations

The two main limitations of this work were the potential impact of selection bias and the small sample size. As this was a pilot project for a proposed national study, the GPs were chosen for convenience and were unlikely to be representative of the national GP population. The GPs that were recruited were from a biased population as they were drawn from those GPs showing an interest in taking part in research. They had a narrow age range, and even within the small recruitment area were predominately based in Lancashire, with only one GP from a rural practice. The practices were all in areas with a low laboratory-confirmed incidence and a low number of hospital admissions for Lyme disease (Chapters 3-4). The GPs therefore may reasonably not recognise a relatively rare disease, for their location, and any of its varied presentations. Hence why they were reluctant to code. This inherent bias could be reversed if the study was performed in an area of high incidence. If GPs had higher awareness and saw more cases, they may feel more comfortable coding a Lyme disease patient. If this study had been performed at a national scale, these biases could have been examined and explored in more depth, providing greater insight in to coding behaviours.

However, as this was a pilot study and the main aim was to test the questionnaire structure and blinding format, these biases and small sample size could be, in the main, overlooked. The objectives were achieved and additional insight in to GPs coding and views about Lyme disease were gained, which will help inform the design of phase 2 of this project. Due to the above, the themes identified have the potential to be unrepresentative of the national GP population. Although it is encouraging that the results from Harkness's work [230] appear to support the themes. They provide a theoretical template that need to be tested in future research studies and on a larger scale.

The most substantial issue faced during this project was not inherent in the design or conduct of the study, but rather was the result of the logistics necessary to gain NHS ethical approval and recruit the GPs, as outlined in Figure 6.1. Tradition dictates that a scientific manuscript should be an impersonal document; however, science is a discipline that involves a huge

amount of personal investment. The adoption of the 'scientific style' can sometimes mask fundamental constraints to a project, and I want to briefly give my personal reflections on the logistics of this project. I feel that this is important as these issues should be formally documented, so that lessons can be learnt, and processes adapted. I want no other researchers to go through the same frustrating process that I did.

On completion of the design of our study, no one within the research team, including myself, were fully aware of the required research ethics for the project. I contacted the Central Research Ethics Support team at the University of Liverpool, who were unable to offer any support other than to say that NHS ethics would be needed, and that this would supersede University ethics. No clear pathway was given, and I was given contacts to seek help, which led to a fruitless chain of emails eventually ending up with the original contact provider. The University Clinical Governance Team was then contacted who again provided no clear route or advice on how to gain ethical approval. After three months of no progress, one of my supervisors mentioned meeting with the Royal Liverpool Hospital's research team. They had a wealth of experience and knew the requisite requirements and documents needed to gain successful ethical approval. They were extremely helpful and guided me through getting the study registered and sponsored by the Royal Liverpool and Broadgreen University Hospitals NHS Trust. The ethics team at the hospital were surprised that the University were unable to provide assistance, as they believed my request was very straight forward. Their help meant the relatively speedy sponsorship of the project. The project then ground to a halt with a disagreement between PHE and the University about who was responsible for the project in terms of insurance and indemnity. This is a fundamental issue that could have been agreed upon before the creation of the Heath Protection Research Units (HPRU). Once this was resolved, the IRAS document took longer than expected to be signed off. On reflection, not all supervisors should have been listed on the project documentation, but I had not been advised to the contrary. Once the IRAS document had been submitted a year had elapsed.

The project's REC and HRA approval was smooth and very quick and the initial recruitment phase began. The University provided no support of how best to recruit GPs and I was left to contacting all GP practices in Merseyside by phone, email and in person. After this recruitment drive, not one GP was willing to take part in the research. It appeared that the non-medical staff in the practices worked as such good gate-keepers that researchers were effectively denied access to GPs. Even with the Merseyside RCGP's branches help I was unable to recruit. This suggests that more needs to be done to train non-clinical members of

primary practices about the importance of research. One supervisor then informed me of the existence of the CRN-North West Coast, whom they had only recently learnt about. In hindsight, the CRN-North West Coast should have been involved with the project from the beginning. The CRN-North West Coast agreed to help, but to get their assistance I went through a labourious six-month process of increased bureaucracy and delays. Once everything had been signed off they were exceptionally helpful in providing contact details of GPs who were willing to take part in the project. From completion of the study design to the end of recruitment took 27 months.

This whole process involved a vast amount of administrative paperwork and multiple email chains with a wide range of organisations; there appeared to be a significant lack of training provided by any organisation. The Royal's research team provided support in gaining sponsorship and IRAS submission. However, if I had not had a member of their staff on the research team I would have been unable to utilise their services. I was one of the first University of Liverpool students that they had worked with. The CRN were exceptionally useful once the paperwork was completed. Their system for project adoption was very complicated and the same documents were sent to them multiple times, with a large number of people involved. For example, proof of funding was sent to five separate people within the same organisation. The University and PHE also struggled initially to find this documentation, which resulted in one supervisor writing a letter to confirm my funding. During this 27 month period it was discussed at multiple supervisor meetings whether we should drop this part of thesis. I felt that this chapter was important in placing Chapter 5 in context and would benefit other researchers; if it was not for my personal tenacity phase 1 would not have been completed. The organisation that I feel the most let down by is the University of Liverpool; the fact that the main assistance I received came from outside of the University framework is unfortunate.

The time taken from project planning to interview completion has meant that phase 2 of the project could not be done within the time period of this PhD. It remains to be seen whether this can be picked up as either part of a post-doc or via PHE. The ratio between the volume of paperwork needed and time invested, compared to the resultant eight interviews is undesirable. There is a substantial failing of the system to support qualitative or mixed-methods researchers embarked in NHS based research. I have spoken to other researchers and colleagues working in human health about these issues, based both in the Institute of Infection of Global Health (Epidemiology and Population Health department) and Institute

of Psychology Health and Society, and in the HPRU's Emerging and Zoonotic Infections, and Gastrointestinal Infections research units. It appears to be a consistent problem when the project is not a hospital-based study, does not involve patients but NHS staff, or there are elements of qualitative research involved. This one project, that should have been a small part of the PhD both in terms of time and results, has proved to be the most frustrating part and one of the biggest users of my time. My feelings have been succinctly summarised by another researcher experiencing the same problems at a different University, highlighting that this appears to be a universal problem; [256]

'This [emotional] *exhaustion was exacerbated by the insufficient acknowledgement of, planning for or engagement with the emotional labour required in conducting this type of research.'*

Considering all of the above, I have four recommendations:

- The University of Liverpool ethics team and Clinical Research Governance team need to be trained in how to approach NHS ethics, with a focus on how to deal with non-clinical trial, non-hospital based qualitative research. They need to be able to appropriately train/inform early career researchers in this process and have this service fully signposted.
- The University of Liverpool and the CRN need to have a more cohesive relationship, with the CRN's profile heightened within the University.
- Researchers, and especially post-graduates, need to be made aware of the ethics pathway to be taken and the time needed for a successful project. Ideally a map or standard operating procedure should be created.
- Supervisors and funders should be made aware of the potential time taken to start these projects and be provided with training. This would enable them to aid in the efficient use of an early career researcher's time and be aware of the difficulties and frustrations in the process of being able to perform qualitative or mixed-methods research. Finally, they should be able to assess whether small projects (like this one) are achievable and worthwhile in the context of a PhD or Master's thesis.

The University of Liverpool needs to urgently address these issues if it wants to produce timely and relevant research and continue to be seen as a leader in health care research.

6.6 Conclusions

There is reluctance by GPs to code with specific diagnostic Read codes when they are presented with a patient with either a vague or unfamiliar symptomology. This needs to be explored further through additional qualitative research, and through the testing of our identified themes. In relation to Lyme disease Read codes, it is likely that the codes have a high specificity, a low sensitivity, and a high positive predictive value. This pilot project has been successful in amending the design of a future national survey to investigate coding behaviour.

In relation to the aims of this thesis, this project does not provide epidemiological data regarding Lyme disease. However, it places the results of the primary care dataset in context; and suggests that it probably underestimates the real incidence of disease. This poses many problems for surveillance in general, not just for Lyme disease, and research is greatly needed to understand what drives GPs' coding behaviour.

Chapter 7 Scoping the potential of Twitter as part of a Lyme disease surveillance system

Previously in this thesis, multiple routinely collected health data have been reviewed for the information they contain about Lyme disease cases. These are likely to form the basis of any future Lyme disease surveillance system. The progression downwards through the Lyme disease surveillance pyramid (Fig 2.1), from laboratory cases, to hospital cases, to primary care cases, has now reached the level of unreported cases in the general population. As discussed in Chapter 2, describing and identifying cases that have no contact with National Health Services is challenging, and validating these cases even harder. However, there is an opportunity to use novel and unexplored datasets and research methods to attempt to describe this population. In the next two chapters such attempts are described. Firstly, the analysis of a social media platform, Twitter, to understand how its temporal and spatial trends may match those of the known epidemiology of Lyme disease. In Chapter 8, the analysis of companion animal electronic health records are explored to provide insight into tick activity across the UK.

7.1 Introduction

Social media are websites and applications that allow users to share and create content, and to participate in building social networks [130]. In the United Kingdom (UK) two thirds of adults now use social media, and 22% of over 15 year-olds (roughly 14.6 million people) use Twitter as a social media platform [257]. Twitter is a micro-blogging website and application that allows users to post messages of up to 280 characters (tweets), and to then share and interact with other users' tweets [258]. Users can 'retweet', repost or forward a tweet of another user, with or without their own comments. Users can also show their appreciation for a tweet by 'liking' it, by selecting a heart-shaped icon alongside the tweet. Tweets not only contain user-composed written information but also data about the time and location of when and where the tweet was sent. Twitter users in the UK are more likely to be male, and from a managerial, administrative or professional occupation [259]. In the Republic of Ireland, 28% of the population over 15 years old use Twitter, this represents around 1.3 million people [260]. Users of Twitter use this platform to network and to discuss a wide range of topics, including aspects of human health [261]. Twitter therefore has the potential to be a large, data-rich resource for research.

In a questionnaire-based study, the reasons for the general public and health professionals' use of social media in a health related context were explored [261]. Patients predominantly used Twitter to stay updated on new healthcare developments, increase knowledge, express emotions, and compare themselves to other patients, in relation to a specific disease or condition affecting them. Conversely, health professionals used it to extend their professional network, self-promote their work, promote their workplace, or share information from medical conferences. It has been hypothesised that by observing the patterns of tweets and their content, one may be able to use Twitter as a disease surveillance tool, and understand patients concerns and emotions regarding a disease [129,133,136].

Many studies have explored Twitter as a surveillance tool [129,133–135]. However, due to methodological restrictions exercised by Twitter these studies have focused on very specific diseases, for a set period of time, in a specific location. Due to the large scale nature of Twitter's data, Twitter only allows free access to a certain number of tweets in a specific time period [262]. If this set limit is exceeded, only a random selection of tweets will be displayed; how these tweets are chosen is unknown to the general public [263]. There are two ways of dealing with this problem; firstly to pay for all tweets resulting from a specific search query; which can be costly. Alternatively, queries must be restricted, via content, geography or time, to avoid reaching the maximum number of tweets allowed. A team at San Diego State University has worked in-depth to develop improved methodologies for surveillance of influenza and pertussis [264,265]. They were unable to perform surveillance at a national level and resorted to performing multiple searches in different cities. This suggests that Twitter surveillance may only be feasible in restricted geographical areas.

To date, the diseases explored using Twitter have had a very high public health importance and tend to have a high incidence or awareness in the researched areas. If the number of tweets is relative to incidence or awareness, then these prior projects had a heightened risk of reaching the tweet collection limit. By investigating Twitter surveillance methodology on a disease of a relatively low incidence, information may be able to be collected at a national level, without reaching the tweet collection limit. A disease that potentially fulfills these criteria is Lyme disease.

In Chapters three to five, the incidence and geographic distribution of Lyme disease has been described through a variety of routinely collected health records. In the Republic of Ireland, only the neuroborreliosis presentation of Lyme disease is under surveillance by the Health Protection Surveillance Center (HPSC) [266]. Whilst the incidence of the more common

erythema migrans presentation is unknown, the HSPC reports an annual incidence for Lyme neuroborreliosis of 0.44 cases per 100,000 population in 2016 [267]. In contrast the incidence of neuroborreliosis in the UK is unknown. The HSPC has produced one map of cumulative crude incidence rate of Lyme neuroborreliosis from 2012-2016, which shows the highest incidence of disease being located in the south-west of the country [268]. Since Lyme disease has a relatively low incidence, social media surveillance, utilising Twitter data, may be achievable for the entirety of the United Kingdom and the Republic of Ireland.

Previous research has been performed utilising social media regarding Lyme disease; however, no studies have been exclusive to the UK and the Republic of Ireland. Two studies have explored the content of YouTube videos, but were primarily focused on content analysis on who influences public health messages [269,270]. Two papers discussed the potential of using Google Trends for understanding seasonality [271,272], and another looked at physicians' browser search history in relation to seasonal trends [273]. No published research to date has focused on Twitter data and considered the extent to which the spatio-temporal data generated through Twitter matches the known epidemiology in a given territory. If Twitter data were to match epidemiological data, it would have the potential to show temporal and spatial trends in real-time and potentially identify new geographical areas with disease.

The aim of this chapter was to explore whether the tweeting habits of British and Irish Twitter users matches the known spatio-temporal epidemiology of Lyme disease in these respective countries.

7.2 Methods

Methodological design was inspired by the work of Nagel et al, and Allen et al [264,265]. Tweets were collected through the TwitteR package [274] in R, for the keyword 'Lyme' using the Twitter Search and Streaming Application Programming Interfaces (APIs). Tweets were collected daily from the 1st of July 2017 till the 30th June 2018 inclusively. To collect tweets that were sent from the UK and the Republic of Ireland, the query was geographically restricted to a 375 mile radius around the geographical centre of Great Britain (Whitendale Hanging Stones: 54.016674, -2.566153) (Fig. 7.1).

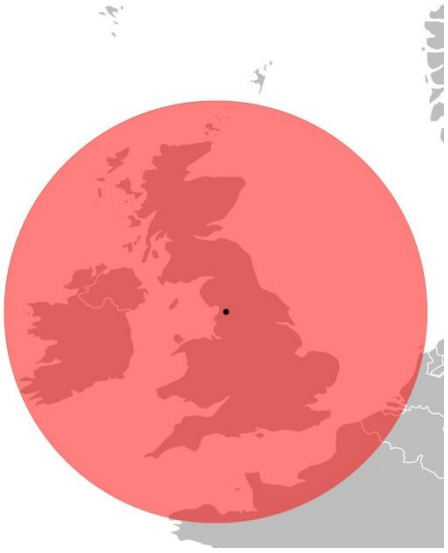


Figure 7.1 The search area queried for 'Lyme' in tweets between 1st July 2017 and 30th June 2018. The search area was centred on Whitendale Hanging Stones.

This did include a small section of northern France, Belgium, and The Netherlands. The query was therefore further restricted to tweets written in English. It unfortunately excluded the Shetland Isles, as to include these islands would have meant expanding the search area to include even larger parts of continental Europe and Scandinavia, increasing the chance of geographical based errors.

Data collected with each tweet included: the user name of the tweet's author, the time and date of the tweet, if the tweet is either a retweet or an original tweet, and the tweet location. The tweet location is based upon the latitude and longitude of where the tweet was sent from a GPS-enabled device. Researchers can only access this information if the user has allowed this data to be shared publicly (i.e. the Twitter account was location enabled). In this study, the users' self-identified location was additionally collected from their user profile. A user's profile location can be any alphanumerical combination; it can therefore be as specific as a grid reference or as vague as 'Planet Earth.'

From this database, tweets were removed which referred to locations that included 'Lyme' within their name, for example Lyme Regis and Newcastle-under-Lyme. This was done to exclude any irrelevant tweets and minimise geographical bias. Time series of daily and monthly count data of both original tweets and retweets were constructed. These were visually compared to the reported seasonality of Lyme disease incidence (Chapters 3-5).

The users' self-identified location was matched with local authority areas in England (n=326), Scotland (n=32), Wales (n=22), and the Republic of Ireland (n=31), and local government districts in Northern Ireland (n=11). Using the Office for National Statistics (ONS) [159] and

the Republic of Ireland's Central Statistics Office (CSO) [275] population data as the denominator population for each area, incidence was calculated. This was defined as the number of users per 100,000 population per local authority per year. Choropleth maps were plotted for the incidence of original tweet users and combined original and retweet users.

Exploratory spatial data analysis (ESDA) [164,165] was used to explore the geographical association between the laboratory-confirmed incidence of Lyme disease (Chapter 3) and Twitter user incidence data at local authority level. This analysis had to be restricted to England and Wales as the available laboratory-confirmed incidence data (RIPL) contained data for only these nations. The bivariate global Moran's I value, and a bivariate Moran's I scatterplot, were created to assess the relationship between each dataset in each local authority. A bivariate LISA (Local Indicators of Spatial Association) plot was constructed to identify spatial clusters and geographic outliers in both datasets.

All statistical and spatial analyses were carried out using R language (version 3.2.0) (R Core Team 2015). Results were deemed significant where $p<0.05$.

7.3 Results

Between the 1st of July 2017 and the 30th June 2018, 56,402 tweets containing the word 'Lyme' were collected through the Twitter API. Data were collected on 98.6% of study period days; connectivity issues with the API led to five days of missing tweets (13th July 2017, 29th July 2017, 30th July 2017, 1st May 2018, and 2nd May 2018). Tweets which included place names relating to the word 'Lyme' were removed; Lyme Regis (n=13,266), Newcastle-under-Lyme (n=9,191), Lyme Bay (n=2,922), Lyme Park (n=1,641), and misspellings of Ashton-under-Lyne (n=61). This led to 49% of tweets being removed, leaving 29,321 to take forward for analysis. Of these tweets, 46.9% (n=13,757) were original, the remaining (n=15,564) were retweets.

7.3.1 Temporal Analysis

To allow a time series of the data to be constructed, the five missing daily tweet counts needed to be imputed. As the level of missingness was small, it was deemed appropriate to use a mean imputation approach. The missing value was estimated to be the mean of the daily tweet count for the month the missing data was located in. The resultant mean number of total daily tweets was 81, the median 65, with a minimum of 15, and maximum of 643. The mean number of original daily tweets was 38, median 33, a minimum of 12 and maximum of 276. The total daily tweets showed some degree of seasonality, with the smallest number in winter (Fig. 7.2). This seasonality was harder to identify when just the

daily original tweets were considered. There was a locus of extreme outliers in both total and original tweets between 21st and 23rd August 2017.

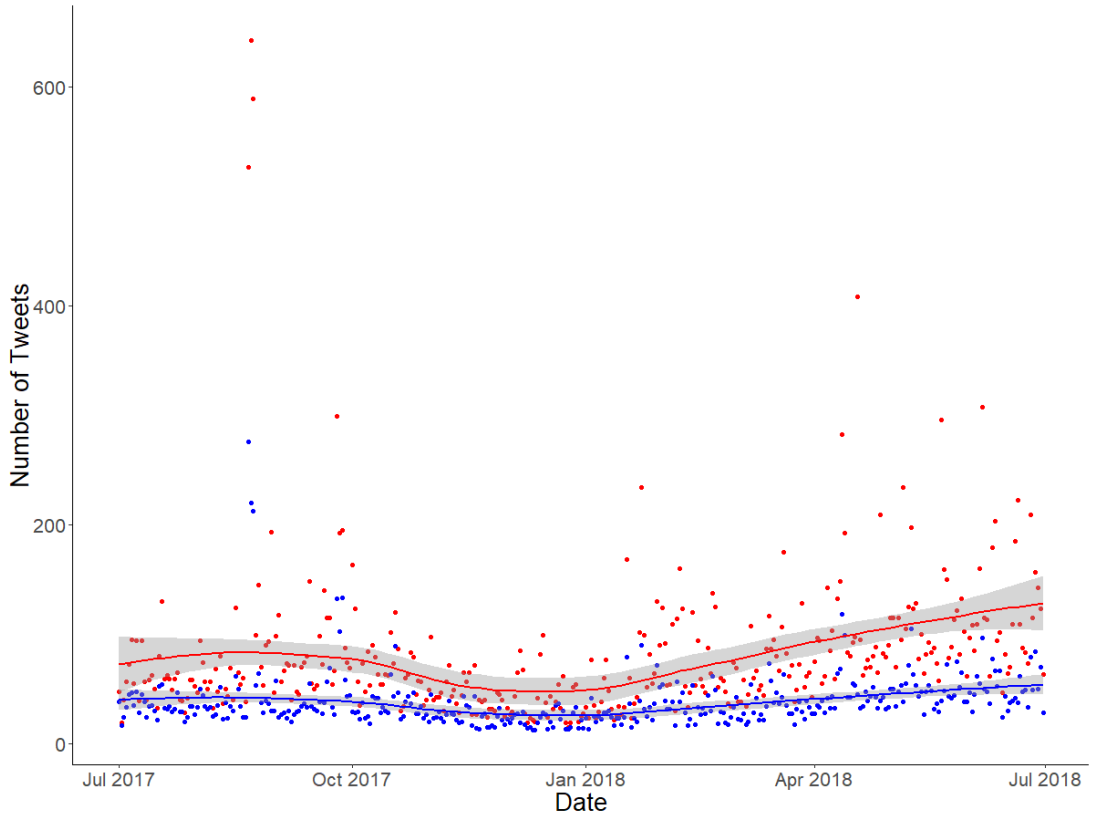


Figure 7.2 Number of daily tweets mentioning 'Lyme' between 1st July 2017 and 30th June 2018. Smoothed lines of best fit (with grey 95% confidence intervals) were fitted with the LOESS method. Red- All tweets, Blue – Original tweets only.

Total tweets and original tweets were aggregated to produce a monthly time series (Fig. 7.3). This suggested a clearer seasonality with peaks in summer months for both datasets.

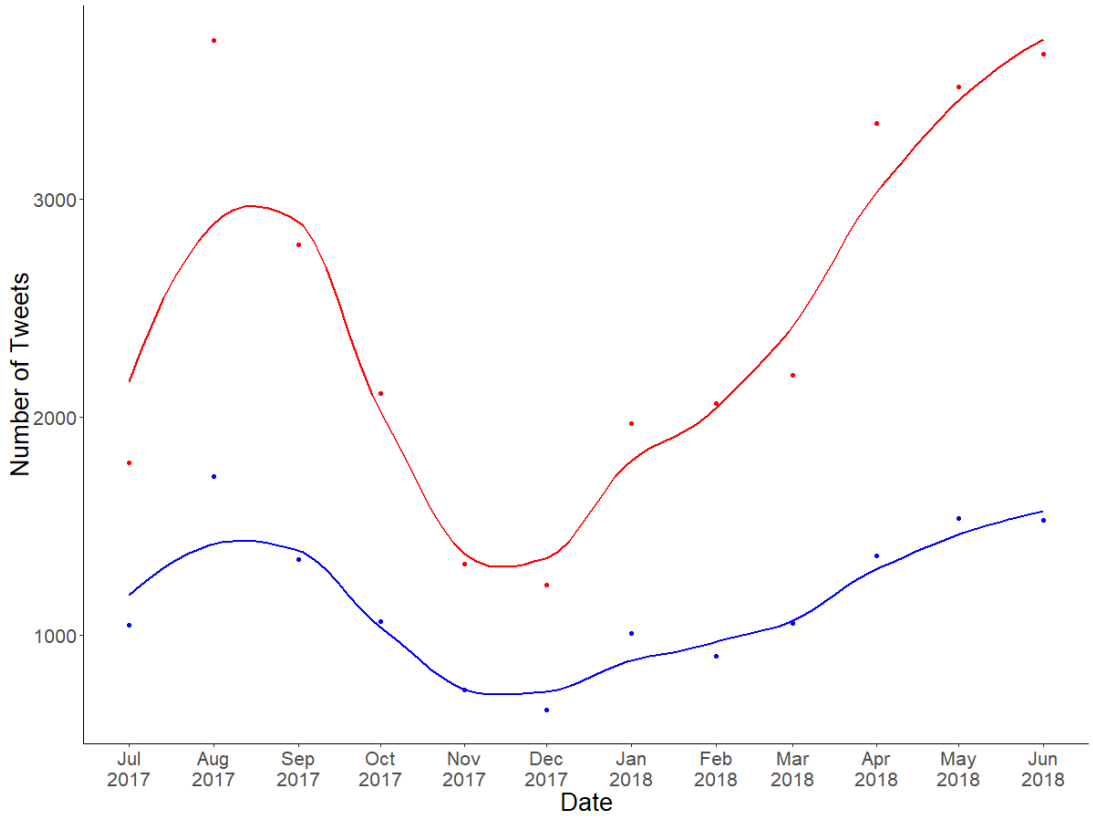


Figure 7.3 Monthly count of tweets mentioning 'Lyme' between July 2017 and June 2018. The smoothed lines were produced using the Loess method. Red- All tweets, Blue – Original tweets only.

7.3.2 Geographical analysis of original tweets

There were 5,212 Twitter users who tweeted original tweets about Lyme disease in the study period, with a mean number of 2.64 tweets per user. Eighty-nine percent (n=4,662) of users had a self-identified location on their user profile that could be matched to a real-world location. The remainder either did not provide a location or were fictional locales; for example, locations mentioned included, Narnia, Westeros, and Gallifrey. Out of the location data, 1.35% (n=63) were outside of the UK and the Republic of Ireland; the top three locations were the USA (n=31), Germany (n=4) and The Netherlands (n=4). In addition, 40.54% (n=1,890) of users had a location in the UK and the Republic of Ireland, which could not be matched to an administrative area as the location was too broad (Table 7.1).

Table 7.1 Number of Twitter users, per broad geographical area, that couldn't be matched to a local authority area

Location	All tweets	Original tweets
UK	**836**	**570**
England	**342**	**225**
London	1012	767
South	158	109
North	109	57
Midlands	26	45
Scotland	**128**	**48**
Wales	**86**	**48**
Northern Ireland	**20**	**11**
Republic of Ireland	**37**	**11**
Jersey	4	10
Guernsey	3	1
Isle of Man	10	4

The remaining 58.1% of tweets (n=2,709) had a location that could be matched to a local authority. There were 6.9% (n=27) local authorities in the UK, and 25.8% (n=8) local authorities in the Republic of Ireland, that had no Twitter users who tweeted about Lyme disease. Users who identified their location with a 'Lyme' related place name, where also removed in case they biased the results; 53 from Lyme Regis and 71 from Newcastle-under-Lyme. There was a mean of 3.7 Twitter users tweeting about Lyme disease per 100,000 per local authority, a median of 2.1 and a maximum of 50.8. Only 110 (0.8%) of all original tweets were location enabled, allowing the exact co-ordinate plotting of where the tweet was sent from. There appeared to be a higher incidence of users in the south-west England and the Highlands of Scotland. Location enabled tweets tended to be in the south of England (Fig. 7.4).

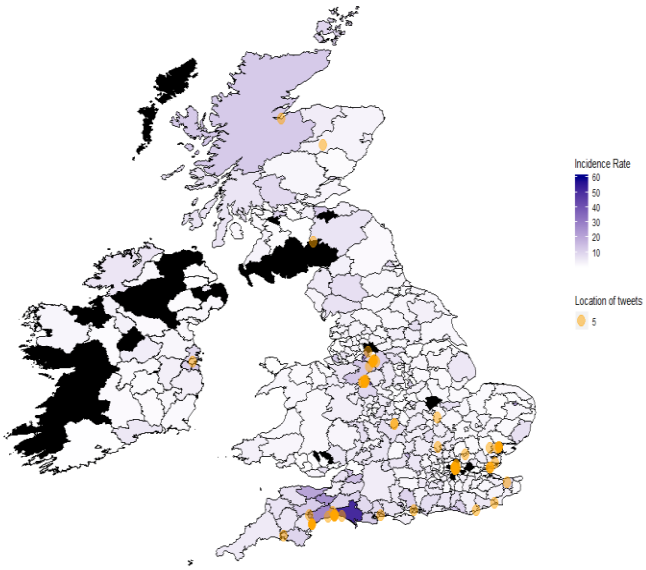


Figure 7.4 The incidence of Twitter users per 100,000 population who tweeted about 'Lyme', and the location of location enabled tweets (in orange) that mentioned 'Lyme'. Black areas had no Twitter users who tweeted about 'Lyme'

When the user locations were compared to the disease incidence of the RIPL dataset for England and Wales (Chapter 3), there was a significant Global Moran's I of 0.17 (p=0.002), indicating an overall significant positive spatial correlation between the two datasets (Fig. 7.5a).

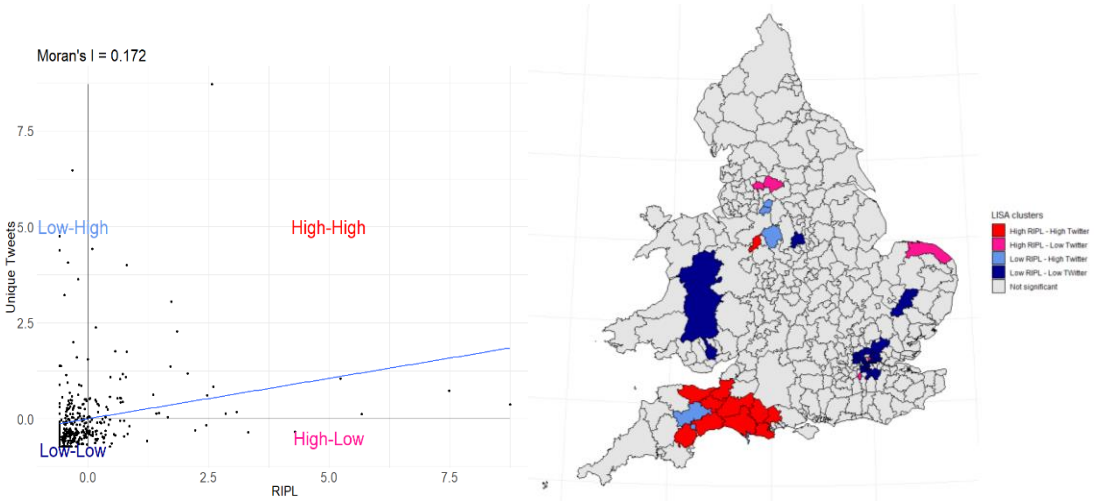


Figure 7.5a Bivariate Moran's I scatter plot between Lyme disease incidence and original tweets in England and Wales local authorities. 7.5b Bivariate LISA cluster map between RIPL laboratory-confirmed (RIPL) Lyme disease incidence and original tweets

This shows a positive linear relationship, in which local authorities (and their neighbours) with a higher laboratory-confirmed incidence tend to have a higher incidence of Twitter users creating original tweets about Lyme disease, and vice-versa. This overall correlation is seen on the bivariate LISA plot (Fig. 7.5b), which shows significant concordant high incidence clusters in the south-west of England.

7.3.3 Geographical analysis of all tweets (i.e. including retweets)

There were 12,168 Twitter users who tweeted or retweeted about Lyme disease in the study period, with a mean number of 2.4 tweets per user. Seventy-six percent (n=9,214) of users had a self-identified location on their user profile that could be matched to a real-world location. Out of the location data, 17.0% (n=1,562) were outside of the UK and the Republic of Ireland, the top three locations were the USA (n=688), Canada (n=181) and Australia (n=76). Almost one-third, 31.8% (n=2,928), of users had a location in the UK and the Republic Ireland, which couldn't be matched to an administrative area (Table 7.1).

The remaining 51.3% of tweets (n=4,724) had a location that could be matched to a local authority. There were 3.1% (n=12) local authorities in the UK, and 9.7% (n=3) local authorities in the Republic of Ireland, that had no Twitter users who tweeted about Lyme disease. Users who identified their location with a 'Lyme' related place name, where also removed; 75 from Lyme Regis and 99 from Newcastle-under-Lyme. There was a mean of 2.0 users per 100,000 per local authority, a median of 1.1 and a maximum of 82.5. Only 110 (0.4%) of all original tweets were location enabled. There appeared to be a higher incidence of users in south-west England and the Highlands of Scotland. Location enabled tweets tended to be located in the south of England (Fig. 7.6).

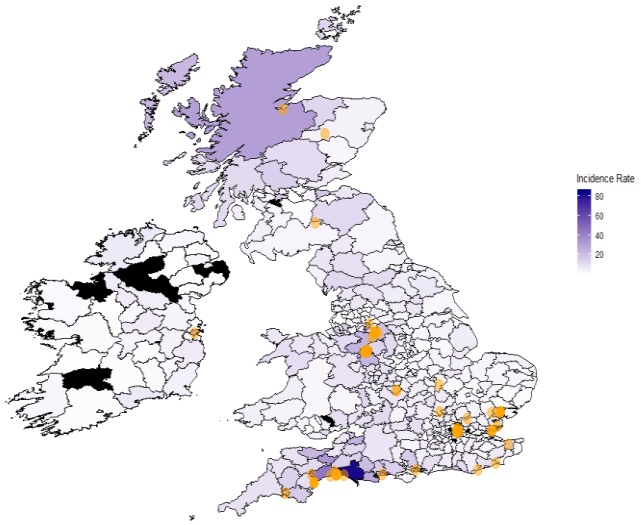


Figure 7.6 The incidence of Twitter users per 100,000 population who tweeted or retweeted about 'Lyme' (i.e. All tweets), and the location of location enabled tweets (in orange) that mentioned 'Lyme'. Black areas had no Twitter users who tweeted about 'Lyme'

When the user locations were compared to the disease incidence of the RIPL dataset (Chapter 3), there was a significant Global Moran's I of 0.18 ($p<0.001$), indicating an overall significant positive spatial correlation between the two datasets (Fig. 7.7a).

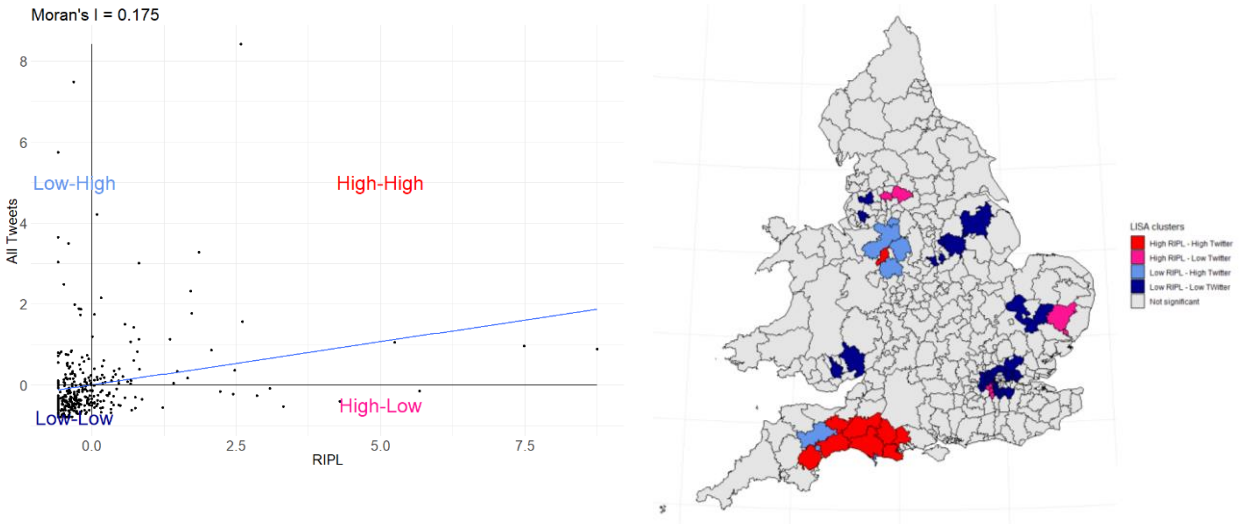


Figure 7.7a Bivariate Moran's I scatter plot between Lyme disease incidence and all tweets in England and Wales local authorities. 7.7b Bivariate LISA cluster map between RIPL laboratory-confirmed Lyme disease incidence and all tweets

This shows a positive linear relationship, in which local authorities (and their neighbours) with a higher laboratory-confirmed incidence tend to have a higher incidence of Twitter users, and vice-versa. This overall correlation is seen on the bivariate LISA plot (Fig. 7.7b), which shows significant concordant high incidence clusters in the south-west of England.

7.4 Discussion

Previous studies have demonstrated that Twitter data has the potential to mirror disease seasonality and outbreak data [134,135,264]. Here we provide further evidence that Twitter can be used for the temporal estimation of disease incidence. However, it is the identification of concurrent disease hotspots, potentially in real-time, that provide exciting new avenues of research that could influence future disease surveillance strategies.

Analysing tweets at a daily resolution level offers only a small insight into seasonal trends. As suggested be Nagel *et al* [265], analysis of only the original tweets appears more robust as the range of the daily tweet count is much smaller and original tweet datasets appear less susceptible to outliers than the combined dataset. However, both original tweets and retweets have extreme outliers in August 2017. These coincide with a high profile press release, which gained significant media attention, as part of a pharmaceutical company's campaign to raise awareness about ticks and to sell a pet ecto-parasiticide [276]. This campaign involved utilising a celebrity sportsman's own case of Lyme disease to raise awareness of Lyme disease and the risk that ticks pose to people and animals [12,277–279], despite the lack of an evidence base for a high incidence of clinically affected pets with Lyme disease in the UK [280]. This shows that Twitter data can be easily skewed by high profile

media and/or celebrity events; this must be considered if Twitter data is used for modelling and surveillance purposes. There were parallels in the research on Lyme disease videos on YouTube, which found that the most viewed and liked videos had included a celebrity component [269,270]. Hence, high levels of disease-related tweets may result from a range of stimuli; these may include increased disease occurrence but may reflect other social signals.

When aggregated to monthly counts, the data showed strong seasonality, with peaks in the summers of 2017 and 2018. Monthly counts were chosen, so that if monthly national figures are released, these figures can be compared. This matches the known seasonality of Lyme disease in the UK (Chapters 3-5). This data currently cannot be matched to the monthly disease incidence, as the UK and Irish governments have not released data covering the entirety of this period and they only record quarters rather than monthly data. Without content analysis the drivers of tweet seasonality remain unknown. It is likely to be multifactorial and to be reflective of, among other things, users tweeting about being diagnosed, the general public's concern, media stories and awareness, and public health awareness campaigns.

One of the exciting potentials of Twitter-derived data is the exploration of the geographical distribution of Twitter users tweeting about a disease. The fact that the user incidence maps produced here reflected the known spatial epidemiology of Lyme disease closely is highly interesting, as no known publications have described this for any infectious disease. The majority of users in the present study had location information that could be matched with a local authority area. The bivariate plots and global Moran's I figures showed a strong spatial agreement with laboratory-confirmed cases for England and Wales. This highlights that Twitter has the potential to map the geographical distribution of an infectious disease, and that it could be used in real time, to identify suspect disease hotspots. By focusing on the LISA cluster map of the original tweets (Fig. 7.6) an understanding can be gained of the relationships between Twitter data and national Lyme disease incidence figures. The majority of the map is covered in non-significant local authorities, showing that in these localities the incidence for both datasets does not differ from a randomly generated distribution of incidence. The high-high areas represent concordant clusters of high incidence. Except one area in the Midlands, these are all neighbouring areas in the south-west of England, an area already identified as a potential hotspot of Lyme disease in this thesis. Whether the higher number of Twitter users tweeting about Lyme disease is due to

increased awareness in the area or due to it reflecting actual cases in the area, or a mixture of the two is unknown. The reverse of this reasoning would explain the low-low areas.

Only seven (2%) local authority areas showed significant discordance between the location of Twitter users who tweeted about Lyme disease and laboratory-confirmed cases of Lyme disease. The areas that have high levels of cases and low level of users could have a low level of awareness of Lyme disease, or simply the Twitter user population is not representative of the general population in these areas. The areas that have a high level of Twitter users and a low level of cases are harder to explain. Reasons for users tweeting about Lyme disease in areas of low incidence are unknown and need to be explored. It could be due to an underreporting of cases, or due to prominent Lyme disease advocates being based in these areas. It must be noted that three of these areas neighbour the main high-high area in southern England. These areas are significant clusters because the incidence is high within their area and in the neighbouring areas. Therefore, these low-high areas in the south-west of England, are still likely to be representative of areas with a higher than average incidence of laboratory-confirmed Lyme disease. Content analysis of the tweets originating from discordant areas may be able to shed light on some of the reasons behind the discordance.

Incidence data at local authority level is unfortunately lacking for Scotland, Northern Ireland, and the Republic of Ireland. Its therefore unknown whether Twitter data matches the known epidemiology of Lyme disease in these areas. However, reports suggest that the incidence is high in the Highlands and Western Isles of Scotland [76] and is high in the west of the Republic of Ireland [268,281]. Better incidence data in these areas would be valuable, although Twitter data for the Republic of Ireland collected in this study was sparse, which limits further research using the methodology described here. The Twitter data has the potential to match the geographical distribution in Scotland and should be explored.

7.4.1 Limitations and future research

The largest limitation of the data presented here is the potential unreliability of the user location data. The search query utilised aimed to identify tweets sent from within a defined geographical area, however it appears that tweets were collected outside of this area. The algorithm for the collection of tweet location is not public, and it can be hypothesised that this is based on a combination of device/GPS co-ordinates (not shared through the API) and network/ (Internet Protocol) IP address location. If this assumption is correct, then all tweets collected should have been posted within the search area. It cannot be ruled out that an IP address is registered in the UK and the Republic of Ireland, but the user posted the tweet or

is based in a location outside of the search area. Only a small fraction of the study tweets were publicly location enabled, however all of these did have a location within the search area. The majority of the tweets locations were derived from the user profile stated location. There are obviously issues with this methodology as a person's location may not be always reliable. This can be as extreme as giving a falsified location, forgetting to update a profile location, or as in the author's case, using a place of work as their location rather than their place of residence. This could account for the number of countries outside of the study location identified as locations. The other alternative is that a Twitter user based outside the UK and the Republic of Ireland sent a tweet in the study area when they were visiting.

However, the fact that maps were produced that had significant similarity to the known epidemiology of Lyme disease suggest that unreliability may be less than initially perceived; given the analyses undertaken. By attempting to reduce bias by removing place names with 'Lyme' in them from the analysed datasets, bias may have been introduced by excluding any trends from these areas. This could especially be true for Lyme Regis, which is a town in an area of high incidence. Places that included 'Lyme' in their name may have also been included, despite the thorough search. This problem is potentially not unique just to Lyme disease. Other diseases with location specific names include, among many others; Rocky Mountain spotted fever, Ross River Fever, Omsk haemorrhagic fever, and Zika. The methodology developed to adapt to this issue is fairly rudimentary and a more robust methodology needs to be explored to reduce this problem.

Two separate studies show that Twitter users are more likely to be male and young adults (15-34) compared to the national population [257,259]. They also tend to be of the upper and middle class demographic, and occupy managerial, administrative, and professional roles. In previous social media health research, it was additionally found that the communities that drive messages and information on social media are often grouped in polarized subsets. Within each subset similar views are shared that reinforce their own beliefs, and act as echo-chambers [282]. The content shared may not stand up to scientific scrutiny, and therefore the spread of false information can be exacerbated [283]. As such, the degree of representation of this study population is uncertain. The views of each Twitter community, and identification of the 'experts', can only be explored through network analysis of Twitter users and qualitative analysis of the tweet content.

One of the assumptions of this research is that the Twitter users are being honest both about themselves and their commentary or information that they are sharing. As shown by the

fictional locales in the user location data, not all users of this platform are always honest, and some want to deceive. This may be for comedic purposes, as was likely in the case in the highlighted users, or for strategic enhancement of self-presentation, or for more malevolent purposes. A study evaluating deception in online dating found that 81% of users had deceptive information on their profile and 18.7% intentionally deceived, all with the aim of strategic enhancement of their self [284]. The author could not find literature discussing malevolent deception, however one could hypothesise that the perceived anonymity of users, and the relative lack of accountability provides a platform where the spread of false information is enabled. The retweet functionality can therefore perpetuate the spread of false information if a user is unaware of the credibility of the information contained within the original tweet. Without in-depth content analysis of the tweets the level of deception and the amount of false information contained within this dataset remains unknown.

This research has focused entirely on the spatial-temporal distribution of tweets relating to Lyme disease; however, the main reason for tweeting is to share information or views. Therefore, the majority of the data contained within a tweet has not been analysed. There is scope for further research on this dataset to explore content analysis through a thematic approach, and to perform a network analysis of the tweeters. This research could provide a deeper understanding of what is being publicly shared about Lyme disease and what type of individual or organisations are the main drivers of this information. This could be used to inform future public health campaigns for Lyme disease and other conditions.

7.5 Conclusions

Twitter has the potential to be used as an adjunct to traditional surveillance methods to understand people's concerns, and to identify seasonal trends and identify geographical activity hotspots in real time. Further analysis is needed to explore its robustness. Future research is also needed to analyse the content of the tweets, and to see whether they can utilised for public health messaging.

The extent to which this work contributes to the overall aims of this thesis regarding Lyme disease surveillance can be summarised as follows:

- I**ncidence**: Disease incidence cannot be defined from Twitter. There were, on average, 38 original tweets a day in the UK and the Republic of Ireland. There were signs of seasonality, with tweets peaking in summer months.
- **Sociodemographics**: No data could be garnered.

- **Geographical hotspots**: These were identified in south-west of England, and the Highlands of Scotland.
- **Patient presentation and management**: No data could be garnered.
- **Additional information:** Twitter offers some scope as a future 'measure of concern' surveillance tool; however, it can be easily biased by the users (both intentionally and unintentionally). Content and network analysis of the tweets is needed to understand who is tweeting and what is being discussed within the tweets. The potential of using this in conjunction with Lyme disease incidence and tick activity surveillance systems to inform health promotion campaigns merits further investigation.

Chapter 8 Companion animal electronic health data as a surveillance tool for tick activity

In the preceding chapter, Twitter data was explored as an adjunct for traditional surveillance systems and showed promise for matching temporal and geographical trends associated with Lyme disease. Another proxy data resource worth exploring is national data on the geographic distribution and temporality of ticks, in particular *Ixodes ricinus*. This data could provide information needed for a formal risk assessment of Lyme disease acquisition in the UK. The Small Animal Veterinary Surveillance Network (SAVSNET) has the potential to provide this information through the analysis of small animal primary care veterinary consultations as recorded in electronic health records. SAVSNET also collects laboratory diagnostic data but unfortunately this was not currently suitable for analysing tick-borne diseases, despite this data being greatly needed [280]. This chapter uses these companion animal primary care electronic health records to describe the incidence and geographic distribution of ticks found on companion animals. It's future potential as a Lyme disease risk resource will be discussed.

The following content was published as: Tulloch, J., McGinley, L., Sánchez-Vizcaíno, F., Medlock, J., & Radford, A. (2017). The passive surveillance of ticks using companion animal electronic health records. *Epidemiology and Infection, 145*(10), 2020-2029.

8.1 Introduction

Ticks are effective vectors of zoonotic pathogens, and tick-borne diseases (TBDs) can be severely debilitating to both humans and companion animals, in some cases leading to death. Lyme disease is the most common TBD in the Northern Hemisphere with, in Western Europe, an unweighted mean for annual incidence rate of 56.3/100,000 persons per year [66]. Tick borne diseases can pose a large burden on health services; a recent study of Lyme borreliosis inpatients in Germany estimated an annual cost in excess of 30 million Euros [194]. Due to this, and increasing public concern, governments and research organisations are trying to heighten and improve their understanding of risk models of ticks [285]. The responses to this call have largely fallen into three categories: the active and passive collection of ticks, the utilisation of digital applications, and the monitoring of electronic health records.

In Great Britain, Public Health England (PHE) coordinates the Tick Surveillance Scheme (TSS) [28,286], which relies on passive submission of ticks by members of the public as well as

medical / veterinary professionals. Between 2005 and 2016 the TSS received a total of 18,000 ticks, primarily found on companion animals and humans (PHE unpublished). Similarly, 'The Big Tick Project', collected data on ticks found on dogs in the United Kingdom (UK) [137,287]; in their most recent study, 6,555 ticks were actively collected from dogs attending select veterinary clinics over a 16 week period from April to July 2015 [137]. Since such systems collect the actual tick, they are able to both identify the ticks and describe their spatial distributions. However, they are labour and time intensive, relying on large amounts of public engagement and involvement and, in the case of the Big Tick Project, do not provide continuous surveillance data.

A very different approach has been developed by the National Institute for Public Health and the Environment (RIVM) in The Netherlands. The Tekenradar website and digital app allows members of the public to record when they have been bitten by a tick (and send it to RIVM), or develop an erythema migrans rash which is pathognomonic for Lyme disease [288,289]. This enables 'live' reporting of tick bites and identifies areas of tick bite and Lyme disease risk. Due to its presence on multiple digital platforms it facilitates easy promotion for public health messaging. It has also been promoted as a resource for researchers of ticks and TBDs [289]. However, the success of such a system is largely reliant on the accurate diagnosis and identification of ticks, tick bites and erythema migrans by members of the public, rather than qualified health care professionals.

In Switzerland, the government has set up a voluntary surveillance system of 150 primary care physicians called Sentinella [212], recording 1,644 cases of tick bites from 2008 to the end of 2011. Rather than collecting and submitting ticks for further analyses, this system relies on the accurate diagnosis and recording of tick bites by medical practitioners within their patients' electronic health records (EHRs), without the actual visualisation or collection of the tick. In a similar way, PHE use routine passive syndromic surveillance based on a predetermined list of clinical codes to monitor the incidence of arthropod bites in near real-time across various clinical settings including general practitioner consultations, emergency department attendance and telephone helplines [205]. However, constraints of the clinical diagnostic codes being used mean tick bites cannot be analysed separately from those of other arthropods.

While each of these systems contribute to different aspects of tick surveillance, none of them currently provide a surveillance system that is low cost and in sufficient temporal and spatial

resolution in near real-time to quickly and efficiently provide large sets of data about generic tick activity.

According to the most recent estimates, there are 11.6 million dogs and 10.1 million cats kept as pets in the UK, with 30% and 23% of households owning a dog and cat respectively [290]. These species have the potential for greater exposure to tick habitats than humans, and often without measures to prevent tick contact. It has been shown that dogs that are regularly walked are likely to acquire ticks, and it is well established that dogs have the potential to act as sentinels for ticks and TBDs [137–141]. Due to owner concern, companion animals with ticks are often presented to veterinary clinics, with the veterinary practitioner frequently recording the presence of ticks within an individual animal's electronic health record (EHR) [145]. The aim of this chapter is to explore the feasibility of using such EHRs from a large sentinel network of veterinary clinics as the basis of a novel surveillance system to provide efficient temporal and spatial estimates of tick activity risk in Great Britain that complement existing tick surveillance schemes.

8.2 Methods

EHRs were collected through the Small Animal Veterinary Surveillance Network (SAVSNET) from volunteer veterinary clinics using a compatible practice management system; currently Teleos ™ and RoboVet™. This study uses over two years of data gathered from 192 veterinary clinics across the UK between 31st March 2014 and 29th May 2016 (Fig. 8.1).

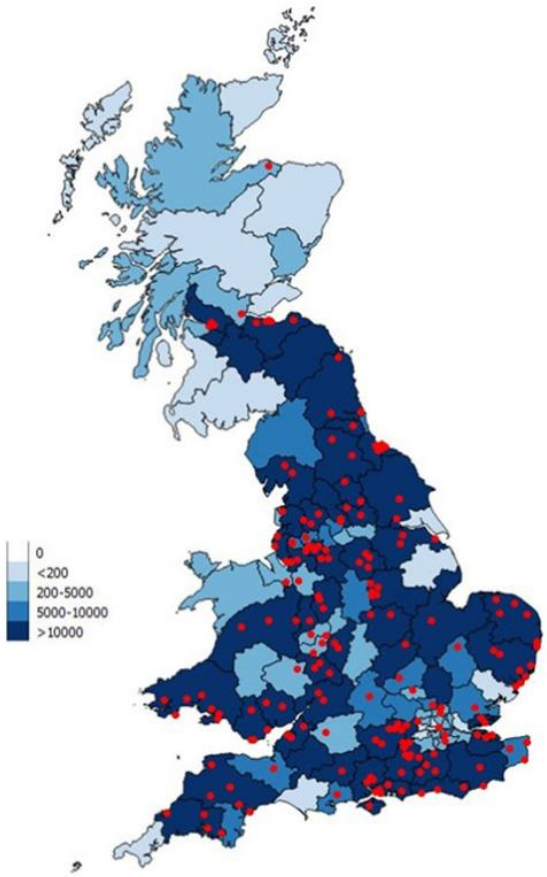


Figure 8.1 The distribution of participating SAVSNET veterinary clinics (red dots), and the total number of electronic health records collected between April 2014 and May 2016 by owners' postcode area.

Each EHR was collected at the end of a veterinary consultation in real-time and included the following data; date of the consultation, postcode of the owner, species of the animal in the consultation and the clinical narrative, which would have been written by the consulting veterinary surgeon or nurse. Whilst data on ecto-parasiticide treatments were collected, they were not included in further analysis; many are active against multiple arthropods and routine prophylactic prescription by veterinarians results in a low specificity for tick infestation.

Initially, a simple free text analysis approach was developed to identify EHRs containing the word 'tick' in the clinical narrative field, whilst excluding records where only the terms 'tickl', 'ticki' and 'sticki' were used. To increase the specificity of such an approach, the resulting EHRs were subsequently read by two domain experts (JT and LM) who verified the reference to ticks based on a strict case definition. A tick was only deemed present in a consultation if, within the associated EHR, 'a veterinary surgeon or nurse confirmed visual sighting or removal of a tick within the consultation.' This case definition was used to avoid any misidentification errors of ticks by owners, or any historical identification of ticks being

included in the current analysis (e.g. tick removed last week). To verify concordant interpretation of the case definition, the two domain experts manually classified a random sample of the EHRs. The amount of agreement (i.e. inter-rate reliability) was measured by Cohen's kappa coefficient (R package 'psych') [291]. Each EHR where an agreement was not achieved was re-examined by both domain experts to develop a more consistent interpretation of the case definition. This process was repeated until an 'almost perfect' level of agreement was achieved, at which point the remaining EHRs were randomly divided between the two domain experts and categorised [292].

Results of these analyses were used to calculate the number of consultations where a tick was recorded in the EHR per 10,000 consultations. This is assumed to be a proxy for activity of ticks and for brevity, refer to this measure as "tick activity". Relative risks were calculated between the two predominant host species (i.e. dogs and cats) with statistical significance ($P<0.05$) measured by a Chi-squared test.

Time-series plots (based on the time of the consultation as recorded in the EHR) were used to identify temporal trends in tick activity and to compare temporal trends by host species. The temporal pattern of tick activity was smoothed using a nonparametric method, the LOESS (locally weighted regression) technique (R package 'ggplot2') [293,294]. Outliers were identified as those outside the smoothed line's 95% confidence intervals. All proportions and 95% confidence intervals (CIs) were calculated using robust standard errors to account for intragroup correlation within veterinary clinics. Statistical analyses were carried out using R language (version 3.2.0) (R Core Team 2015).

Maps were used to describe the spatial distribution of tick activity during each season. Seasons were defined as winter (December-February), spring (March-May), summer (June-August) and autumn (September-November). The spatial distribution of tick activity was stratified by owner's given address. SAVSNET receives full owner postcode for each EHR, which locates each address to one of 1.75 million locations [295]. However, at such resolution it is possible to identify some individual properties, particularly in rural areas. Therefore, postcode area (first half of postcode; n=124) was used to maintain owner confidentiality when presenting the results. When displaying the data, a cautious approach was taken and areas with less than 200 EHRs in each season were excluded, as they were less likely to be representative. A map was constructed displaying all EHRs, aggregated by

owner's postcode area, to show the underlying population distribution (Fig. 8.1). The data was depicted using QGIS version 2.8.2-Wien.

8.3 Results

In total 1,658,857 EHRs were collected during the study period, consisting of 70.5% dogs and 26.4% cats. Of these, 10,155 (0.61%) had a clinical narrative containing the word 'tick'. The two domain experts first independently read and applied the case definition to 365 randomly selected EHRs from these 10,155. After adjusting by the amount of agreement which would be expected by chance, a 'substantial agreement' (*K*= 0.7; 95% CI: 0.63-0.78) with 305 EHRs agreed was achieved. After reappraising this first data set, the exercise was repeated on a new random sample of 365 EHRs, this time achieving an 'almost perfect' agreement (*K*= 0.82; 95%CI: 0.77-0.88) with 332 EHRs agreed; the remaining EHRs were therefore categorised independently by the two authors.

In total, 2,180 EHRs were confirmed as having a tick present, equating to 0.13% of the total 1,658,857, and 21.5% of the 10,155 automatically identified EHRs. Of these 2,180 EHRs, 1,421 were from dogs (65.2%), 728 from cats (33.4%), and 17 from other species (which only included ferrets, rabbits and guinea pigs; 0.8%), with the remaining 14 EHRs lacking an identifiable species label (0.6%). The relative risk of a dog being recorded as presenting with a tick compared to that for a cat was 0.73 (95% CI: 0.67-0.80, $p<0.005$). The main reasons for EHRs being identified by the free text analysis but failing to meet the case definition included; misidentification of ticks by owners (e.g. skin tags, nipples, tumours), ticks observed by owners before the consultation and not confirmed by a veterinary surgeon or nurse within the consultation, and discussions held in the consultation about ticks and TBDs without a tick being present. Only five of the 2,180 (0.2%) EHRs identified as relating to ticks included information at genus and species level; two referring to *Ixodes spp*, one to *I.ricinus*, one to *Dermacentor spp* and one EHR referring to both *Dermacentor and Rhipicephalus spp*.

The mean weekly rate of tick reporting in this population over the entire study period was 15.3 tick based EHRs per 10,000 EHRs. The temporal pattern was similar in both calendar years, with peak tick activity between May and July each year, and highest levels recorded in mid-June (Fig. 8.2).

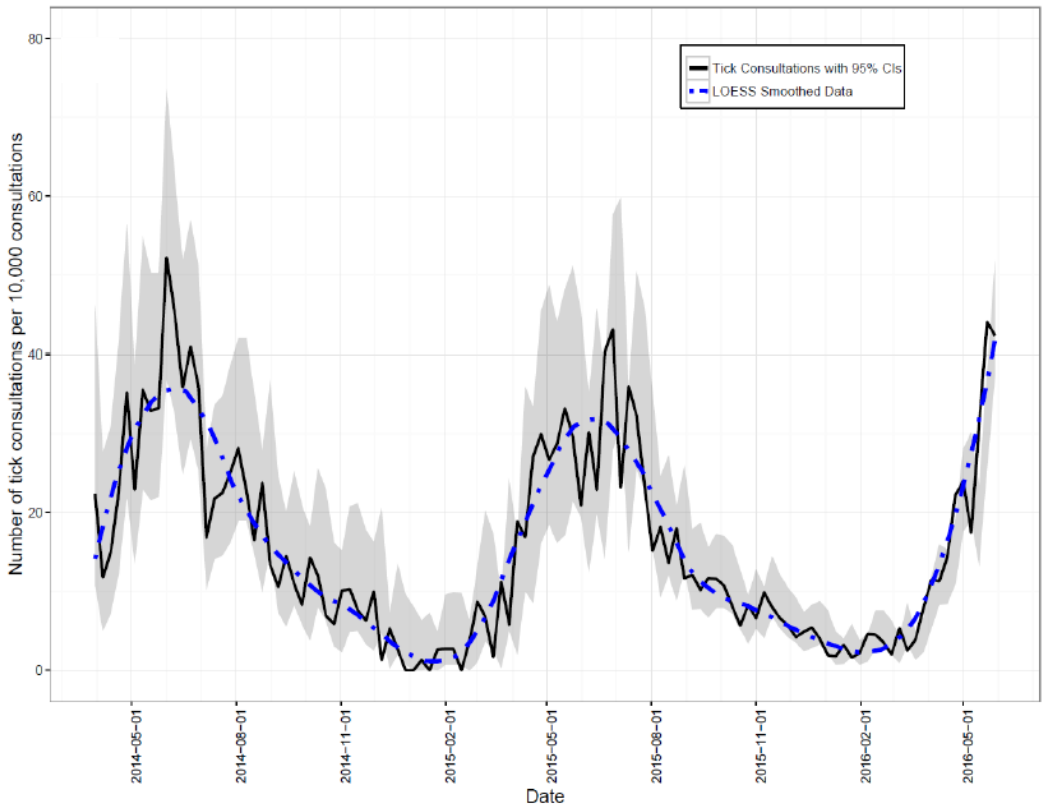


Figure 8.2 Time series plot showing the weekly number of tick based electronic health records (EHRs) per 10,000 EHRs between April 2014 and May 2016

Minimum tick activity was between December and February, with the lowest activity in January in both calendar years. The temporal pattern of tick activity in dogs was similar to the overall population, with peak activity in June (maximum of 65.5 tick based EHRs per 10,000 EHRs over a single week) and lowest levels between December and February (Fig. 8.3). In contrast, cats seemed to have an earlier peak in weekly tick activity in May (with a maximum of 87.2 tick based EHRs per 10,000 EHRs), with a secondary smaller peak in the autumn, and their lowest levels in February (Fig. 8.3). In the winter of 2015-2016, ticks were still recorded in every week on cats, whilst for two separate weeks none were recorded on dogs. The mean weekly rate of tick activity was lower for dogs (14.8 tick based EHRs per 10,000 EHRs) than for cats (18.3 tick based EHRs per 10,000 EHRs).

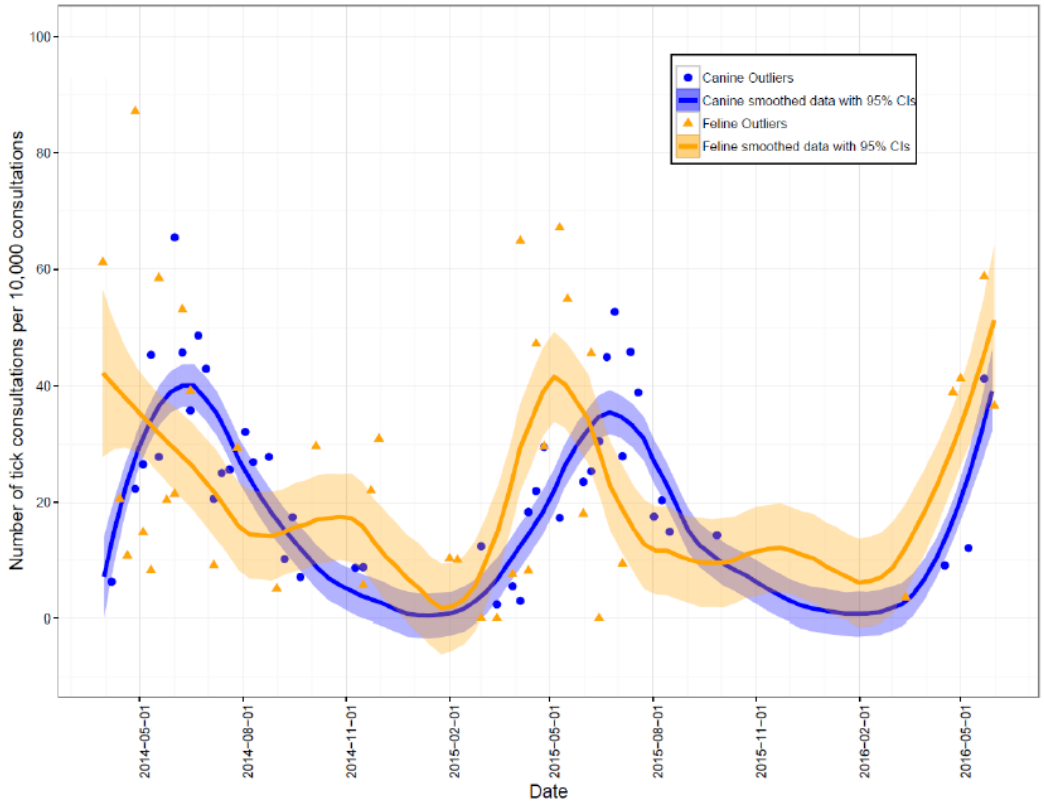


Figure 8.3 Time series plot showing the weekly number of tick based electronic health records (EHRs) per 10,000 EHRs in dogs and cats between April 2014 and May 2016

There was considerable variation in the spatial distribution of tick activity across each postcode area in Great Britain (Fig. 8.4). The ten postcode areas with highest tick activity across all seasons were (in descending numerical order); Bournemouth, Hemel Hempstead, Southampton, Falkirk, Salisbury, Guildford, Croydon, Llandudno, Reading and Lancaster. Of these, Southampton, Bournemouth, Guildford, Reading, Llandudno and Lancaster peaked in spring; the remainder peaking in summer. No postcode areas had their peak activity in autumn or winter. The areas with no tick activity across all seasons were Hereford, Oldham and Wolverhampton.

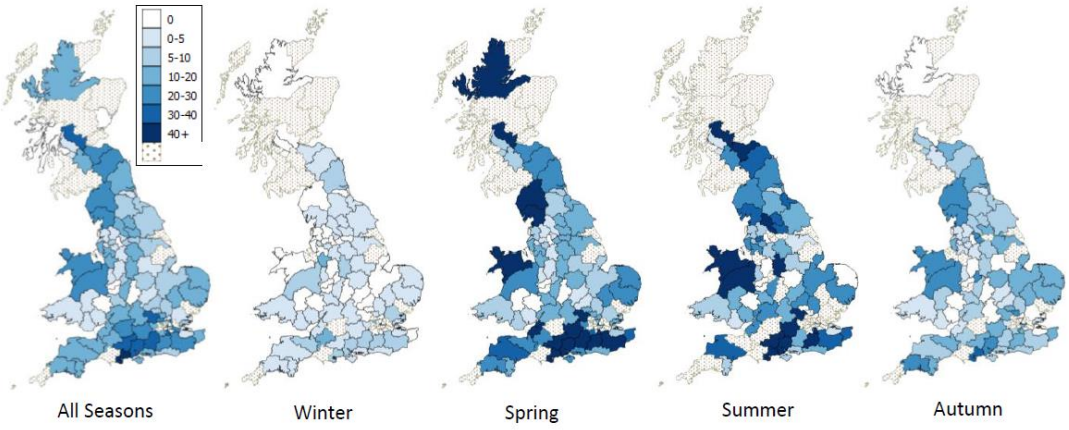


Figure 8.4 Geographical distribution of tick based electronic health records (EHRs) per 10,000 EHRs by owners' postcode area for each season between April 2014 and May 2016. The dotted postcode areas represent areas with less than 200 EHRs in total during the relevant time period.

8.4 Discussion

Ticks are an important vector of disease but continuous near real-time surveillance has proved challenging. Here it is shown how text mining of real-time companion animal EHRs from a large sentinel population of veterinary clinics may provide a novel form of passive surveillance to describe temporal and spatial trends in tick activity across Great Britain. This could provide an early warning system to the risk of veterinary disease and human health risks associated with ticks.

The SAVSNET data showed a seasonality to tick activity consistent with previous reports with peaks of activity at the start of the summer and minimal activity during winter [28,296–298]. Although identification of the life cycle stage of the ticks referred to within the EHRs was not possible, it was assumed this seasonality largely reflects that of adult ticks, and to a lesser extent nymphs, of *Ixodes ricinus,* as these are the most common ticks found on companion animals in Great Britain [28,137,287,297,299]. This tick is susceptible to desiccation and its host seeking behaviour (questing) is greatly influenced by changes in temperature and humidity [300]. It therefore has an annual variation of peak activity, with levels rising in early spring and peaking between April and July, with low levels in winter [297], although this may show regional variation [301]. The fact that the seasonal profile observed in dogs was similar to that in the total population, is a reflection of the demographic predominance of dogs in the data and in other such veterinary visiting populations [302,303].

Interestingly, tick activity on cats showed marked differences to that seen in dogs, raising several important questions relating both to ectoparasite biology, as well as owner and

veterinary surgeon behaviour. This study is the first to suggest that cats are more likely to present to veterinary clinics with ticks than dogs. Previous studies based on tick submissions either excluded cats [137,287] or lacked suitable population denominator data to calculate a relative risk [28]. Whether this represents a genuine increased risk of ticks on cats, or that ticks on cats are more likely to be observed by owners and presented to the veterinary surgery, or whether veterinary surgeons are more likely to record ticks on cats in their EHRs, remains to be determined. As well as this overall increased risk, cats continued to present with ticks during the winter of 2015-2016, in contrast to dogs where there were short periods where ticks were not identified in this population. During these months, dog owners may be less inclined to take their dogs for exercise where they may be exposed to ticks due to shortened day length, cooler temperatures and higher rainfall. However, domestic cats in the UK may remain susceptible to ticks, albeit at lower levels, due to their ability to explore outside habitats at their own free will due to the common use of cat flaps.

Ticks on cats also showed a different temporal pattern of tick activity with an earlier main peak in the spring and some evidence for a second smaller peak in the autumn. The precise reason for this apparent difference remains unknown but may relate to differences in host susceptibility to different tick species. Cats are significantly more likely to carry *Ixodes hexagonus* than *I. ricinus* [299] and *I. hexagonus* is more frequently found on cats than dogs [28,286,299]. The activity of *I. hexagonus* is closely linked to the density and behaviour of its primary host, the European hedgehog (*Erinaceus europaeus*) [304]. *I. hexagonus* is more prevalent earlier in the year than *I. ricinus* [304,305], coinciding with the emergence of hedgehogs from hibernation [306], and possibly explaining the earlier peak of tick activity identified in cats. The second autumnal peak could represent interaction between cats and hedgehogs at a time when hedgehogs are preparing for hibernation and juveniles are gaining independence, leading to greater hedgehog numbers being seen [306], all at a time when *I. hexagonus* is also at great abundance on the hedgehogs themselves [304,305].

The spatial distribution described generally mirrors previous work on tick distributions in Great Britain [28,137,286,287,296]. Comparing it to the most recent study published using data from a shorter, but overlapping time period (16 weeks between April – July 2015), both studies identified the highest levels of tick activity in southern postcode areas of England, with high levels also in the south of Scotland [137]. However, in contrast, higher levels of activity were seen in north and mid-Wales, and north-west England, and less clear areas of high activity in north Norfolk and the north-east of England. These observations are likely to

reflect differences in methodology used by the two projects including veterinary clinic recruitment, the period of sampling, and potentially tick distribution [137].

Collection of continuous surveillance data over two years across GB has allowed us to begin to describe a complex mosaic of tick activity across the country in different seasons. However, broad trends can be identified. In winter, low levels of tick activity remain throughout England and Wales, challenging the belief of some vets, who recorded in their EHRs that ticks pose no risk in winter (unpublished observations). The results also showed that the timing of peak activity varied by postcode area, with the majority of areas peaking in the spring, the remainder peaking in the summer. The data set described here represents a rich research tool in which to explore the varied impact of climate, and other environmental and ecological factors, on tick activity.

To maintain a high specificity, a very restrictive case definition was applied, only including ticks that were seen by a veterinary surgeon or nurse and recorded during the consultation. Therefore, it is clear that not all ticks on cats and dogs will be included in this study. Many ticks on companion animals will not present to the veterinary practice either because the owner is not concerned, or removed the tick themselves, or the ticks were not noticed. Equally ticks on animals in a veterinary consultation may not be noticed, or not recorded, especially where they are incidental findings in relation to what may be a more serious clinical need. Indeed, where dogs had a bespoke thorough clinical examination as part of a research study to identify tick carriage, reported tick prevalence was much higher (30%) [137]. This study was however carried out during peak tick activity (April-July) and as the authors stated, practitioners participating in the study may have been more likely to sample animals with observed ticks on them. Although it is clear that the values reported are therefore an underestimate of overall tick activity on companion animals, one is confident that they can describe relevant levels of relative risk. It must be acknowledged that health scares and media coverage could influence owner behaviour and veterinary recording behaviour. This has been previously discussed in relation to the *Babesia canis* outbreak seen in early 2016 [145]. However, in this particular case, this outbreak did not appear to influence the overall temporal trends of this data (data not presented).

Arrival of exotic ticks has been of great concern to both the veterinary and medical professions as they have the potential to carry pathogens not currently transmitted in the UK [175,307–309]. This has driven a need for species level surveillance of ticks such as

provided by PHE [286] and the Big Tick project [137]. Within this data, only five EHRs included information at the genus and species level; two referring to *Ixodes spp*, one to *I. ricinus*, one to *Dermacentor spp* and one EHR referring to both *Dermacentor and Rhipicephalus spp.* Although these numbers are clearly low and in the absence of microscopic confirmation need to be treated with some caution, they still raise important questions. Whilst a few foci of *Dermacentor* are known to exist in the UK [286], *Rhipicephalus sanguineus* has only been reported in dogs that have travelled in the rest of Europe [137,309], such that reference to *Rhipicephalus spp* in even one EHR could be significant. The infrequent mention of tick species likely reflect time constraints of a short consultation, the challenge of identification, especially if the tick is engorged, and veterinary surgeons deeming it clinically irrelevant. In the future, the reference to rare and exotic tick species identified by EHR surveillance, could be followed up by submission of the tick to relevant health authorities with tick identification capabilities; such as PHE. Surveillance systems based on EHRs would be improved if veterinary surgeons were encouraged to record within the EHR any recent travel history and information about tick species where they are confident to do so.

The limitations of this study are inherent to its methodology. Since recruitment of practices is not random, there may be selection bias in the results, meaning generalisability to the entire UK population of veterinary visiting dogs and cats is not possible. In addition, population statistics for companion animals in the UK are generally poor or unavailable, such that the results cannot be described as incidence; this may change as compulsory microchipping of dogs has recently come into legislation [310]. Some postcode areas have relatively small amounts of data and were excluded from analysis. However, the fact that 56% (70 of 124) of postcode areas contributed more than 5,000 EHRs during the study period, and that 42% (52 out of 124) of areas contributed more than 10,000 EHRs suggests that there is already good data coverage for large parts of Great Britain. As SAVSNET continues to expand through clinic recruitment, it is believed that the spatial distribution of clinics and the number of EHRs collected will become more homogenous. SAVSNET data will always underestimate true tick activity on companion animals, and veterinary surgeons or nurses rarely record the tick species in EHR. In addition, like other studies that define a tick's location by the pet owner's postcode [137], the results should be seen as a proxy for tick activity at a given geographical area, rather than the location where the animal necessarily acquired the tick.

Despite these limitations, it is believed that this form of surveillance offers some real benefits. Using EHRs is very passive in nature, as once a veterinary clinic has been enrolled, no changes in clinician behaviour need occur for the data to be captured. This data is collected in real-time, with the only rate-limiting step currently being the time taken to verify the strict case definition. Compared to systems that rely on the general public identifying a tick, ticks recorded in EHRs are identified by a qualified health care professional [289]. There is also minimal labour required, except the upkeep of a system to collect the EHRs on which it relies. As the results are similar to previous surveillance and field work performed in the UK, it is believed that they do provide a novel and complementary approach for tick surveillance that could be adopted by other countries where mature pet animal EHRs exist. As more clinics are recruited the representativeness of such systems can be improved. Linking data through postcode to other data sources will provide new opportunities to understand the effect of climate change and land use changes on the distribution and activity levels of ticks [26,27,175,307,311].

8.5 Conclusions

In summary, this study shows how the passive real-time collection of companion animal EHRs can provide efficient, accurate and novel data on tick activity in a large national sentinel population of companion animals. We highlight for the first time temporal differences of tick exposure between domesticated cats and dogs. As the availability of EHRs increases, such methodology can provide a comprehensive temporal and spatial understanding of tick activity, and in combination with other systems already in place, we believe can further inform real-time tick and TBD risk models, aiding a 'One Health' approach for public health messaging and tick control.

The extent to which this work contributes to the overall aims of this thesis regarding Lyme disease surveillance can be summarised as follows:

- I**ncidence**: Not applicable. However, seasonal trends were seen in tick activity, with peaks in summer months.
- **Sociodemographics**: Not applicable. However, cats and dogs were identified as the main species that could be used in future sentinel surveillance systems
- **Geographical hotspots**: These were identified in southern England and north Wales.
- **Patient presentation and management**: Not applicable.
- **Additional information:** SAVSNET has the potential to act as a sentinel surveillance system for tick risk. However, this needs to be formally assessed (Chapter 9). The

laboratory data contained within SAVSNET needs to be assessed, as companion animal TBD surveillance data is severely lacking [280]. SAVSNET's laboratory data for zoonotic infections, like Lyme disease, could be a more useful sentinel surveillance system than tick activity data.

Chapter 9 A comparison of datasets

9.1 Introduction

Over the last six chapters, various health datasets, and the information they could add to Lyme disease surveillance in the United Kingdom, have been discussed. For each dataset the following questions, as set out in Chapter 1, were considered:

- For each health dataset appraised (see Chapters 3-8);
    - What is the incidence of Lyme disease?
    - What are the sociodemographics of the patient population?
    - Can any geographical hotspots be identified?
    - Can any data about patient presentation and management be extracted?
    - Is there any additional information about the epidemiology and management of cases which is unique to this dataset?

This leaves the two following research questions unanswered:

- Using the current laboratory-confirmed based surveillance system (RIPL) as a reference point, how does each of the other datasets compare? How complete is each dataset, and is there a stable multiplication factor that can be applied to RIPL, that can provide an improved overall annual incidence estimate?
- Based on the datasets analysed, what policy (or policies) described by EPPI should the public health authorities of the United Kingdom adopt? (Chapter 10)

Within this chapter datasets are compared, in terms of demographics, spatial, and temporal patterns, and investigated as to whether stable multiplication factors exist between the datasets.

9.2 Methods

The Zoonoses report [153], part of which was analysed as the RIPL dataset in Chapter 3, represents the British Government's official laboratory-confirmed case numbers for Lyme disease in the UK. This shall be the reference dataset when comparing official UK incidence figures to that within THIN. The data in the Zoonoses report is not stratified by any sociodemographic or geographic variables, and only reports new national case counts. For the remaining comparisons the RIPL dataset will be used as the reference dataset and will therefore focus on English and Welsh data. It must be noted that SGSS data has been excluded from analysis, as it was not a reliable dataset for Lyme disease (Chapter 3). Likewise,

only HES and PEDW admissions data will be used as a comparator as HES A&E and outpatients data were deemed too sparse and unreliable (Chapter 4).

9.2.1 Methods – Demographic comparisons

Only three datasets had demographic data available for comparison; namely, RIPL, HES and THIN. Neither SAVSNET nor Twitter contained information about patient demographics.

To visually compare the structure of the Lyme disease case populations captured within each dataset, population pyramids were constructed based on case count data and compared to the national population pyramid [159]. The sex ratio for each dataset was calculated for 5-year age-bands and overall. Chi^2 tests were performed for each age-band, and overall, to compare the sex ratio of each dataset with those in RIPL reference dataset.

Measures of ethnicity, rural-urban status and societal deprivation were compared descriptively. Where a dominator population was not available for analysis, these factors were compared to 2015 national population demographics as provided by the ONS [159].

9.2.2 Methods – Spatial comparisons

Four datasets had geographical data with a fine enough geographical resolution to perform comparative spatial analysis; these were the RIPL, HES, Twitter and SAVSNET datasets. THIN contained no degree of geographical resolution, and as discussed in Chapter 3, SGSS has been removed from further analysis. For all datasets, annual information was sparse; data was therefore aggregated and analysed at an average annual incidence per geographical unit. Only geographical data for England and Wales could be analysed, as the RIPL dataset only provides information for these two countries. Local authority was chosen as the geographical unit for analysis, as this offered finer resolution than postcode area, without producing large numbers of areas with zero incidence. Comparisons between Twitter and RIPL geographical data have already been discussed in Chapter 7. The remaining datasets to be compared were RIPL, HES and SAVSNET (Fig. 9.1).

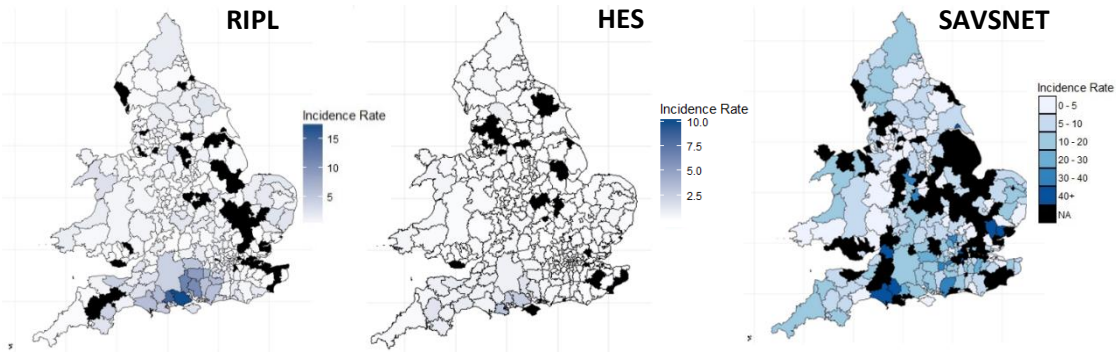


Figure 9.1 Geographical distribution of Lyme disease incidence (RIPL, HES), and ticks found on companion animals (SAVSNET). Areas with no positive cases are shaded black.

The geographical analysis of HES and THIN data has so far, in this thesis, focused on descriptive mapping of their incidence data. The RIPL data was explored further through exploratory spatial data analysis (ESDA) [164,165] using global Moran's I value and LISA (Local Indicators of Spatial Association) analysis (Chapter 3). These methodologies were used to identify global and local outliers, identify global trends, examine local variation and examine spatial autocorrelation. By performing this analysis, further insight into the geographical differences in the incidence of laboratory-confirmed Lyme disease were gained. ESDA methodology can be adapted to compare how multiple variables are distributed over the same space [164,165]. They can therefore be used to explore geographical associations between the laboratory-confirmed incidence of Lyme disease (RIPL) and hospitals admissions data (HES) or tick activity on companion animals (SAVSNET). The two methodologies used were the calculation of a bivariate global Moran's I value and the calculation of local indicator of spatial association (LISA) values for each local authority.

The bivariate global Moran's I value describes the direction and strength of a relationship between two variables (e.g. RIPL Lyme disease incidence and HES Lyme disease incidence) in each geographical unit and measures the overall level of clustering. As the degree of spatial autocorrelation in a dataset is not always globally uniform, local statistics were critical. LISA provides this, by identifying locations of spatial clusters and outlier data.

The significance of these statistics was tested by comparing them to a reference distribution created by 2000 random permutations of the underlying spatial reference distribution [164,165]. A Queen's contiguity was used to define spatial neighbours; this defines a neighbouring area as one with shared borders and vertexes. The global Moran's I value can be visualised in a bivariate LISA plot, which shows the correlation between one variable (for example, RIPL incidence) and a different variable (HES incidence) in a local authority and

neighbouring areas. If the slope of the plot significantly differs from zero, there is a relationship between the two variables. Information from the scatter plot was used to create a cluster map to show spatial areas with significant concordant or discordant relationships. The local authorities with significant spatial correlation were identified by type of spatial autocorrelation and mirror the quadrants shown in the Moran scatter plot. Essentially the high-high and low-low areas represent positive spatial autocorrelation, and so spatial clusters, while the high-low and low-high areas represent negative spatial autocorrelation and are spatial outliers. A cluster is defined as an area with a value more similar to its neighbours than would be the case under spatial randomness. As the cluster map displays only the core of the cluster, the cluster is likely to extend to neighbouring locations.

This methodology was used to explore the associations between the laboratory-confirmed incidence of Lyme disease and the incidence of hospital admission, and the incidence of ticks found in companion animal electronic health records.

9.2.3 Methods – Temporal comparisons

To compare case counts and disease incidence, datasets need to include figures for the same time period. For UK data, the Zoonoses Report will act as the reference dataset and it covers the years 2007 to 2016 inclusive. Therefore, comparator datasets must provide case counts and incidence figures for the same period. The only dataset matching these criteria was THIN (Chapter 5). Since THIN only provides age-standardised incidence rates, case counts have been extrapolated using the ONS population figures [159]. The UK's case counts, and incidence figures, were first compared between the Zoonoses Report and THIN dataset.

To formally assess the similarities between the datasets a linear regression model for each dataset, and a combined model, was constructed. Slopes were compared by examining the interaction of year by dataset using an ANOVA. The residual variance of the two datasets was compared utilising a Fisher's R-to-z transformation of the Pearson correlations of the two datasets.

English and Welsh annual count and incidence figures were compared using the same methodology for the years 2007 to 2016. The datasets analysed were: the Zoonoses Report, HES and PEDW admissions, and THIN.

Monthly incidence data were only available for comparison for the period 2013-2015, for the RIPL, HES and THIN datasets. Comparisons were only possible for English data, as there was no Northern Irish or Scottish data, bar THIN, and Welsh data was too sparse to perform

analytics. The increased number of data time points meant that time series analysis was possible. Time series analysis provides an understanding of the underlying forces and structure that produce the observed incidence data. By decomposing a time series, an enhanced understanding of these structures can be gained. The trend component reflects the long-term progression of the time series, this can be increasing or decreasing and can be non-linear. The seasonal component reflects any seasonal variation, in this case over the twelve months of a year. The random component represents the residuals of the time series once trend and seasonal components have been removed. This should be stationary and random, to show that factors influencing the time series have been accounted for by the other components. By performing time series analysis of each Lyme disease dataset, and analysing each component separately, the similarity of the underlying processes defining the time series can be compared. If they are similar, then it could be concluded that each dataset's incidence data is driven by the same underlying processes and therefore reflective of similar disease and reporting dynamics.

Time series analysis was performed by decomposing the time series of the RIPL, HES and THIN data, using a Loess smoothing approach [312], to their trend, seasonal and random components. These components were compared visually and distance measures calculated to provide a formal assessment of the similarities between the time series [313,314]. The combined seasonal and random components were described, to visually identify the relative importance of the trend component on the data. To assess whether the random component of each dataset contained any further trends an augmented Dickey-Fuller (ADF) test was performed [315]. This tested the null hypothesis that the random component of each time series was not stationary.

The distance similarity measures chosen were Euclidean distance and dynamic time-warping (DTW) distance. The Euclidean distance is one of the most commonly used measures and provides a figure based upon the straight-line distance between points in two time series at the same fixed point in time. This works well when there is limited lag between the datasets. However, if a temporal lag is present, the distance can become unrepresentative despite the shape of the time series being similar. To correct for this a DTW distance was also calculated. This is calculated in a similar manner to the Euclidean distance, but it takes any lag into account by optimally aligning the points in the time series (Fig. 9.2) [314,316].

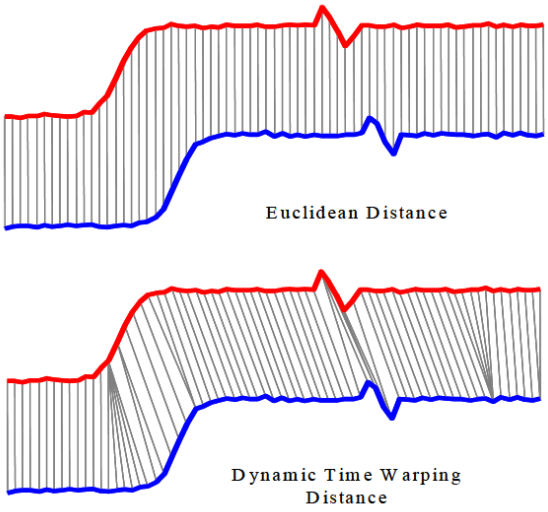


Figure 9.2 The different basic assumptions to calculate Euclidean distance and dynamic time warping distance. Permission to use the image was kindly provided by its creator, Professor Eamonn Keogh [303].

To reduce computational time calculating the DTW distance, the data was constrained using a Sakoe-Chiba band [313,314,316]. The accepted bandwidth of 10% of n (i.e. 10% of the 36 months analysed) was chosen.

However, as the aim of this comparison was to compare the shape of each components' distributions, a problem of data magnitude would arise. This was due to the incidence being markedly different in each dataset, for example the incidence of Lyme disease in the datasets in 2015 was, 1.91 cases per 100,000 in RIPL, 0.45 cases per 100,000 in HES, and 5.23 cases per 100,000 in THIN. Therefore, data transformation was needed to allow comparison of the time series' shapes.

The three traditional methods of data transformation (log, square-root, and arcsine transformations) were not appropriate as none of these methodologies deal with issues of differences in magnitude between datasets. A new magnitude transformation was created, which aimed to maintain relative proportions of the confidence intervals and mean of each curve. Effectively the ratio between a comparison dataset's upper confidence interval of the curve and the RIPL dataset's confidence interval of the curve, was used as a multiplier to the comparison curve.

$$new\ curve = \left\langle \frac{\{mean(original\ HES\ or\ THIN\ curve) + 1.96 \times SD(original\ HES\ or\ THIN\ curve)\}}{\{mean(RIPL\ curve) + 1.96 \times SD(RIPL\ curve)\}} \right\rangle \times (original\ HES\ or\ THIN\ curve)$$

This allowed all curves to be framed within the same upper and lower limits whilst maintaining the relative shape of the curves and keeping the temporal points fixed. In summary the time series were first transformed, then decomposed, and finally Euclidean and dynamic time-warping (DTW) distances were calculated.

To assess all the datasets, some of which could not be included in the above analysis, each dataset's monthly data was aggregated. With this data, a plot was created displaying the percentage of cases, out of the entire year, occurring in each month for each dataset.

9.2.4 Methods – Dataset completeness

Despite the proliferation of research based upon electronic health records, no standardised definition or process has been described to assess dataset completeness [317]. As Weiskopf et al importantly note;

> *'Completeness is contextual and is determined through an understanding of specific data needs. The number of complete records available for analysis is dependent upon the definition of completeness being used.*[317]'

The simplest and easiest measure, to calculate and understand, was adopted [318,319]. For each dataset variable, the percentage of total records present that could be used for analysis was calculated. An overall score was given, equaling the mean of each respective dataset's percentages.

9.3 Results

9.3.1 Results – Demographic comparisons

Population pyramids of each dataset and the ONS national population for England and Wales were constructed (Fig. 9.3)

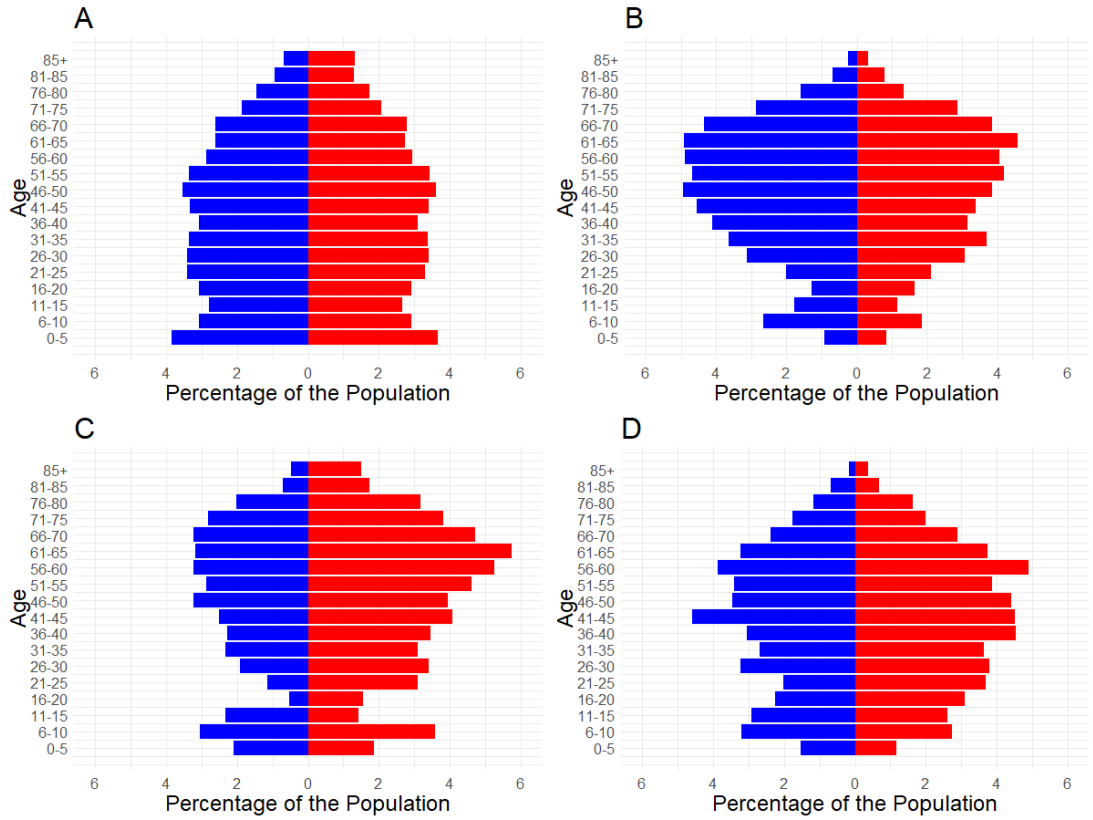


Figure 9.3 Population Pyramids for England and Wales. A = ONS population official figures 2015. B = Laboratory-confirmed cases of Lyme disease (RIPL, 2013-2016). C = hospital admissions of Lyme disease (HES & PEDW, 1998-2015). D = Lyme disease cases presenting in primary care (THIN, 1998-2016). Blue = Male, Red = Female.

The ONS pyramid is evenly balanced between sexes and its most populous age-band is 0-5 years. The pyramid is relatively stable from 6 to 70 years old, at which point the population starts to decrease. In contrast the RIPL pyramid is noticeably skewed towards males, although its general shape is symmetrical and bimodal with peaks at 6-10 and 46-65 years old. The HES data matches this distribution, except that it is skewed towards females rather than males and has a more pronounced bimodal distribution. The THIN pyramid is less obviously bimodal, with potential peaks at 6-10 and 56-60 years old, and as described in Chapter 5, no overall predominance in sex. However, its shape is more similar to RIPL and HES than to the ONS data.

The sex ratio for each dataset was plotted (Fig. 9.4) and compared using a Chi^2 test (Table 9.1).

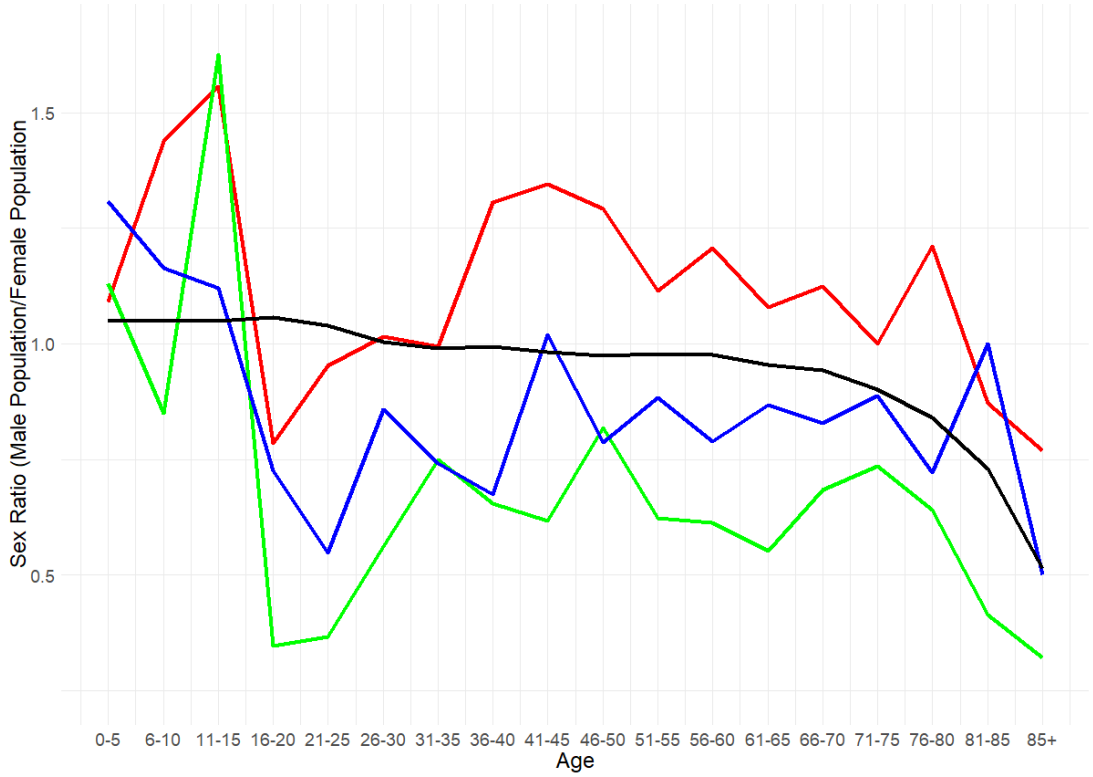


Figure 9.4 The sex ratio for each dataset over varying age bands. (Red = RIPL, Green = HES, Blue = THIN, Black = ONS data)

Table 9.1 The sex ratio for each dataset by age band. When a Chi² showed a significant difference to the corresponding RIPL sex ratio an asterisk is present *<0.1, **<0.05, *<0.01.**

Age Group	RIPL (Reference Value)	HES & PEDW	THIN	England and Wales (ONS)
0-5	1.09	1.13	1.31	1.05
6-10	1.44	0.85**	1.16	1.05**
11-15	1.56	1.63	1.12	1.05**
16-20	0.78	0.35*	0.72	1.06
21-25	0.95	0.37***	0.55**	1.04
26-30	1.02	0.56**	0.86	1.00
31-35	0.99	0.75	0.74	0.99
36-40	1.31	0.66***	0.67***	0.99**
41-45	1.35	0.62***	1.02	0.98***
46-50	1.29	0.82**	0.79**	0.97**
51-55	1.12	0.62***	0.88	0.98
56-60	1.21	0.61***	0.79**	0.98*
61-65	1.08	0.55***	0.87	0.95
66-70	1.13	0.68**	0.83	0.94
71-75	1.00	0.73	0.89	0.90
76-80	1.21	0.64**	0.72	0.84*
81-85	0.87	0.41	1.00	0.73
85+	0.77	0.32	0.50	0.51
Overall	1.14	0.66***	0.84***	0.97***

The overall sex ratio in the HES, THIN and ONS datasets all differed to that of RIPL. RIPL had a male bias, whilst the others displayed a female bias.

The two age groups that show agreement between datasets were the 11-15 year olds and those between 16 and 35, even though the proportions differ somewhat between the datasets. The 11-15 year olds were mainly male at a ratio that does not differ between datasets but were significantly different to the national population. Between 16 and 35 there is a predominance of females in the Lyme datasets, whilst the ONS has an even sex ratio.

In all datasets, age groups past the age of 36 show stability in their sex ratios, however differences between datasets were present. The HES and THIN datasets, in the main, have a female bias with a relatively stable ratio, until after 75 years old, when they become increasingly more female. During the 36-75 age range the ONS sex ratio also remains stable, with a slight female predominance, and then sharply becomes more skewed towards female at around 75 years old. In contrast the RIPL population has a male bias until 81 when this switches to female.

Due to the structure of the datasets assessed, only two datasets had ethnicity data, HES and THIN. The lack of denominator ethnicity data meant that both these datasets were compared to the ONS 2015 census population, in which 86% of the English and Welsh population identified with being white. In the HES and PEDW datasets, there was a significant difference to the ONS with 96% identifying with a white ethnicity (Chapter 4). Likewise, the English and Welsh population in the THIN dataset had significantly greater proportions of white ethnicity; 94% compared to 86% (Chapter 5).

In all three datasets there was a significant relationship between Lyme disease status and measures of deprivation. The RIPL and HES datasets were compared to the census figures of the Index of Multiple Deprivation, a measure of societal deprivation. In both these datasets, as deprivation increased, so the number of Lyme disease cases decreased (Chapters 3, 4). THIN provided a material measure of deprivation, the Townsend score. In this database a significantly higher incidence of Lyme disease coded cases was found in the least deprived areas.

In the RIPL and HES datasets a significantly higher proportion of cases lived in rural areas compared to the national population (Chapters 3,4). In THIN dataset there was roughly a twofold significant increase in incidence between Lyme disease coded cases found in rural areas than in urban areas (Chapter 5).

9.3.2 Results – Spatial comparison

The global Moran's I value for RIPL and HES incidence was 0.47 ($p<0.001$), indicating an overall significant positive spatial correlation between the two datasets (Fig. 9.5).

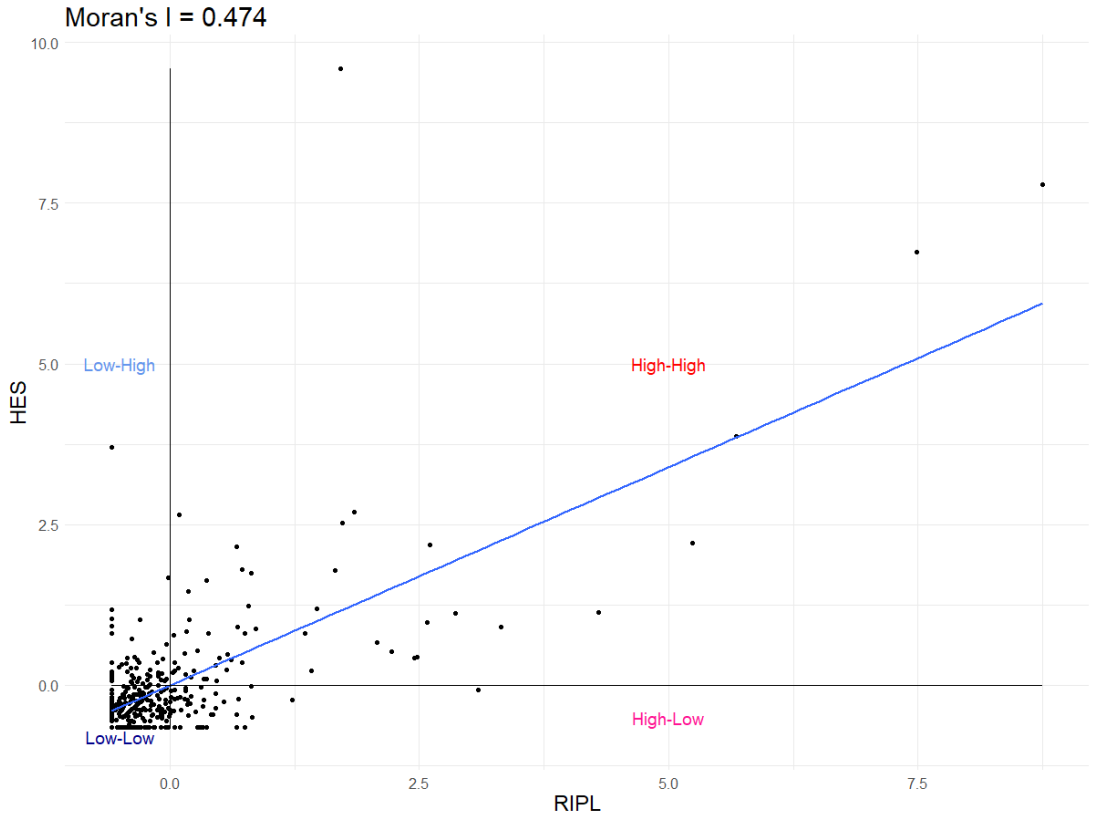


Figure 9.5 Bivariate Moran I's scatterplot between RIPL and HES Lyme disease incidence at local authority level.

The bivariate scatterplot shows a significant positive linear relationship in which local authorities (and their neighbours) with a higher laboratory-confirmed incidence tend to contain hospitals with a higher incidence of Lyme disease admissions, and vice-versa.

This overall positive correlation was seen on the LISA cluster map (Fig. 9.6)

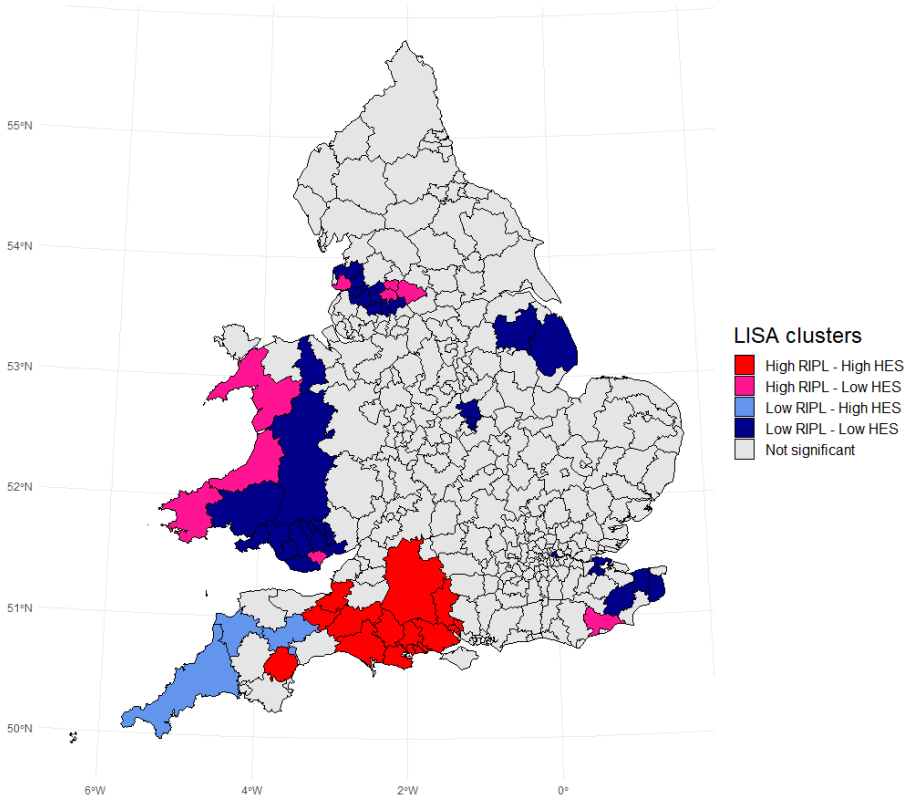


Figure 9.6 Bivariate LISA cluster map between RIPL and HES Lyme disease datasets.

The majority of the map is covered in non-significant local authorities, this shows that in these localities the case incidence for both datasets does not differ from a randomly generated distribution of incidence. The high-high areas represent concordant significant clusters of high incidence. The low-low areas represent the reverse and are areas where a significantly low incidence of disease is recorded in both datasets.

Only thirteen (4%) local authority areas show significant discordance. The areas that have a high level in RIPL and a low level in HES were scattered throughout the country. The areas with a low level in RIPL and high level in HES were all located in neighbouring local authorities in the south-west of England.

The global Moran's I value for RIPL and SAVNSET incidence was 0.05 (p=0.08), indicating a lack of spatial correlation (Fig. 9.7).

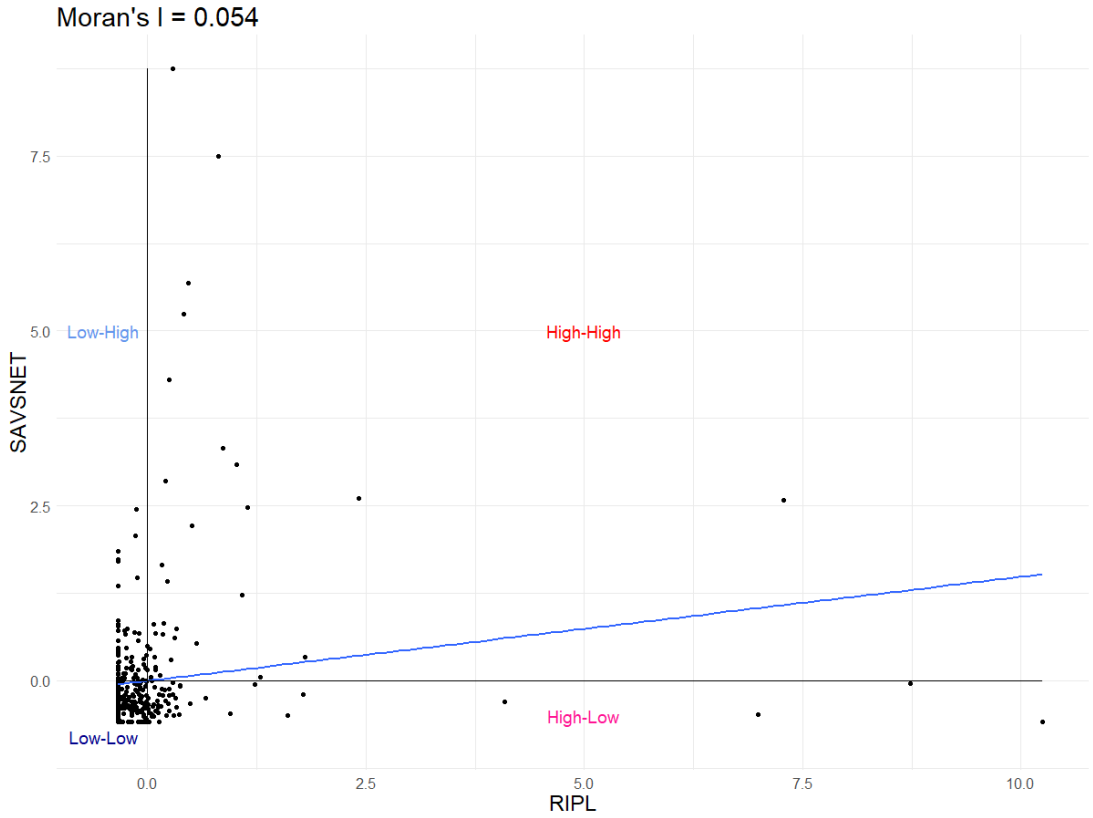


Figure 9.7 Bivariate Moran I's scatterplot between RIPL Lyme disease incidence and SAVSNET tick activity at local authority level.

The bivariate scatterplot shows a small positive non-significant relationship, which indicates that there is no spatial relationship between the laboratory-confirmed incidence of Lyme disease and tick activity based on companion animal electronic health records. As no significant spatial relationship was identified a bivariate LISA cluster map was not constructed.

9.3.3 Results – Temporal comparison

Case counts and the incidence of Lyme disease cases in the UK based on the Zoonoses report and THIN dataset were calculated (Fig. 9.8).

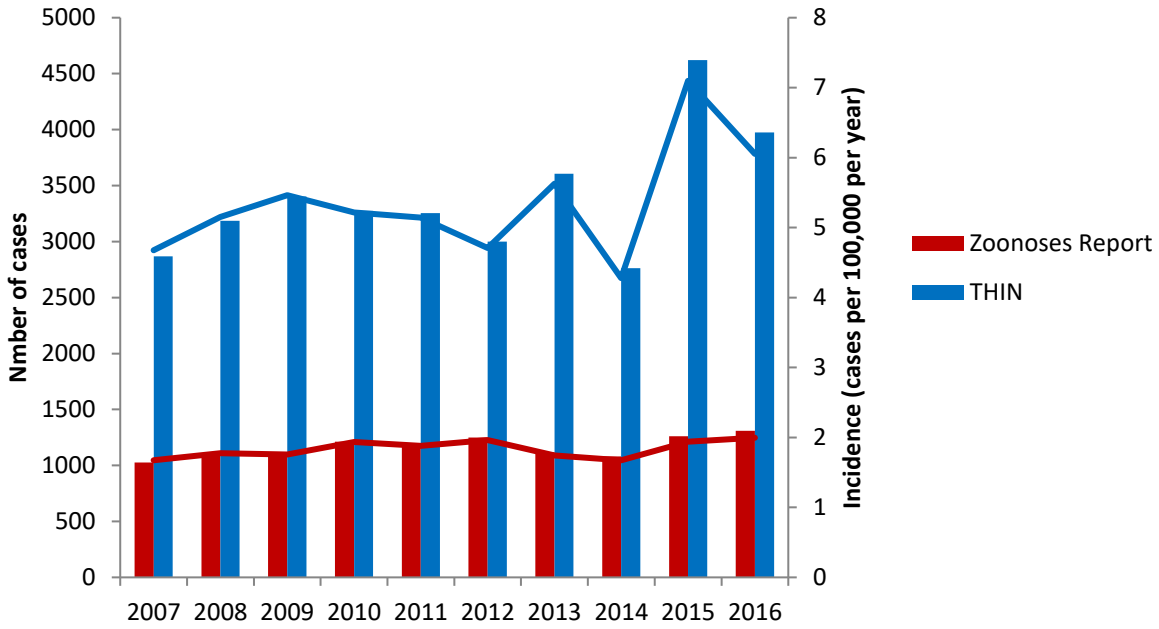


Figure 9.8 Case count (bars) and incidence (lines) of Lyme disease in the United Kingdom based on laboratory-confirmed cases (Zoonoses Report) and cases identified in a primary care database (THIN).

On visual inspection both datasets show an increasing trend of cases and disease incidence, with the fluctuations in numbers apparently mirroring each other. The relative ratio of incidence between the datasets also appears to be stable (Table 9.2).

Table 9.2 The ratio of annual incidence between primary care Lyme diseases cases (THIN) and laboratory-confirmed Lyme disease cases (Zoonoses Report), in the UK.

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	Mean Ratio
THIN Incidence	4.67	5.15	5.46	5.22	5.14	4.71	5.63	4.23	7.09	6.05	N/A
Zoonoses Report Incidence	1.67	1.78	1.76	1.93	1.88	1.96	1.74	1.67	1.94	2.00	N/A
THIN:Zoo Ratio	2.80	2.89	3.10	2.70	2.74	2.40	3.23	2.56	3.66	3.03	**2.91** (2.20-3.62)

There was no significant interaction in the relationships of incidence to year for the Zoonoses Report (β=0.02), nor THIN dataset (β=0.13); $F(1,16)=1.68$, $p=0.21$. The Fisher's R-to-z comparison indicated that there was no significant difference ($p=0.97$) of the Pearson correlation between the Zoonoses Report ($r=0.47$) and the THIN data ($r=0.48$), (Fig. 9.9).

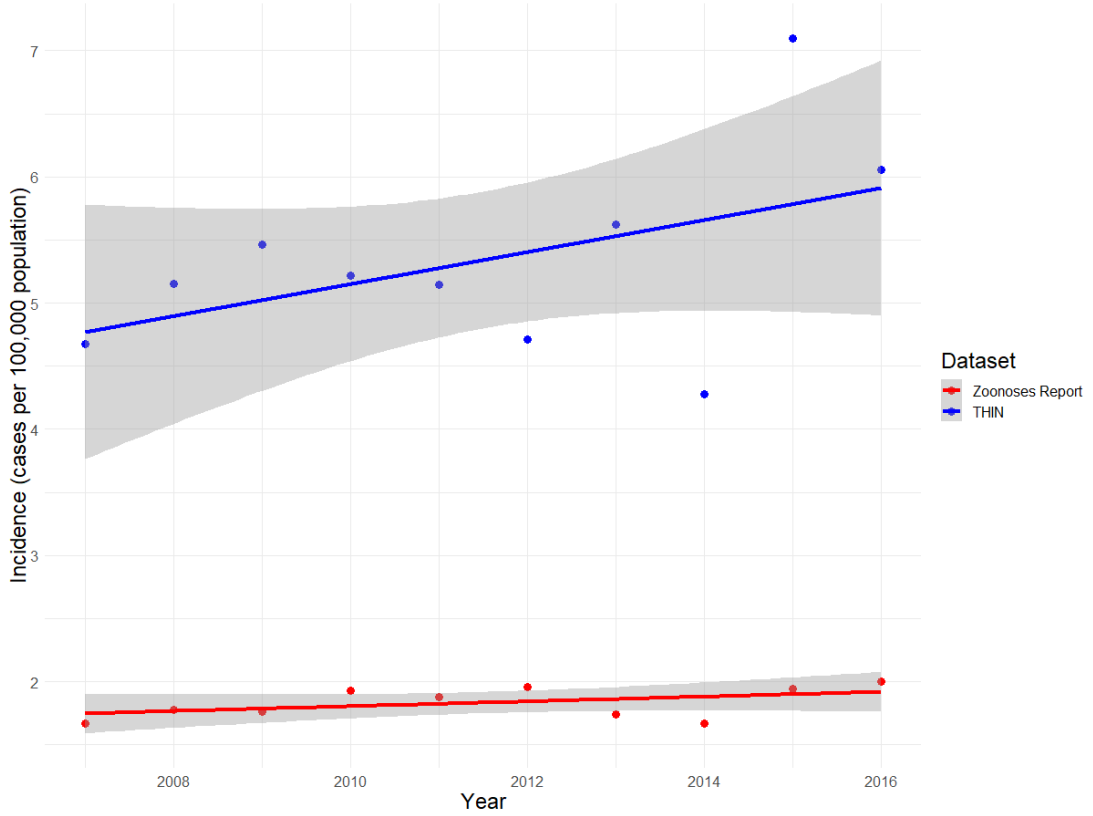


Figure 9.9 The incidence of laboratory-confirmed and primary care diagnosed cases of Lyme disease in the United Kingdom, with associated linear regression lines and 95% confidence intervals; 2007-2016.

Case counts and the incidence of Lyme disease cases in England and Wales based on the Zoonoses report, HES admissions and THIN were calculated (Fig. 9.10).

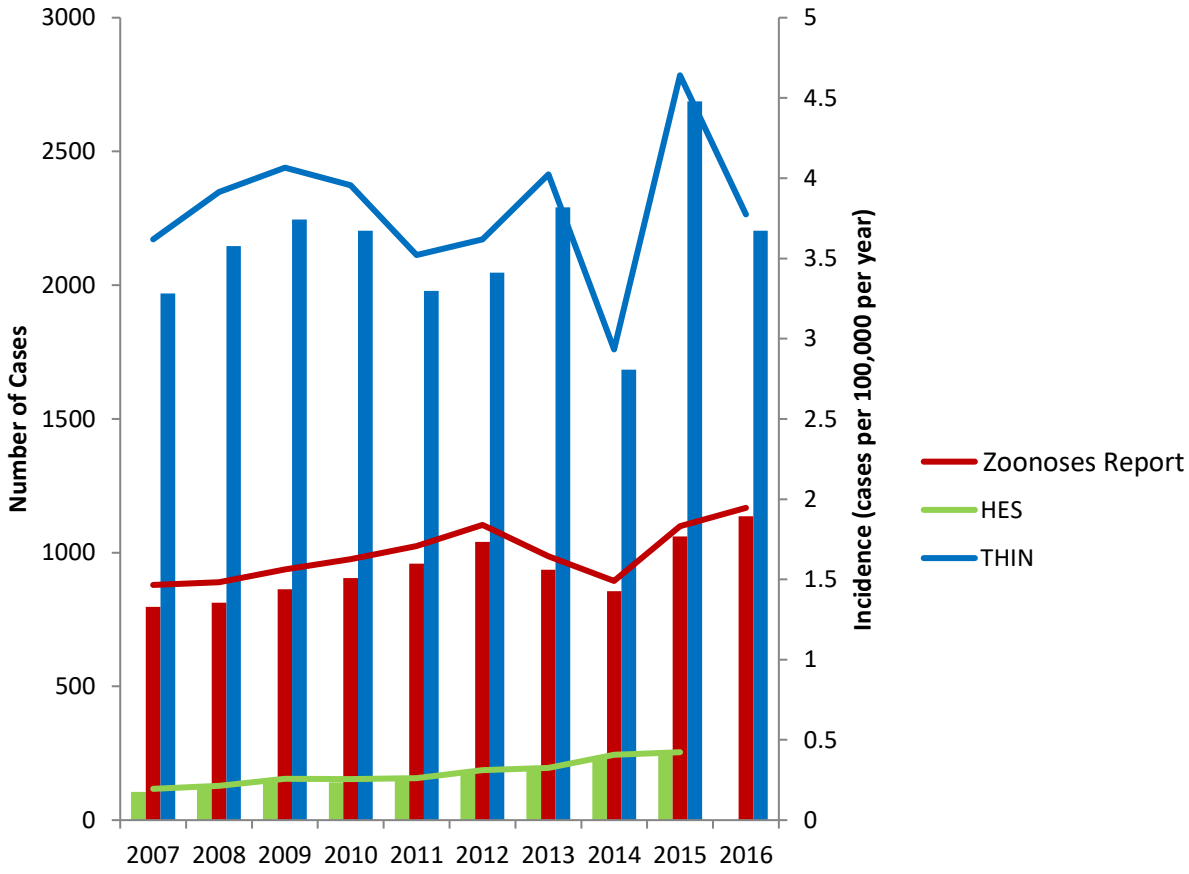


Figure 9.10 Case count (bars) and incidence (lines) of Lyme disease in England and Wales based on laboratory confirmed cases (Zoonoses Report), hospital admissions (HES and PEDW), and cases identified in a primary care database (THIN). NB. No hospital data for 2016 were available.

On visual inspection all three datasets showed an increasing trend of cases and disease incidence, with the fluctuations in numbers apparently mirroring each other. The relative ratio of incidence between the datasets also appears to be stable (Table 9.3).

Table 9.3 The ratio of annual incidence between primary care Lyme diseases cases (THIN), hospital admissions (HES), and laboratory-confirmed Lyme disease cases (Zoonoses Report), in England and Wales.

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	Mean Ratio
THIN Incidence	3.62	3.91	4.07	3.96	3.52	3.62	4.02	2.93	4.64	N/A
HES Incidence	0.20	0.23	0.27	0.26	0.27	0.33	0.35	0.43	0.44	N/A
Zoonoses Report Incidence	1.47	1.48	1.56	1.62	1.71	1.84	1.64	1.49	1.83	N/A
HES:Zoo Ratio	0.14	0.16	0.17	0.16	0.16	0.18	0.21	0.29	0.24	**0.19** (0.09-0.29)
THIN:Zoo Ratio	2.46	2.64	2.61	2.44	2.06	1.97	2.45	1.97	2.54	**2.35** (1.81-2.88)

There was no significant interaction in the relationships of incidence to year for the Zoonoses Report (β=0.03), nor HES (β=0.03), nor THIN dataset (β=0.01); F(2,21)=0.08, p=0.93. The Fisher's R-to-z comparison indicated that there was no significant difference (p=0.1) of the Pearson correlation for the Zoonoses Report (r=0.59), and the THIN data (r=0.07). However, there was a significant difference (p<0.01) for the Zoonoses Report (r=0.59), and the HES data (r=0.96) (Fig.9.11).

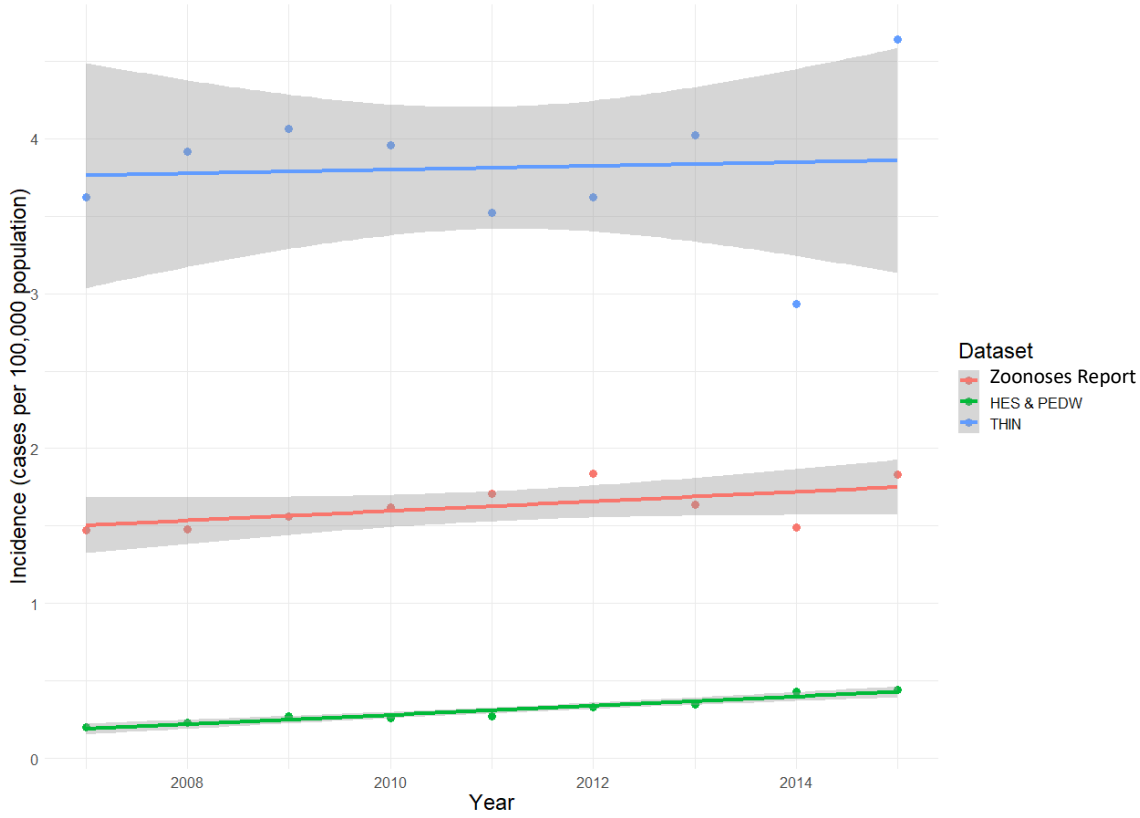


Figure 9.11 The incidence of laboratory-confirmed, hospital admissions and primary care diagnosed cases of Lyme disease in the England and Wales, with associated linear regression lines and 95% confidence intervals; 2007-2015.

Time series decomposition of the 2013-2015 English monthly incidence data for RIPL, HES, and THIN, provided in three decomposed components; seasonal, trend and random components. The data transformation, described above in section 9.23, minimised differences in magnitude and enabled visual and formal comparisons, using Euclidean distance and DTW distance. The raw monthly incidence and transformed incidence data were visually compared (Fig. 9.12).

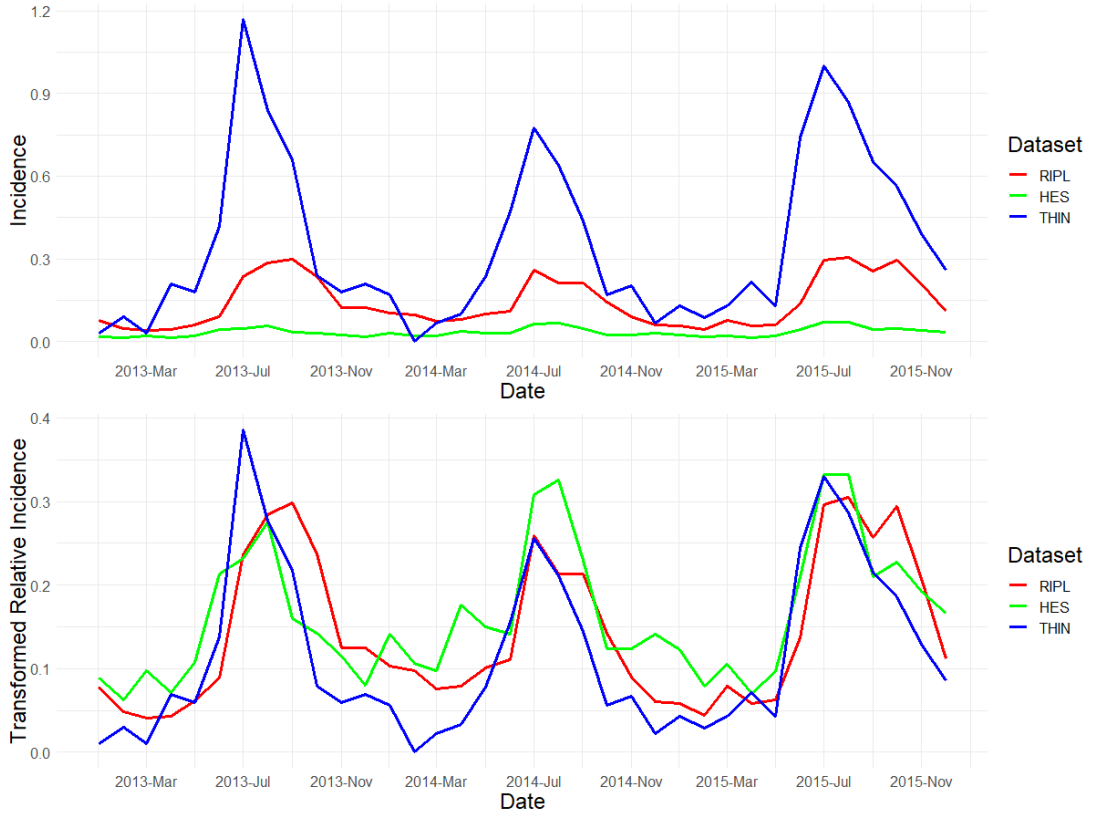


Figure 9.12 The monthly incidence of Lyme disease cases in three health datasets, and below the post-magnitude transformation of the datasets

This shows that the method of transformation was able to reduce the differences in magnitude, without compromising the relative shapes of each dataset's time series. Seasonality was already apparent prior to decomposition, with peaks in all three datasets during the summer months.

The trend component of each dataset time series comparison (Fig. 9.13), had an Euclidean distance between RIPL and HES of 0.10, and a DTW distance of 0.55, and an Euclidean distance between RIPL and THIN of 0.09, and a DTW distance of 0.46.

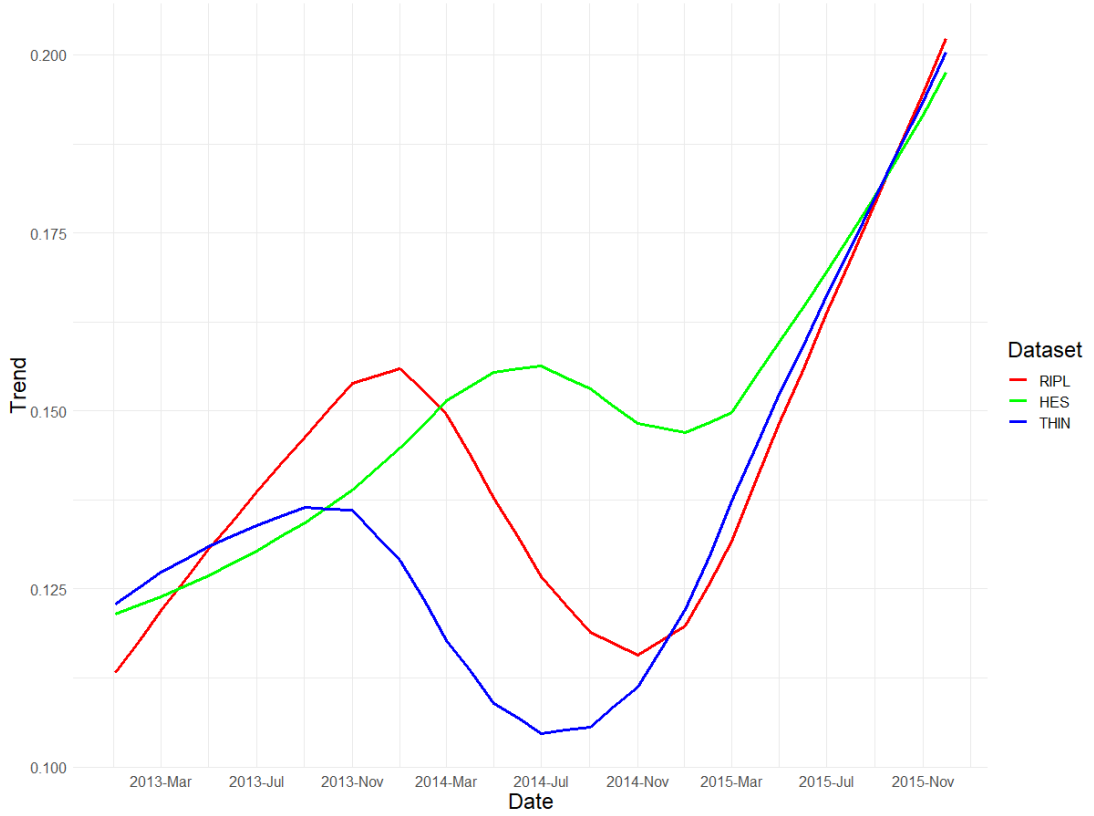


Figure 9.13 Trend component of the decomposed time series analysis of Lyme disease health datasets relating to laboratory-confirmed cases, hospital admissions and primary care; 2013-2015

Overall all three datasets show an increasing trend over the study period. The Euclidean distances between datasets was similar, however, the DTW distance is quite different. Visually RIPL, THIN and HES share a similar wave shape but with differences in lag and amplitude of the wave. From July 2015 onwards, all datasets showed an almost identical trend.

To explore the impact of trend on the time series, trend was removed from the overall time series leaving a component made up of the seasonal and random time series components (Fig. 9.14).

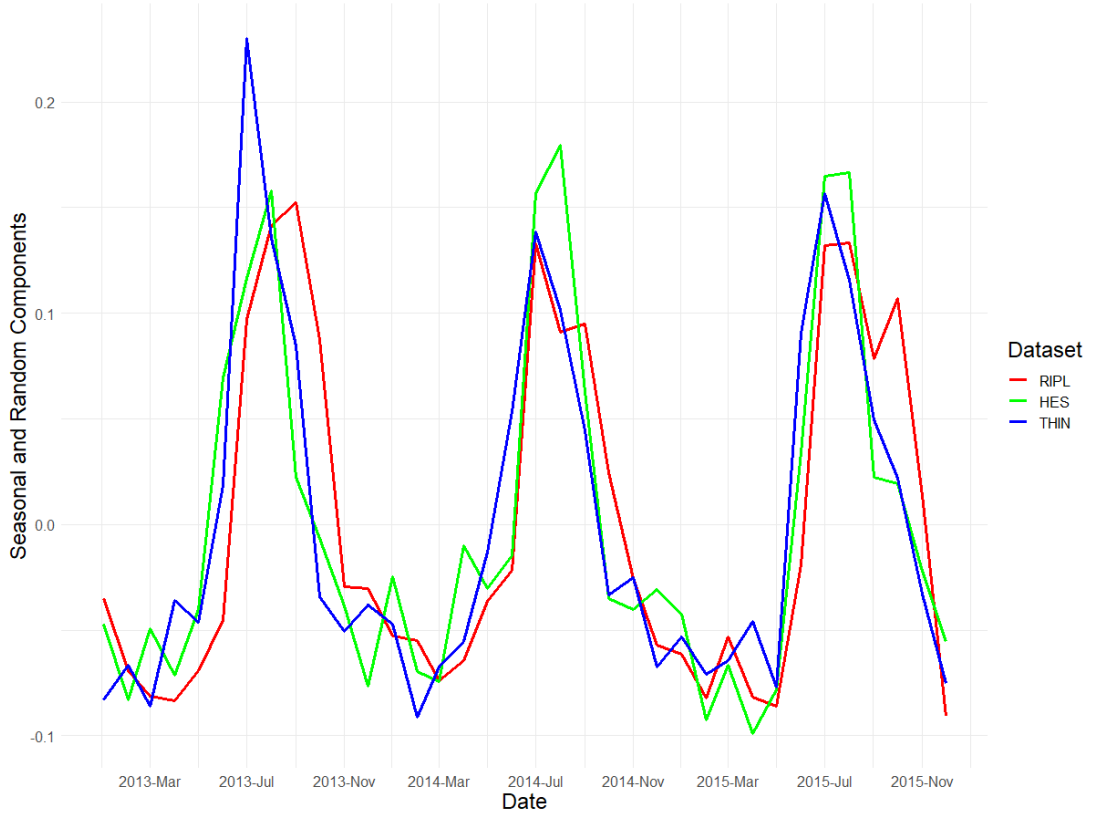


Figure 9.14 The combined seasonal and random time series components of three healthcare datasets for Lyme disease.

By eliminating the trend component from the time series, the seasonality of the datasets appear to match each other more closely than before and follow a similar pattern each year. This indicates that the overall trend has only a small impact on the monthly incidence of Lyme disease.

The seasonal component displayed similar distributions and similar distances (Fig. 9.15). The Euclidean distance between RIPL and HES was 0.24, and the DTW distance was 1.32, and the Euclidean distance between RIPL and THIN was 0.26, and the DTW distance of 1.08.

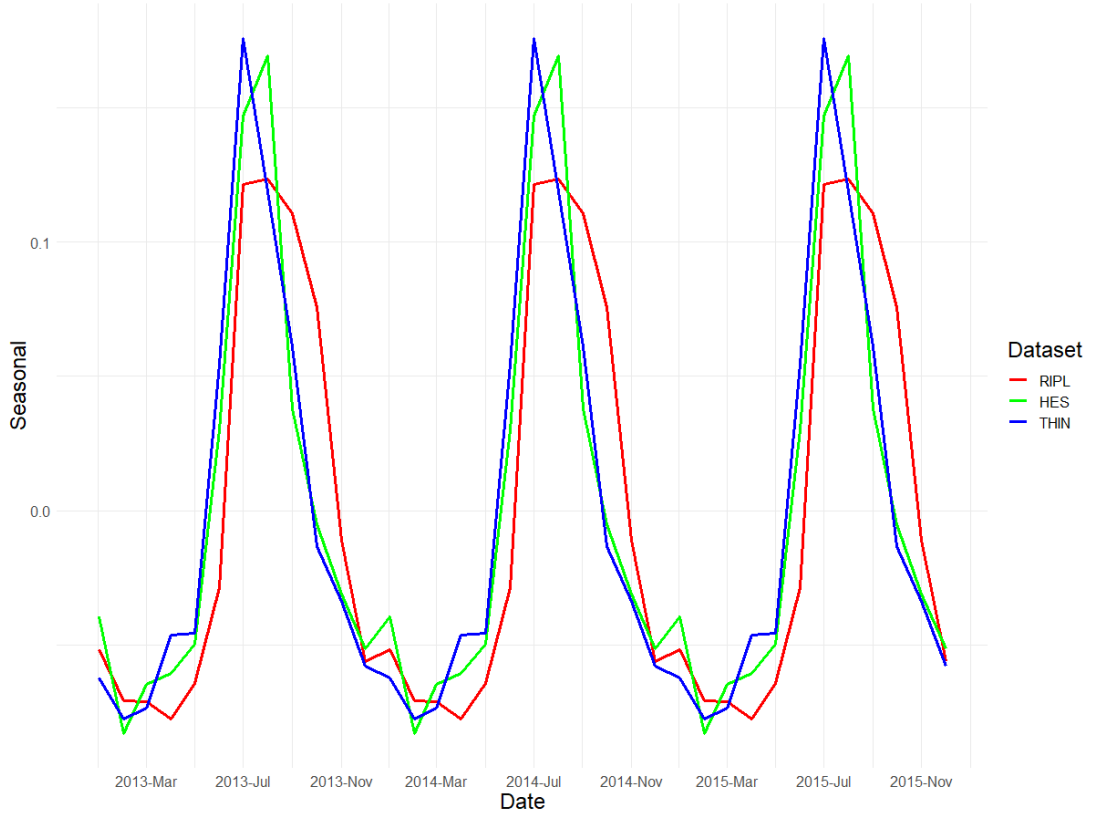


Figure 9.15 The repeated seasonal component of the decomposed time series analysis of Lyme disease health datasets relating to laboratory-confirmed cases, hospital admissions and primary care; 2013-2015

The seasonality seen indicates that cases tend to peak in the THIN dataset in July, followed by hospital admissions and laboratory-confirmed cases in August.

The random component of each dataset's time series comparison (Fig. 9.16) had an Euclidean distance between RIPL and HES of 0.16, and a DTW distance of 0.73, and an Euclidean distance between RIPL and THIN of 0.14, and a DTW distance of 0.73.

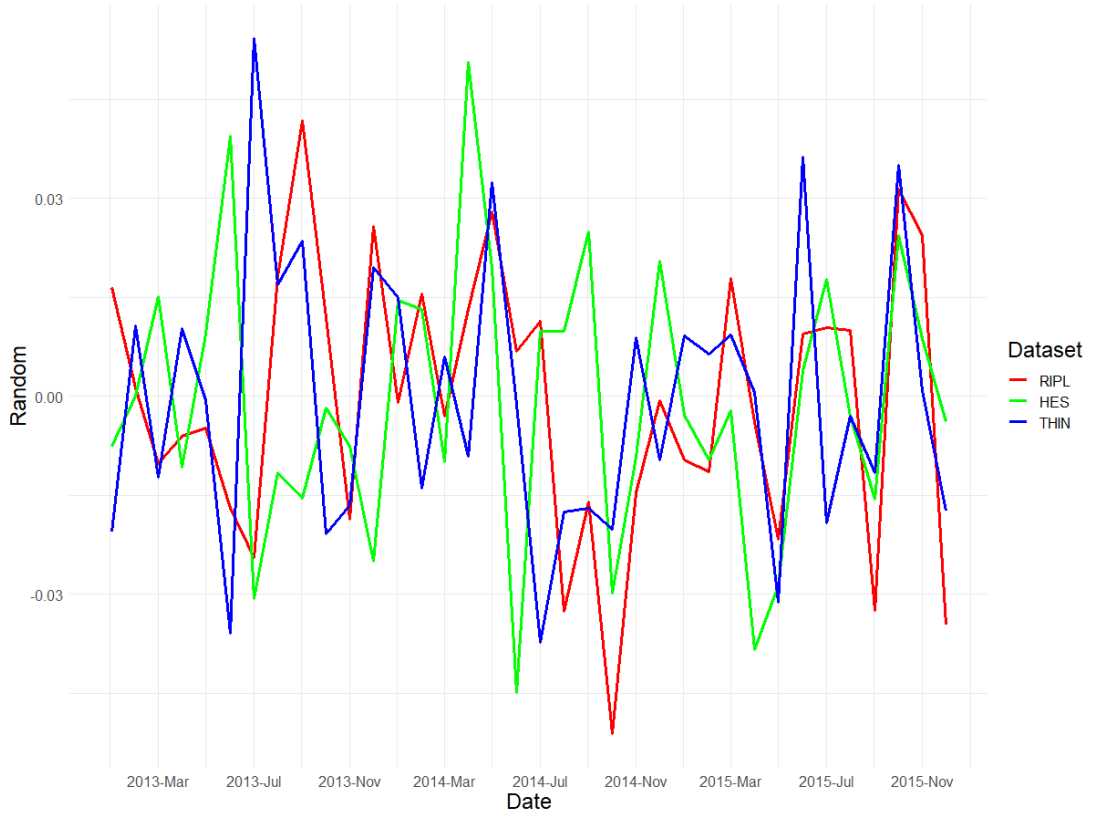


Figure 9.16 Random component of the decomposed time series analysis of Lyme disease health datasets relating to laboratory-confirmed cases, hospital admissions and primary care; 2013-2015

The similarity in distances, both for Euclidean distance and DTW, between the datasets show that they have comparable random components of their time series. The ADF test for the RIPL dataset, showed that the random component was not stationary (Dickey-Fuller test statistic=-2.22, p=0.49). The ADF test for the HES dataset, showed that the random component was not stationary (Dickey-Fuller test statistic=-2.80, p=0.26). The ADF test for the THIN dataset, showed that the random component was not stationary (Dickey-Fuller test statistic=-3.27, p=0.09).

The percentage of cases, out of the all aggregated years, occurring in each month for each dataset was described (Fig. 9.17)

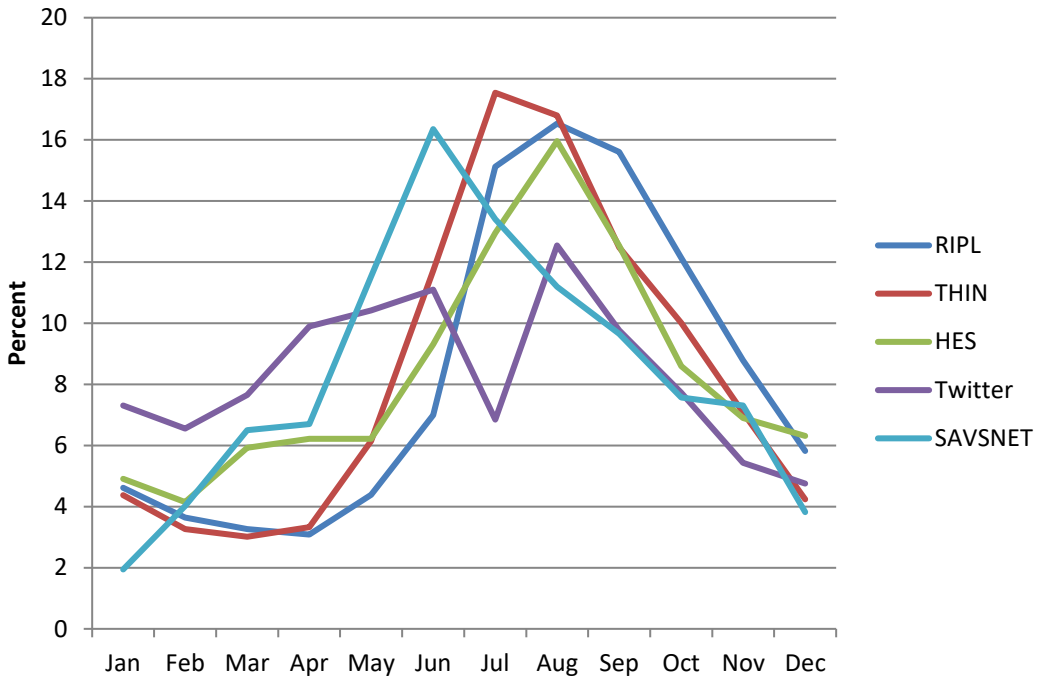


Figure 9.17 The percent of Lyme disease cases within each dataset by month

For all the datasets described, the months with the highest percentage of cases were in the summer. SAVSNET in June, THIN in July, and RIPL, HES, and Twitter in August. The lowest percentages were in winter and spring, Twitter in December, SAVSNET in January, HES in February, THIN in March, and RIPL in April.

9.3.4 Results – Data completeness

The completeness of the main variables under analysis for each dataset was calculated, with an overall completeness score calculated (Table 9.4).

Table 9.4 Summary table and heatmap of the degree of completeness of variables analysed in various Lyme disease health datasets.

Completeness Variable	RIPL	HES	THIN (England and Wales)	Twitter: Original Tweets Only	SAVSNET
Number of records	3,986	2,361	2,220	13,757	2,180
Date	98.7% (3,935)	100% (2,361)	100% (2,220)	100% (13,757)	100% (2,180)
Sex	98.7% (3,935)	70.9% (1,673)	100% (2,220)	N/A	N/A
Age	98.7% (3,935)	70.9% (1,673)	100% (2,220)	N/A	N/A
Postcode area/ Local Authority	58.3% (2,321)	88.0% (2,078)	N/A	52.0% (2,709)	100% (2,180)
Deprivation	56.6% (2,257)	96.5% (2,278)	94.5% (2,106)	N/A	N/A
Ethnicity	N/A	79.5% (1,877)	24.3% (540)	N/A	N/A
Rural Urban Status	56.6% (2,257)	97.1% (2,292)	93.4% (2,073)	N/A	N/A
Overall (An average of the above scores)	66.8%	86.1%	73.2%	76.0%	100%

There is a range in completeness between the datasets. RIPL is the least complete (66.8%), whilst SAVSNET is the most complete (100%). HES is the only dataset that contains information on all variables of interest and has a completeness of 86.1%. The variable with the poorest completeness is ethnicity in the THIN dataset (24.3%).

9.4 Discussion

This chapter has compared all the datasets analysed in this thesis using the variables of demographics, geography, time and dataset completeness. The closest similarities were between RIPL and HES, and RIPL and THIN. It is likely that these three datasets represent the same diseased population, which are interacting with different levels of the National Health Service. This offers scope for these datasets to be utilised within any future surveillance systems. Each of the variables compared will now be discussed in more depth, and their potential utilisation within surveillance systems will be explored in Chapter 10.

9.4.4 Discussion – Demographics

The general age distribution for Lyme disease cases, regardless of surveillance system, differs from that of the general population. It is likely to be a bimodal distribution with peaks in

cases seen in preteens and retirement age. The reasons for the overall differences in sex ratio are unknown, and have been discussed in Chapters 3-5, but may be due to differences in health seeking behaviour dependent on disease presentation.

There were only two age groups where all the datasets concur in terms of the direction of the sex ratio. Firstly, there were more cases seen in preteen boys than girls. This may suggest a different exposure risk for boys. The underpinning reasons for this need further exploration but could be due to differences in play activities or attitudes and awareness of a tick bite. Secondly, in early adulthood there were more cases in women than men. The reasons for this are unknown, and could include differences in health care seeking behaviour, health care prejudices, and unknown sex specific risk factors. All datasets also share a sudden change in ratio in geriatrics to being predominately female. This is likely, at least in part, a reflection of the national population becoming predominately female in this age group. Causal associations cannot be drawn from the current data and should be explored in future research.

The datasets providing information about ethnicity, deprivation, and rural status, all agree in their findings that Lyme disease patients were more likely to identify with being white and from areas with the least amount of deprivation, which were rural.

Limitations and Conclusions:

One of the limitations of the demographic data is that none of the datasets have been linked, the degree of overlap between them is unknown. However, the datasets will invariably overlap and the proportion of cases found in multiple datasets may cause bias in the conclusions drawn from each individual dataset. This could be resolved through the analysis of linked datasets, resulting in a more accurate understanding of the affected population.

Secondly, most of the demographic variables were studied univariably, due to the structure of the databases under analysis. It is likely that these variables confound and interact with each other, and a study enabling multivariable analysis is needed.

In this chapter, it has been assumed that each dataset has referred specifically to sex rather than gender. In fact, RIPL collects information on gender, whilst HES and THIN collect information on sex. As gender is decided through self-identification, there could be bias in the results and conclusions, as the gender of a patient may differ from the biological sex of a patient. There is the potential that the differences that were seen between RIPL and the

other datasets were a result of this bias rather than differences in any of the reasons discussed above.

Ethnicity data in the THIN dataset proves a challenge, due to the high level of missing data. Here, it has been assumed that the findings described in Chapter 5 are true, using ethnicity data from HES to help justify this. However, this may be a false assumption. There are significant challenges with obtaining ethnicity data in primary care [208,320], and until these are resolved the interpretation of any findings related to ethnicity in primary care must be treated with a degree of caution. This in turn could lead to issues with targeting public health messaging, with the wrong demographics being targeted.

In conclusion, all the three main datasets were different to the national population in terms of age and sex structure, ethnicity, rural-urban status, and measures of deprivation. Overall, these datasets are likely to be representative of the same diseased population accessing differing parts of the health care system.

9.4.3 Discussion – Spatial comparison

The spatial distribution of hospital admission cases was significantly similar to laboratory-confirmed cases of Lyme disease, both sharing positive spatial autocorrelation. The bivariate LISA plot identified areas with significant concordant and discordant clusters of incidence. The high-high clusters were all in southern to south-western England, highly suggestive of these being true hotspots of high disease incidence in England and Wales. The reverse of this is also true, that areas with low-low clusters are likely to represent areas with significantly low levels of Lyme disease.

The local authorities with discordant clusters with a high incidence in RIPL and low incidence in HES, could represent regions where diagnostic samples were received from cases managed within primary care without being admitted to hospital. This could be especially true in Wales, where there is a dispersed rural population and very few hospitals. It is possible that in these regions' cases were predominantly managed in primary care. The low incidence in RIPL, and high in HES, could be a result of doctors in primary care being aware of Lyme disease and confidently treating erythema migrans presentations, without taking diagnostic samples and sending them onwards to RIPL. This could result in a higher incidence of more serious cases being admitted into hospitals.

The ideas explaining these discordant areas are only hypotheses, and the underlying reasons for discordance remain unknown. Geographical data, at the same resolution, from primary

care would help identify areas with true discordance. Qualitative research to explore how patients with Lyme disease access health care, and how they are managed within the NHS, may provide reasons for these disparities. This type of research would need to take place over multiple geographies to ensure it was representative of the clusters, outliers, and areas with a non-significant background distribution.

The SAVSNET data did not show significant spatial concordance with RIPL data. There are at least a few potential reasons why this may have occurred. Firstly, the location of where there is a high presence of ticks on companion animals may not be associated with high Lyme disease incidence. This could be related to the prevalence of *Borrelia burgdorferi* within these tick populations. SAVSNET may have captured tick populations with a low prevalence, and so reflect tick populations that pose a limited Lyme disease risk to humans. It would be interesting to see whether SAVSNET's spatial distribution is concordant with PHE's Tick Surveillance Scheme, or whether a spatial distribution of within-tick Lyme prevalence is concordant with the RIPL dataset. Missing data is however likely to explain the differences better. There were many local authorities with no data about ticks on companion animals. This is most likely to be due to under-reporting, as discussed in Chapter 8, rather than a genuine lack of ticks. The time period under analysis only represents two years' worth of data, and the volume of data on ticks may not be large enough for SAVSNET to reflect the distribution of ticks. Currently, SAVSNET cannot be recommended as a surrogate surveillance system to produce risk maps in relation to ticks and Lyme disease. However, SAVSNET should be reassessed once there is a larger volume of data, as its potential to be part of a national tick and tick-borne disease surveillance system is large.

For a discussion about the significant concordance between RIPL and Twitter data please refer to Chapter 7. As an aside, due to one paper based on data from the 1990s [154], Thetford Forest and the Lake District have persisted as 'known' hotspots of Lyme disease [12], and are regularly mentioned in press releases about Lyme disease. None of the data presented in this thesis support these areas as hot spots. In no dataset do these areas appear as areas with a high incidence of Lyme disease. There is the possibility that the main cases originating in these areas are visitors to them, however one would still expect a higher base-line incidence in the local population compared to surrounding areas. Unless cases are only captured within primary care records, then the current narrative needs to change to highlight that the areas with the highest incidence are all in southern England.

Limitations

There are three main limitations of the geographical data assessed. Firstly, data for Northern Ireland and Scotland were either unavailable or have not been published. This will prove a challenge for any future public health interventions or strategic planning regarding Lyme disease in these nations. Scotland, in particular, has a more pressing need for this data as the incidence of Lyme disease in primary care (Chapter 5) is much higher than the other UK nations. Previous research suggests that the incidence of disease is not uniformly distributed across Scotland, with the Highlands and Western Isles having the highest laboratory incidence[76] and highest prevalence in blood donors [47]. Published Scottish geographic incidence data is needed with high granularity to ensure that public health resources are targeted in the most appropriate way.

Secondly, with the datasets that show significant spatial correlation (RIPL, HES and Twitter), the geographical location is based upon the patients' or users' home address. These data were therefore acting as proxies. This has been a consistent theme with the geographical data throughout this thesis, and it does lead to a conservative bias in exposure-response relationships relating to ecological fallacy. However, this is only true if one is trying to understand the riskiest area in terms of tick bite exposure. When looking at burden to the National Health Service, patient postcode is highly valuable information, as the maps produced throughout this thesis will show which localities potentially have the highest number of patients accessing their health services in relation to Lyme disease. This is important in terms of public health resource allocation, and highlights locations where future research could be conducted.

Finally, there is a lack of geolocation data for primary care Lyme disease cases. In absolute numbers, primary care has the largest diseased patient population, and understanding where they reside is important for any successful public health management and interventions. Currently no primary care database available to PHE provides access to geographical data other than at country level. If these resources are to be used for disease surveillance in the future, then the unavailability of geographical data needs to be overcome. Information should be provided at least at local authority level, and ethical measures can be put in place to ensure that no issues occur relating to patient identification due to small numbers.

Conclusions

Hospital admissions data and Twitter data have a strong spatial agreement with laboratory-confirmed cases of Lyme disease in England and Wales. As such, they offer potential as an additional geographical surveillance resource. However, hospital admissions data consistently have an incidence over five times lower than the incidence of laboratory-confirmed cases. Despite showing strong spatial agreement, it does not make sense to base a geographical surveillance tool on a dataset that captures less data than another. This makes HES potentially redundant for geographical surveillance. Twitter on the other hand, despite its limitations, offers some interesting potential as an adjunct surveillance tool (Chapter 7). In particular, its ability to identify people tweeting about Lyme disease in near-real time could be used to locate hot spots of activity. Its surveillance potential needs to be explored further to test its capability. These include its ability to measure levels of public concern about disease, or their appreciation and knowledge about the risk of ticks and their bites.

In conclusion, until primary care data are released that provide higher resolution geographical data, RIPL should remain the primary resource for geographic surveillance of Lyme disease in England and Wales.

9.4.3 Discussion – Temporal comparison

The comparison of UK national Lyme disease incidence trends between the Zoonoses Report and THIN datasets from 2007 and 2016 in the UK show a significantly similar trend in terms of the degree and variability of the increase. This suggests that primary care data reflects the temporal trends seen in laboratory-confirmed cases at a national scale. By calculating an average ratio of incidence between the two datasets, further information on the stability of the multiplication factor between the two can be gained (Table 9.2). The overall ratio of incidence between the THIN dataset and that in the Zoonoses Report was 2.91 (95% CI: 2.20-3.62).

The English and Welsh national trend for increasing Lyme disease incidence was seen in the Zoonoses Report, HES, and THIN datasets from 2007 to 2015. They were significantly similar in terms of the degree of increase. The variability in the increase was similar between the Zoonoses Report and THIN, however there is a significant difference between the Zoonoses Report and HES. The increased variability in HES may be due to the impact of the small case numbers on relative incidence figures; HES has the lowest incidence of all the datasets. This suggests that primary care data is more representative of Lyme disease in the community than hospital data in terms of temporal trends seen in laboratory-confirmed cases at a

national scale. By calculating an average ratio of incidence between each dataset, further information on the stability of the multiplication factor between them can be gained (Table 9.3). This resulted in an incidence ratio of 0.19 (95% CI: 0.09-0.29) between the incidence in the HES dataset and that in the Zoonoses Report. The incidence ratio between the THIN dataset and the Zoonoses Report was 2.35 (95% CI: 1.81-2.88).

Exploratory time series analysis of the RIPL, HES and THIN datasets provided some interesting insight. The similarity in distances, both for Euclidean distance and DTW, between the seasonal decomposition of each datasets time series show that they have comparable seasonality (Figs. 9.14 and 9.15). The most interesting element displayed is the consistent lag in peak month between datasets across the three years. Incidence peaks first in primary care (THIN) during July, followed by HES and RIPL in August. This makes sense both in terms of the clinical progression of Lyme disease and patient management pathways within the NHS. Clinical signs of early localised infection, classically erythema migrans with associated flu-like symptoms or borrelial lymphocytoma [21,23], occur within days to weeks after a tick bite. These presentations would probably not be severe enough to warrant a visit to a hospital and cases are likely to be dealt with in primary care. Weeks to months after a tick bite infection has disseminated and presents as Lyme neuroborreliosis or carditis. These presentations are more severe and are likely to be admitted at hospitals. The late dissemination of arthritis or ACA, are likely to present months to years after a tick bite and will present at any time of year. Although it cannot be confirmed in this study, it is likely that the THIN peak in July mainly represents acute EM rashes, whilst the HES admissions represents more complex/serious presentations. Serological testing is only recommended for non-EM presentations, and samples should be taken four to six weeks after onset of symptoms [23]. Thus, if cases are predominately presenting in July, since a diagnostic sample is recommended to be taken at least four weeks after symptom onset [23], samples would arrive at RIPL four weeks later in August. It is remarkable how closely the data matches this hypothesis. To explore this further, clinical presentation needs to be explored in different health datasets and their seasonal incidence described.

The trend components of the time series analysis were not as obviously similar as the seasonal components. In 2014, RIPL and THIN, but not HES, showed a dramatic decrease in annual incidence. This is likely to be driving the wave-like shape in the trend component of the time series. The fact that THIN data closely mirrors RIPL data, and has a smaller DTW distance, is indicative that it is likely to be representative of the trends shown in RIPL. The

data compared consists of only three years' worth of data, and it is likely that if more temporal data were provided then one could understand the degree of annual fluctuations in trend. It would show whether this is a standard feature of the trend component of Lyme disease time series in England and Wales.

The random component of the time series analysis showed similar Euclidean and DTW distances between the datasets, yet, all were non-stationary. This suggests that the datasets were alike but each having other underlying trends explaining the patterns of the random components of the time series. It was beyond the scope of this thesis to delve further into this analysis. However, if these data were to be used for forecasting purposes any further underlying trends need to be understood and described. Overall, it visually appears that the seasonal data is the main component of each of the time series, and the relative importance of the non-stationary random component of the time series currently remains unknown.

The summary of the proportion of monthly cases is a crude method as it ignores the differences between years and any overall trends. The twitter data represents only 12 months split between six months of two separate years. However, strong seasonality is seen again, with each dataset peaking between June and August. SAVSNET, representing tick activity peaks first in June, followed by THIN in July, and HES and RIPL peak in August. This appears to sequentially show the progression of disease presentation and NHS patient management, with the added peak of tick exposure a month before cases present in primary care. The Twitter dataset is probably too limited, in terms of number of years of data, to draw conclusions other than that mentions of Lyme disease on social media in the UK and Ireland peak in summer months. These trends reflect those already described within the timeseries analysis. SAVSNET data has the potential to accurately describe peak tick activity, which could be utilised to predict the peaks of Lyme disease cases annually. Further work is needed to explore this potential.

Conclusions:

The seasonality of all the datasets is striking, particularly as they peak sequentially in the summer months. First, the presence of ticks on companion animals (a proxy for tick numbers in the wild, and potentially for tick bites) peak in June, then cases in primary care peak in July, followed by hospital admissions and laboratory-confirmed cases peak in August. This displays the progression of disease and care pathways of an individual bitten by an infected

tick. The sequential series of peaks could be utilised to predict when differing sections of the health system will see the highest burden of Lyme disease cases each year.

The incidence of Lyme disease coded cases in primary care electronic health records are reflective of laboratory-confirmed cases, used for current surveillance figures. At a UK level there is a stable multiplication factor of 2.91 (95% CI: 2.20-3.62) converting laboratory-confirmed incidence to the UK incidence of Lyme disease in primary care, in England and Wales this is 2.35 (95% CI: 1.81-2.88). English and Welsh hospital admission records were less representative as they have larger relative differences in variability compared to laboratory records. The multiplication factor, converting laboratory incidence to hospital admissions incidence, is 0.18 (95% CI: 0.09-0.29) (Fig. 9.18).

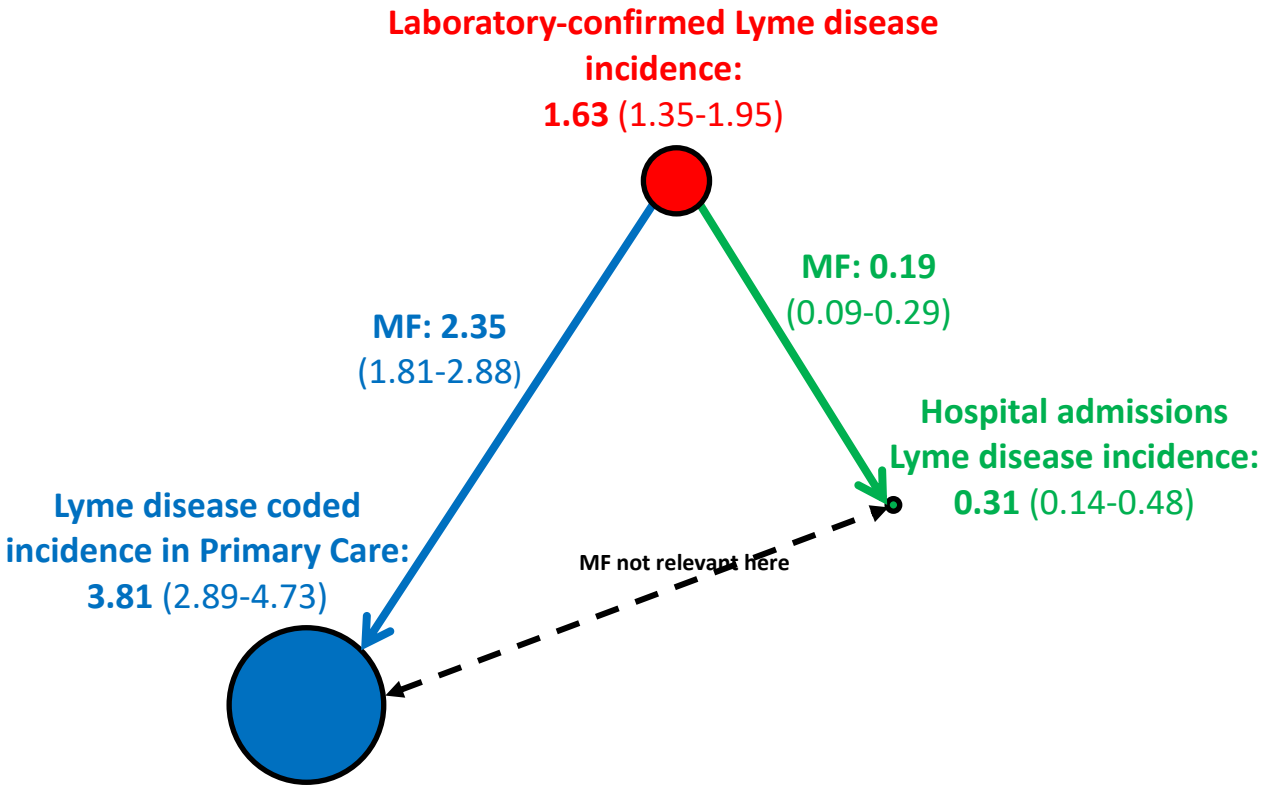


Figure 9.18 The average annual incidence (cases per 100,000) of each health dataset for England and Wales (the size of circle is proportional to incidence), and the multiplication factor (MF) between each. Blue = THIN (primary care), Red = RIPL (laboratory), and Green = HES (hospital admissions).

Primary care Lyme disease coded cases are the most similar to laboratory-confirmed cases in terms of annual and seasonal incidence. This is reflected in the narrow confidence intervals of the multiplication factor calculated between the two. Using this multiplication factor, one could estimate the incidence of Lyme disease activity (as currently THIN captures non-validated cases of Lyme disease) by applying the multiplication factor to RIPL. This process

could also be used to estimate the number of hospital admissions; however, this would be less reliable than the THIN multiplication factor due to the relatively larger confidence intervals. If these multiplication factors are routinely assessed for their stability, they could provide a low-cost addition to the existing surveillance system and provide PHE with robust evidence-based annual estimates of Lyme disease for England and Wales.

9.4.4 Discussion – Completeness

For each dataset over half of the records were complete for all the variables captured, except ethnicity in THIN where less than 25% of records were complete. Twitter and SAVSNET provide complete temporal data. SAVSNET has excellent spatial data, whilst Twitter is less complete. Since both are proxy datasets and provide no information about the Lyme disease patient population, they will not be discussed further in terms of completeness. Out of the human health datasets, HES performs the best for completeness, with all variables having more than 70% of completeness. It is also the only dataset that contains data on all measured variables. THIN performs the next best, and 100% of records are complete in terms of baseline demographic information. However, it has two significant issues. The ethnicity data were highly incomplete, and no high resolution geographical data were present. These issues have been discussed previously in Chapter 5. The poorest performing health dataset is the RIPL dataset. As discussed in Chapter 3, RIPL are entirely reliant on the referring hospital laboratory to provide all relevant data. As these results indicate, this is currently far from optimal. The poor completeness scores are predominately a reflection of deprivation and rural status data being linked to provided patient postcode data. The majority of cases do have a postcode provided, but greater completeness would lead to greater confidence in any geographical conclusions drawn. The RIPL dataset does not collect any ethnicity data, how important this is currently unknown. This could be easily rectified by the addition of an ethnicity question to PHE's Lyme disease test request form [158].

Basing the selection of a dataset for surveillance purposes solely on data completeness wrongly assumes that completeness is related to data quality, data accuracy and the degree of representativeness of the Lyme disease population. Instead this information should be used to assist in identifying which datasets provide data on certain variables, and the degree to which they could be explored with confidence.

9.5 Summary and conclusions

In this chapter, comparisons have been performed between the datasets described in this thesis in terms of agreement and completeness about several demographic, temporal and spatial variables (Table 9.5).

Table 9.5 Summary table of datasets evaluated in this thesis (* represents a significant difference (p<0.05) between either the specified reference population or the national population, except in the Global Moran's I category where it represents a significant spatial agreement; ^{NS} not significant)

Variables		Datasets				
		RIPL	HES	THIN	Twitter	SAVSNET
Temporal Comparison	Slope	0.03 Ref	0.03^{NS}	0.01^{NS}	N/A	N/A
	Pearson Correlation	0.59 Ref	0.96*	0.07^{NS}	N/A	N/A
Peak month of cases		Aug	Aug	Jul	Aug	Jun
Multiplication factor		Ref	0.18 (95% CI: 0.09-0.29)	2.35 (95% CI: 1.81-2.88)	N/A	N/A
Spatial	Global Moran's I	Ref	0.47*	N/A	0.17*	0.05 ^{NS}
Demographic factors	Sex Ratio	1.14 Ref	0.66*	0.84*	N/A	N/A
	Index of Multiple Deprivation	Number of cases increase as the level of societal deprivation improves*			N/A	N/A
	Ethnicity	N/A	96% white*	94% white*	N/A	N/A
	Rural-urban status	More rural than the national population*				
Completeness		66.8%	86.1%	73.2%	76.0%	100%

In Chapter 2, a surveillance pyramid was constructed to aid in the identification of potential datasets capturing Lyme disease (Fig. 2.1). The comparisons of incidence have highlighted, that if following the traditional structure, the tiers of the datasets are in the wrong order. The datasets with the lowest incidence were the hospital datasets and should in theory be placed on top of the pyramid, followed by laboratory data then primary care data. This, however, ignores the underlying fact that cases could occur in multiple datasets and are not unique to a single tier of the surveillance pyramid. One could conclude that a surveillance pyramid is potentially a poor model for Lyme disease surveillance. In reality it is better represented by a Venn diagram-like structure, similar to that proposed in the IID2 study [121]. However, such a diagram cannot be constructed in this study, as the degree of overlap

between the Lyme disease datasets remains unknown. At best the data can be summarised as in Fig. 9.18 and it can be acknowledged that the proportion of patients found in each dataset is currently unknown. Future research on Lyme disease needs to be able to quantify this overlap so more accurate estimates of disease incidence can be calculated.

Regarding specific datasets, Twitter and SAVSNET have both offered potential for Lyme disease proxy or sentinel surveillance. However, temporal data was not long enough to allow comparison to current incidence surveillance data, and only Twitter showed significant spatial agreement with current incidence surveillance data. No meaningful data were available regarding case demographics. Both approaches offer exciting avenues for future research, but they do not offer data that is comparable to the existing surveillance system across many variables. As such, they have been excluded from further discussion.

The remaining datasets (RIPL, HES and THIN) are comparable across multiple variables, including time, geography and demographics. They appear to be representative of the same diseased population, although the degree of overlap between the datasets is unknown as data linkage was not possible. In comparison to the current surveillance system (RIPL) HES additionally captures ethnicity data and is comparable except for sex ratio and variability in incidence trends. Despite these facts, and its high degree of completeness, there are some concerns about using it as a surveillance system. The dataset represents the smallest proportion of Lyme disease cases, and without further assessment of the clinical records of the patients it is difficult to know the validity of the ICD-10 codes used to define cases. There is potentially a large overlap of cases with the RIPL dataset, as late disseminated or complicated cases of Lyme disease admitted to hospitals are nearly all likely to receive diagnostic investigation, which would result in samples being sent to RIPL.

The THIN dataset also captures ethnicity, though poorly, and is comparable to RIPL data except for the sex ratio. It has the largest representation of Lyme disease cases. One of its flaws as a potential surveillance system is the unavailability of geographical resolution. Validation of Lyme disease Read codes and coding behaviour of GPs in primary care need to be further explored (as discussed in Chapter 6) to provide higher levels of confidence in the data. This dataset is also likely to capture cases that do not appear in the other databases, specifically cases presenting with erythema migrans and receiving a clinical diagnosis, and thus may describe a wider diseased population.

It must be remembered that only RIPL data is currently performing as a surveillance system. The work presented in this thesis are observational studies, and do not formally test or evaluate the datasets as potential surveillance systems. Before any policy changes take place, formal trial and evaluation work is needed to check that the assumptions made, and conclusions drawn in this thesis are correct. Taking these factors all in to account, primary care activity data offers the greatest potential as a new or adjunct surveillance system for Lyme disease in England and Wales.

Chapter 10 Discussion

10.1 What this thesis adds to the knowledge base

In this thesis I have used a mixture of routine health data, and novel datasets, to provide new insights in to the current epidemiological situation of Lyme disease in the UK. The populations of Lyme disease cases that can be found in English government laboratory data, English and Welsh hospitals, and UK primary care, were described. A new mixed methods technique was described to validate the diagnostic codes (Read codes) used to identify Lyme disease in primary care. Twitter and a companion animal electronic health record system (SAVSNET) were evaluated as adjunctive surveillance tools for Lyme disease. By drawing comparisons between all these datasets, new options for Lyme disease surveillance were identified.

In the routine health data studies (laboratory, HES and primary care) that make up this thesis several novel findings were identified. They all show a significant increase in Lyme disease incidence across their respective study periods. Notably, as discussed in Chapter 9, the described incidence was increasing at the same rate across various health care settings. This shows that each tier of the NHS is seeing a similar increasing rate of new Lyme disease cases. Whether this increase is due to a real increase in cases, increased public and/or physician awareness, or a mixture of the two was outside the scope of this thesis and should be the focus of further research.

The populations affected by Lyme disease showed similar bimodal age distributions, peaking in preteens and around retirement age. This distribution was significantly different to the national population structure in England and Wales. This reflects other Lyme disease populations described in Europe [67,176,177,180], and suggests that exposure behaviour to *Borrelia spp* carrying ticks is similar across Europe. However, there were differences between the sex ratio of these populations across datasets. Reasons for this are unknown but could be due to differences in accessing the differing tiers of healthcare, healthcare prejudices, or different risk exposure behaviours. Ethnicity data was limited but suggest that the population affected by Lyme disease is not diverse, and predominately of a white ethnicity.

This is the first UK based research to have explored trends in socioeconomic deprivation in relation to Lyme disease. All datasets showed similar trends, with the primary care data displaying the most compelling evidence that the incidence of Lyme disease-related diagnosis is highest in areas with the least deprivation. The reasons for this are likely to be

multifactorial and need to be explored through future research. This is critical as any public health interventions need to be targeted towards the appropriate population.

This is the first time that spatial Lyme disease data has been produced for England and Wales at a high geographical resolution. Several datasets identified the south and south-west of England as the main areas with a high incidence of Lyme disease. Interestingly, previously recorded hotspots of disease, Thetford Forest and the Lake District [154], were not identified in any of the datasets evaluated. These data will be useful to local authority public health teams for targeting disease awareness and intervention strategies. They also provide locations where researchers interested in Lyme disease epidemiology, at a local level, can focus their efforts.

A novel approach to validate diagnostic codes, Read codes, was developed for use in primary care electronic health records. Traditionally Read codes are validated through expensive and time-consuming methods, which may not provide additional insight in to coding behaviour. By conducting mixed-methods interviews with GPs, the validity of specific Lyme disease Read codes was described, and the thought processes used in the selection of a code explored. Themes (Chapter 6) were identified that describe the decision making behaviour for code selection in primary care. This highlighted that the conscious and unconscious behaviour driving coding is not solely limited to clinical acumen and awareness but is intertwined with personal experience and the national political landscape. The results suggested that the incidence described is likely to be an underestimate of the true incidence of Lyme disease in primary care. These findings are crucial for understanding the patient-doctor dynamic in primary care, which is incredibly important for such a highly politicised disease as Lyme disease. The research will hopefully provide a platform for further qualitative research on Lyme disease and coding behaviour in general.

The remaining datasets analysed did not contain routinely collected human health data. One set of data was collected from a social media platform, Twitter, and the other was routinely collected companion animal health data, SAVSNET. As far as the author is aware, this is the first time that Twitter data for an infectious disease has been explored to this degree of geographical resolution. The data showed significant spatial concordance with that of national surveillance data for England and Wales. This highlights that Twitter has the potential to be utilised in a surveillance system and to identify locales with the greatest related social media activity. The data needs to be further assessed to see whether this activity is related to disease concerns, health promotion, or individuals discussing their own

cases. SAVSNET has the potential to produce temporal and spatial figures for tick activity in the UK. Currently the data is spatially sparse and is not concordant with known areas of high Lyme disease incidence. Once further data is collected using this surveillance system it could play an important role in the surveillance of both ticks and tick-borne diseases.

One of the most important findings from this thesis has been the derivation of a multiplication factor that can be applied to laboratory surveillance figures to give a national estimate of Lyme disease incidence within primary care. This gives public health organisations better evidence of the potential burden that Lyme disease has at a national level, and on an annual basis. This multiplication factor has the potential to be a low-cost, easily implementable tool that can be used within the current Lyme disease surveillance framework. There is scope for further validation of this factor to identify whether it can be used at a local level and calculating its degree of stability over time.

This thesis has produced new epidemiological data that was specifically requested by both the NICE guidelines [23] and EPPI recommendations [9,22,72–74]. It enables future researchers and policy makers to inform their planning and decision-making within a more evidence-based context.

10.2 Lyme disease surveillance policy options

To place the findings of this thesis into a surveillance framework, one must identify how these datasets could be implemented practically through national surveillance programmes. In Chapter 1, several Lyme disease surveillance policy options were presented; created by the EPPI on behalf of the Department of Health and Social Care (DHSC) [9]. These options will now be addressed, based upon the thesis findings, and a proposed surveillance strategy will be proposed.

10.2.1 Maintain the existing system with no change

The current surveillance system for England and Wales is based on laboratory-confirmed cases identified by RIPL and published in the Zoonoses Report. This is the easiest strategy to follow, both in terms of cost and management. As discussed in multiple chapters, this system is likely to underestimate the incidence of Lyme disease. This thesis has described multiplication factors that can be applied to the RIPL dataset to estimate the incidence of Lyme disease coded patients in primary care, and the incidence of Lyme disease related hospital admission (Fig 9.18).

These could provide a simple methodology for estimating the national incidence of Lyme disease, and would enable PHE to issue, with relative confidence, statements such as 'for each laboratory-confirmed Lyme disease case, two may be seen in primary care'. Further assessment of these multiplication factors is needed to test their robustness over time.

As a tertiary reference laboratory, RIPL currently lacks case ethnicity reporting, and geographical reporting of cases is incomplete. If this policy option is chosen, then these two issues need to be addressed. Further confusing the situation is the potentially missed laboratory-diagnosed cases that could either appear in SGSS or private laboratory databases (chapter 3). This needs to be rectified.

There are two potential solutions. Firstly, to legislate *Borrelia burgdorferi* as a notifiable causative organism (instead of *Borrelia spp)* through the 2010 Health Protection (Notification) Regulations [77]. In theory, all laboratory-confirmed cases of Lyme disease would then be captured through SGSS. This would have the additional benefit of the collection of ethnicity and case home address data, as they are mandatory information to be provided in SGSS. Therefore, more complete demographic information would be collected. However, as discussed in chapter 3, 91.6% of all *Borrelia spp* reports in SGSS were already classified as *Borrelia burgdorferi*, so it could be questioned whether such a change in legislation would have any significant impact. It is likely that the inconsistencies between SGSS and RIPL would remain. An audit of the Lyme disease tests performed by referring laboratories, and which results they report to SGSS, would help to highlight some of the reasons for the differences. Until there are no discrepancies between the RIPL and SGSS data, national surveillance figures should continue to be based on RIPL data.

Secondly, PHE could recommend that all confirmatory Lyme disease testing is performed solely at RIPL. This would ensure that RIPL captures as many laboratory-confirmed cases as possible. A stricter version of this policy has been recently adopted by Health Protection Scotland (HPS), who have stated,

> *'Following this* [sic 1st October 2018], *all diagnostic laboratories* [in NHS Scotland] *will submit samples to the Reference Laboratory* [Scottish Reference Laboratory for Lyme Disease and Tick-borne Infections, Inverness] *for screening and confirmation, and will result in more complete data on laboratory confirmed cases.* [321]'

Therefore, all suspect Lyme disease diagnostic samples in NHS Scotland will be sent directly to the reference laboratory, and HPS surveillance figures will contain limited missing laboratory data. Routes for diagnostic testing are not as explicitly described for England [322], and refer to protocols within the NICE guidelines [23]. Requesting that all diagnostic laboratories in NHS England and Wales send positive screening samples to RIPL for confirmatory testing should be feasible. A statement similar to that which appears on HPS's website could be added to PHE's website [322] and/or be included in future amendments to the NICE guidelines [23].

10.2.2 Introduce mandatory clinician reporting for all Lyme disease cases

This approach has been taken in other countries in Europe (e.g. Netherlands, Belgium, Slovakia) [9]. These countries all have a much higher incidence than the estimates generated in this thesis and are likely to have a greater awareness of Lyme disease amongst the general population and medical professionals. There are some issues with this option. Firstly, it requires a significant change in legislation to the Health Protection (Notification) Regulations 2010 [77]. Since this legislation has been passed no new diseases have been added to the notifiable disease list. Since its preceding legislation, The Public Health (Infectious Diseases) Regulations 1988 [323], only five diseases have been added (botulism, brucellosis, invasive group A streptococcal disease, Legionnaires' disease, and severe acute respiratory syndrome (SARS)), and three have been removed (AIDS, leptospirosis, and ophthalmia neonatorum). To introduce mandatory reporting, Lyme disease would have to become Schedule 1 notifiable disease in addition to *Borrelia spp* being a Schedule 2 causative agent.

The differences between the interpretation of a Schedule 1 notifiable disease and Schedule 2 causative agent lie in the respective public health actions to a positive case. The differences are essentially that of the degree of urgency and the scale of a response [77]. A suspected Schedule 1 disease must be notified to a 'proper officer' of a local authority within three days, or 'orally as reasonably practical.' A 'proper officer' is someone given statutory power to initiate actions regarding notifiable diseases [323]. These public health and disease control actions could include, among others, the quarantine of individuals, the initiation of contact tracing, mass vaccination, or the withdrawal of food items from sale. A laboratory identifying a Schedule 2 causative agent must notify PHE within 7 days, this is usually through SGSS. If the

> '*diagnostic laboratory reasonably believes that the Health Protection Agency* [sic PHE] *has already been notified in accordance with this regulation by the*

operator of another diagnostic laboratory in relation to the same causative agent being found in a sample from the same person.[77]'

This means that if a confirmatory test has been performed by a PHE laboratory (for example RIPL), there is no need to notify PHE. Schedule 2s are not notified to a 'proper officer' but are notified directly to PHE. PHE holds no statutory power to implement any public health actions, without the assistance of a 'proper officer'. As such, Schedule 2 pathogens require surveillance but do not need immediate public health interventions on detection. In summary, Schedule 1 require urgent public health action, whilst schedule 2 require surveillance.

In comparison to the notifiable disease list, Lyme disease's characteristics are very different. It has no human to human spread, it is not vaccine preventable, it is not highly contagious with the potential for epidemic spread, it is not fatal in the vast majority of cases, and it is not a disease virtually eliminated from, or exotic to, the country. The National Institute for Public Health and the Environment (RIVM), the Netherlands, have made a decision aid for policy makers to decide whether an infectious disease should be notifiable [324]. By following the criteria in this decision aid Lyme disease would not qualify for being notifiable. The criteria for a disease becoming notifiable in England are not explicit, but in my opinion, it is difficult to justify it becoming notifiable based on the above, and would set a new, potentially unwanted precedent.

Secondly, it is hard to think of an immediate public health intervention that could be actioned if a positive case of Lyme disease were identified. Control measures for Lyme disease typically revolve around tick bite prevention measures and education of the general public and medical practitioners [21,325,326]. These are all measures that take time and would need to be coordinated at a localised level. They lack the sense of urgency and elevation of scale, (i.e. the necessity for a 'proper officer' wielding statutory power) that a Schedule 1 disease traditionally demands. The author is unsure what public health action could occur immediately after notification which would have a direct impact on Lyme disease control and prevention. It could be argued that by changing the status of Lyme disease to a Schedule 1, local authorities in high incidence areas would be spending an inappropriate amount of resources, compared to other Schedule 1 diseases, dealing with the notification of Lyme disease cases.

Finally, reporting of Lyme disease is likely to be poor, due to the lack of awareness of the clinical presentations of Lyme disease within the British medical profession (chapter 6). There is the risk that this could result in any rash with an uncertain diagnosis being reported as Lyme disease. There are also many examples in the literature where GPs would first await a laboratory confirmation before notifying the appropriate authorities about a disease [327–331]. A systematic review of the completeness of infectious disease notification in the UK found that the range of reporting completeness was from 3 to 95% [331]. For example, food poisoning was reported as having 47% completeness of notification. This is likely due to GPs sending food poisoning samples for diagnosis first before notifying, or assuming that the diagnostic laboratory would notify the 'proper officer' on identification of the causal agent. The case is then likely to be notified as the pathogen, rather than as the notifiable condition 'food poisoning.' One can see that this could happen with Lyme disease; the GP would send a diagnostic sample for confirmation and may not notify a case as they would correctly assume that RIPL would do so if the diagnostic result were positive. Essentially, GPs are likely to only report laboratory-confirmed cases rather than "on suspicion". Therefore, underreporting of cases would still be likely to remain.

Therefore, to ensure appropriate reporting, mass education of health professionals about Lyme disease and the appropriate notification process would be needed, with the potential addition of making it a QOF-able condition. This is known to be successful [332], but these measures would be costly and time-consuming approaches for the DHSC.

The systematic review of Lyme disease surveillance systems, discussed in chapter 1, concluded that the introduction of mandatory clinician notification would increase the number of cases captured by surveillance, but it would not necessarily be a more reliable methodology than laboratory-confirmed cases alone. In my opinion there needs to be a cost-benefit analysis of this policy option before adoption, as it currently seems difficult to justify.

10.2.3 Introduce mandatory clinician reporting for late or disseminated Lyme disease cases

This option has the same issues as above. Given that the awareness of the differing clinical presentations of Lyme disease varies by location, and is potentially already poor in some areas, it may be challenging for clinicians to identify late or disseminated Lyme disease cases. Many of these cases would be managed in a hospital setting and be likely to have diagnostic samples taken. If so, they would already appear within the RIPL dataset. This option could therefore produce a change in legislation but not impact the number of Lyme disease cases reported. However, if it were adopted, it would be synergistic with the European Centre for

Disease Prevention and Control's (ECDC) surveillance list [333,334], where Lyme neuroborreliosis is under surveillance.

10.2.4 Include Lyme disease in clinician sentinel networks

This would be one of the easiest options to adopt. In this thesis, Read codes have been identified that could be used in primary care sentinel networks (chapter 5). Currently two systems could be adapted for this process, the Real Time Syndromic Surveillance Team (ReSST) [335], as part of PHE, and the RCGP research and surveillance centre [336,337]. The ReSST collects anonymised GP consultation data from two clinical databases, QSurveillance and SystmOne [335] (See Chapter 2). The ReSST covers 55% of England's population in primary care and the RCGP network only covers 2.9%. The geographical distribution of the ReSST is undisclosed, whilst that of the RCGP network is very unrepresentative (it only has two practices located in the disease hotspots identified in this thesis [337]). As discussed in chapter 2, there are a variety of research databases that could be used for surveillance, such as THIN (chapter 5). They are all more representative than the RCGP, however they are not primarily designed as surveillance systems, and do not provide geographical distribution. If this policy option were adopted then the ReSST would be the obvious network to choose, with the caveat that geographical information is also collected.

10.2.5 Introduce enhanced surveillance using clinician questionnaires

This option will not identify further cases but would provide more information about the confirmed cases identified by RIPL. Information about the case presentation and management of cases within the UK is needed, as there is a sparsity of data [9,23]. It would enable the collection of exposure data and would potentially be useful for identifying risk factors. This surveillance method is time-consuming and expensive, and the cost-benefit ratio of this approach needs to be assessed. This is especially true for Lyme disease in England and Wales as diagnostic samples are sent to RIPL via hospital microbiology units, rather than directly from primary care; this results in data attrition (Chapter 3). Some of the data lost would include primary care clinicians' contact details. A significant amount of resource and goodwill would be needed from the referring hospital laboratory to facilitate RIPL contacting a referring GP to gain access to additional patient details.

A systematic review of notifiable diseases found that even diseases under enhanced surveillance had sub-optimal recording of cases [331]. Much of the demographic information captured in this thesis is likely to be replicated during enhanced surveillance. The questionnaire could provide additional information on exposure risk to individuals and help identify risk behaviours that could result in a tick bite, and subsequently Lyme disease.

10.3 Surveillance policy recommendations

The traditional surveillance pyramid is an imperfect model as the population in each tier are often linked, with significant degrees of overlap. The extent of this overlap is unknown, and to survey only one tier is likely to represent a biased view of the diseased population. Therefore, a combination of health datasets is likely to offer more representative coverage of the diseased population. As stated by Lorenc et al;

> '*a combination of methods gives more complete coverage in terms of the identification of cases than any single method alone, but also that no combination can guarantee full coverage of all cases.'* [9]

This thesis has concluded that the THIN and RIPL datasets are likely to be the most representative of the health datasets available for Lyme disease surveillance, and so any surveillance policies should be based on these or similar systems. I believe that a Lyme disease surveillance policy should:

- Provide reliable estimates of the incidence of Lyme disease in primary care, calculated from RIPL figures and the multiplication factor described in this thesis. This should be a very achievable goal.
- Maintain current surveillance performed by RIPL and legislate that it be the only laboratory, in England and Wales, allowed to perform confirmatory testing. This would remove the need for the flawed SGSS reports, at least regarding Lyme disease.
- Include Lyme disease Read codes within ReSST. These will provide estimates of the incidence of Lyme disease in primary care, at much greater demographic coverage than any research primary care electronic health record database. Pursue geographical data of these cases. If this is not possible, then further assessment of primary care databases like THIN or CPRD should be performed to assess the stability of the proposed multiplication factor. Access to geographical data in these datasets should be sought.
- Tailor the RIPL sample submission form to include more questions, that would be traditionally collected within an enhanced surveillance questionnaire. Incomplete forms of positive cases should be followed up. Information greatly needed include geographical data, case management and treatment choices, ethnicity, and tick bite history. As noted above, there are likely to be resource issues associated with the follow up process, and if enforced, potential resistance from referring clinicians. However, I feel it would provide additional data, and so it would be prudent to try.

10.4 Limitations and future research

10.4.1 Limitations

Seven datasets have been independently assessed for their ability to describe Lyme disease incidence and case demographics. This analysis has not allowed the datasets to be linked. This presents a challenge, as the extent of the overlap of cases between the datasets remains unknown. In chapter 4, it was found that over 25% of hospital admissions originated from primary care. However, without official linkage of the datasets the exact degree of overlap is unclear. Ethical approval was granted to link THIN to HES but it was technically impossible to interface the two systems. Linking the RIPL servers to any other database is not allowed due to the sensitive nature of some of the other conditions contained within it. Since the RIPL-THIN multiplication factor is based upon two unlinked databases, we do not know the degree of overlap between them and therefore how the multiplication factor is affected. Ideally any future work needs to be able to describe and define the degree of overlap between different health datasets.

The structure of the datasets and the inability to stratify some of the denominator populations meant that only univariate analysis could be performed, without multivariable analysis. For example, within both RIPL and HES, a denominator population was not provided and so the ONS population was used. As these two datasets are independent, it would be remiss to perform multivariable analysis as the numerator datasets are nested, in an unknown manner, within the denominator population and they could not be stratified to the same levels. For example, one could stratify the national population by age, sex and year, but not additionally by postcode area, ethnicity and rural-urban status. THIN did have a denominator population, and so could be stratified by all variables except ethnicity. The lack of ethnicity stratification may be explained by the low level of record completeness for this variable. The study design of the analysis of THIN was solely ecological, and so multivariable analysis was not planned. This would be recommended for any future research on this or similar primary care databases. Without the ability to perform multivariable analysis these demographic factors cannot be disentangled to identify any interaction or confounding effects of the variables under investigation and the bias inherent in them remains unknown.

From a surveillance perspective, the relative lack of geographical data is a fundamental limitation. As shown within this thesis, Lyme disease is a condition that has strong positive spatial autocorrelation. However, the two main datasets recommended for future surveillance policies have either no geographical data (THIN) or data is only available for

around half of all cases (RIPL). This needs to be urgently addressed. For any future modelling of Lyme disease in the UK these data need to be more robust than at present. In addition, it has been discussed across multiple chapters how the geographical data is reliant on case residence postcode rather than tick bite location. Nonetheless, in high risk areas the great majority of cases are probably acquired locally during rural activities, and the error rate is likely to be small. In cities, cases are likely to originate from many centres and may result in a higher proportion of "out-of-area" acquisitions. The extent that this impacts tick bite/risk modelling is unclear and needs to be explored. The current data is still useful as it indicates where Lyme disease places a burden on health care, and so can be used for appropriate resource allocation and to target public health messaging and interventions.

A potential limitation of all the datasets explored is a lack of information about the background prevalence of Lyme disease in the UK population. Since the datasets are interrelated it is difficult to interpret whether the observed incidence and dataset synergies relate to the background risk of exposure and prevalence or are due to deficiencies in the clinical management and diagnostic systems in different locations. This offers scope for future research. A Lyme disease seroprevalence study like that performed in Scotland [47], would provide an understanding of the underlying spatial distribution of prevalence, and potentially disease risk. If maps generated from such a study had significant concordance with data produced through national Lyme disease surveillance, we could be more confident that the current surveillance system is fit for purpose and not missing previously unidentified hotspots of disease.

The final limitation is around the identification of a Lyme disease case within a dataset. Only RIPL offers strict clarity, as cases are identified based on the results of diagnostic tests. The remaining datasets were based on diagnostic coding and are therefore reliant on appropriate coding by a health professional. As discussed in chapter 6, this is not as reliable as initially perceived, as there may be a lack of awareness of a certain condition, or there may be a reluctance to code for a variety of reasons. This means that the diseased populations described are still likely to be biased and not represent the entire population seen either in hospital or primary care. This is a known issue across all research and surveillance performed on datasets reliant on diagnostic codes such as ICD-10 and Read codes. To date no solution to this problem has been identified. As previously discussed even making a code QOF-able does not result in more reliable coding. Validation of Lyme disease codes needs to be further investigated to avoid the maxim of 'garbage in, garbage out'.

10.4.2 Future research

The most immediate potential for future research is to assess whether the ReSST or one or more primary care database, can be used as a continuing surveillance system for Lyme disease. A formal surveillance system assessment needs to be performed to decide on their potential. Alongside this, the stability of the multiplication factor identified needs to be assessed to decide whether this number can be utilised as an estimate for the number of Lyme disease cases seen in primary care in the UK.

The analysis of the datasets in this thesis has focused solely on disease incidence and demographics, however, this misses a large amount of the data contained within them. The HES and THIN datasets could be examined further to identify the prevalence of the various clinical presentations of Lyme disease in the UK and how they are clinically managed. This could include analysis of treatment and referral choices. This information is currently not present for the UK and its analysis would fill important gaps in the evidence base regarding Lyme disease. Following this work, health economic analysis would be strongly recommended, to calculate the burden that Lyme disease currently places on the NHS.

There is a need for a UK-wide seroprevalence survey to understand the background risk of Lyme disease, therefore providing exposure data for the UK population. To date, similar work has only been performed in Scotland [47]. Such research would allow better interpretation of regional variations in incidence, such as those identified in the thesis. For example, are the areas with low case numbers reflective of a low exposure risk or poor data recording or poor case recognition? It would place the findings of the thesis into better context, and provide the opportunity to identify risk factors associated with higher Lyme disease seroprevalence, such as occupational/recreational practices. The resultant data could be linked to tick distribution maps, produced by the Tick Surveillance Scheme [24,286], to produce high resolution Lyme disease risk maps for the UK. It would enable better targeting of public health advice and future Lyme disease interventions. It would provide more robust evidence of the true burden of Lyme disease in the UK and produce a definitive answer to the level of Lyme disease exposure to the UK population.

This research, excluding the SAVSNET chapter, has excluded the inclusion of ticks in its analysis. Ticks are obviously critical to the spread and risk of Lyme disease. There are multiple research groups within the UK who are currently performing valuable research regarding ticks, though mainly from an ecological perspective. Important epidemiological and management questions that still need answering include, ‘can the incidence of tick bites,

both temporally and spatially, be used to predict the incidence of Lyme disease?' and 'how are tick bites currently managed in the UK?' Answers to these could offer scope for tick-based surveillance methods for Lyme disease, and may highlight the need for education and awareness around the management of tick bites.

The final area that needs exploring is patient and health care professionals behaviours, experience and knowledge around Lyme disease. As highlighted in Chapter 6, the reasons behind GP coding behaviour is more complex than perhaps perceived. The interview process, rather than the questionnaire, drew out how GPs feel and behave. Despite the bias in the group, it provided some fascinating insights. Similar qualitative research is needed to further explore coding behaviour, and to understand health professionals' attitudes and beliefs to Lyme disease. For patients, research is needed to understand how, when and why they access various elements of the health care system. It is also needed to understand the views and beliefs within the Lyme disease affected population, both those with solely an EM rash and those that identify with having 'chronic Lyme disease'. By understanding personal drivers to attitude and behaviour, public health strategies can be designed that reflect and respond to these drivers better than present.

10.5 Conclusions

Lyme disease offers a unique set of surveillance challenges in the UK. Little is known about its UK epidemiology, not all cases need diagnostic confirmation, and the awareness of its various presentations amongst health professionals is variable. I have identified a variety of health datasets that could be used as part of an integrated disease surveillance system.

The NICE guidelines stated there was a lack of epidemiological data about Lyme disease in the UK. This thesis has filled elements of this evidence gap, by presenting new incidence figures, providing new Lyme disease incidence maps for England and Wales, and for the first time, describing the socio-demographics of the Lyme disease affected population in the UK. This work will provide a new reference point for future surveillance and research projects for Lyme disease in this country.

There are two key policy recommendations resulting from this research. Firstly, current geographical data is relatively sparse, despite the identification of significant hotspots of disease in southern England. Due to this geographical restriction any surveillance systems adopted by Public Health England need to ensure that the capture of geographical data is deemed a vital part of the system. This will allow the assessment of disease spread and the appropriate allocation of public health resources and interventions. Secondly, a combination

of health datasets for surveillance is likely to be the most representative of the diseased population. Primary care and laboratory datasets appear to be the most representative, and a multiplication factor of 2.35 has been calculated between the two. This multiplication factor can be used to provide estimates of national and local Lyme disease incidence in primary care and assist in the allocation of public health resources for Lyme disease control, presentation, and education. The use of routinely collected primary care electronic health records data alongside laboratory data, should be considered as the core of any future Lyme disease surveillance policy.

References

1 Steere AC, Malawista SE, Snydman DR, Shope RE, Andiman WA, Ross MR, et al. Lyme Arthritis and Adults in Three Connecticut Communities. *Arthritis Rheum.* 1977;20(1):7–17

2 Burgdorfer W, Barbour AG, Hayes SF, Benach JL, Grunwaldt E, Davis JP. Lyme disease-a tick-borne spirochetosis? *Science.* 1982;216(4552):1317–9

3 Baird AG, Gillies JC, Bone FJ, Dale BA, Miscampbell NT. Prevalence of antibody indicating Lyme disease in farmers in Wigtownshire. *BMJ.* 1989;299(6703):836–7

4 Muhlemann MF. Thirteen British cases of erythema chronicum migrans, a spirochaetal disease. *Br J Dermatol.* 1984;111(3):335–9

5 Haywood GA, O'Connell S, Gray HH. Lyme carditis: a United Kingdom perspective. *Heart.* 1993;70(1):15–6

6 Parke A. From New to old England: The progress of Lyme disease. *Br Med J (Clin Res Ed).* 1987;294(6571):525–6

7 Williams D, Rolles CJ, White JE. Lyme disease in a Hampshire child - medical curiosity or beginning of an epidemic. *Br Med J (Clin Res Ed).* 1986;292(6535):1560-1

8 Muhlemann MF, Wright DJ. Emerging pattern of Lyme disease in the United Kingdom and Irish Republic. *Lancet.* 1987;1(8527):260–2

9 Lorenc, T, Jones-Diette, J, Blanchard, L, Sutcliffe K, Stansfield C, Harden M, et al. Incidence and surveillance of Lyme disease: systematic review and policy mapping. 2017. London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London

10 Rizzoli A, Hauffe HC, Carpi G, Vourc'h GI, Neteler M, Rosà R. Lyme borreliosis in Europe. *Euro Surveill.* 2011;16(27):pii=19906

11 Dawkins R. *The Selfish Gene*. Oxford: Oxford University Press; 1976.

12 Usborne S. How afraid should we be of ticks and Lyme disease? *Guardi*an. 2017. Available from: https://www.theguardian.com/lifeandstyle/2017/aug/22/ticks-lyme-disease-matt-dawson-harm. [Accessed 1st January 2019]

13 Elsheikha H. Lyme disease is a ticking time bomb. *Mail Online*. 2016. Available from:

http://www.dailymail.co.uk/health/article-3563844/Lyme-disease-ticking-time-bomb-Leading-expert-explains-life-wrecking-illness-spreading-protect-yourself. [Accessed 1st January 2019]

14 ABC/AAP. Lyme disease does not exist in Australia, researchers say. *ABC News.* 2016. Available from: http://www.abc.net.au/news/2016-10-31/lyme-disease-doesnt-exist-in-australia,-researchers-say/7979158. [Accessed 1st January 2019]

15 Collignon PJ, Lum GD, Robson JMB. Does lyme disease exist in Australia? *Med J Aust.* 2016;205(9):413–7

16 Afzelius A. Erythema chronicum migrans. *Acta Dermato-Venereologia.* 1921;2:120–5

17 Hollstrom E. Successful treatment of erythema migrans afzelius. *Acta Dermato-Venereologia.* 1951;31(2):235–43

18 Stanek G, Wormser GP, Gray J, Strle F. Lyme borreliosis. *Lancet.* 2012;379(9814):461–73

19 Keller A, Graefen A, Ball M, Matzas M, Boisguerin V, Maixner F, et al. New insights into the Tyrolean Iceman's origin and phenotype as inferred by whole-genome sequencing. *Nat Commun.* 2012;3:698

20 Walter KS, Carpi G, Caccone A, Diuk-Wasser MA. Genomic insights into the ancient spread of Lyme disease across North America. *Nat Ecol Evol.* 2017;1(1):1569–76

21 Steere AC, Strle F, Wormser GP, Hu LT, Branda JA, Hovius JW, et al. Lyme borreliosis. *Nat Rev Dis Prim.* 2016;2:16090

22 Stokes G, Blanchard L, Sutcliffe K, Dickson K, Brunton G, Burchett H, et al. A systematic evidence map of research on Lyme disease in humans. 2017. London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University College London

23 National Institute for Health and Care Excellence. Lyme disease NICE guidelines [NG95]. 2018. Available from: https://www.nice.org.uk/guidance/ng95. [Accessed 1st January 2019]

24 Cull B, Pietzsch ME, Hansford KM, Gillingham EL, Medlock JM. Surveillance of British ticks: An overview of species records, host associations, and new records of Ixodes

ricinus distribution. *Ticks Tick Borne Dis.* 2018;9(3):605–14

25 Mysterud A, Easterday WR, Stigum VM, Aas AB, Meisingset EL, Vijugrein H. Contrasting emergence of Lyme disease across ecosystems. *Nat Commun.* 2016;7:11882

26 Medlock JM, Hansford KM, Bormane A, Derdakova M, Estrada-Peña A, George JC, et al. Driving forces for changes in geographical distribution of Ixodes ricinus ticks in Europe. *Parasit Vectors.* 2013;6:1

27 Scharlemann JP, Johnson PJ, Smith AA, Macdonald DW, Randolph SE. Trends in ixodid tick abundance and distribution in Great Britain. *Med Vet Entomol.* 2008;22(3):238–47

28 Jameson LJ, Medlock JM. Tick surveillance in Great Britain. *Vector Borne Zoonotic Dis.* 2011;11(4):403–12

29 Mead PS. Epidemiology of Lyme Disease. *Infect Dis Clin North Am.* 2015;29(2):187–210

30 Cook MJ. Lyme borreliosis: a review of data on transmission time after tick attachment. *Int J Gen Med.* 2014;8:1–8

31 Eisen L. Pathogen transmission in relation to duration of attachment by Ixodes scapularis ticks. *Ticks Tick Borne Dis.* 2018;9(3):535–42

32 Hofhuis A, van de Kassteele J, Sprong H, van den Wijngaard CC, Harms MG, Fonville M, *et al.* Predicting the risk of Lyme borreliosis after a tick bite, using a structural equation model. *PLoS One.* 2017;12(7):e0181807

33 Centers for Disease Control and Prevention. Lyme disease charts and figures:historical data. *CDC.* 2017. Available from: https://www.cdc.gov/lyme/stats/graphs.html. [Accessed 1st January 2019]

34 Abdullah S, Helps C, Tasker S, Newbury H, Wall R. Prevalence and distribution of *Borrelia* and *Babesia* species in ticks feeding on dogs in the U.K. *Med Vet Entomol.* 2018;32(1):14-22

35 Davies S, Abdullah S, Helps C, Tasker S, Newbury H, Wall R. Prevalence of ticks and tick-borne pathogens: Babesia and Borrelia species in ticks infesting cats of Great

Britain. *Vet Parasitol.* 2017;224:129–35

36 Hansford KM, Fonville M, Gillingham EL, Coipan EC, Pietzsch ME, Krawczyk AI, et al. Ticks and Borrelia in urban and peri-urban green space habitats in a city in southern England. *Ticks Tick Borne Dis.* 2017;8(3):353–61

37 Hall JL, Alpers K, Bown KJ, Martin SJ, Birtles RJ. Use of mass-participation outdoor events to assess human exposure to tickborne pathogens. *Emerg Infect Dis.* 2017;23(3):463–7

38 James MC, Gilbert L, Bowman AS, Forbes KJ. The heterogeneity, distribution, and environmental associations of Borrelia burgdorferi Sensu Lato, the agent of Lyme borreliosis, in Scotland. *Front Public Heal.* 2014;2:129

39 British Infection Association. The epidemiology, prevention, investigation and treatment of Lyme borreliosis in United Kingdom patients: A position statement by the British Infection Association. *J Infect.* 2011;62(5):329–38

40 Stanek G, Fingerle V, Hunfeld KP, Jaulhac B, Kaiser R, Krause A, et al. Lyme borreliosis: clinical case definitions for diagnosis and management in Europe. *Clin Microbiol Infect.* 2011;17(1):69–79

41 Leeflang MM, Ang CW, Berkhout J, Bijlmer HA, Van Bortel W, Brandenburg AH, et al. The diagnostic accuracy of serological tests for Lyme borreliosis in Europe: a systematic review and meta-analysis. *BMC Infect Dis.* 2016;16:140

42 Waddell LA, Greig J, Mascarenhas M, Harding S, Lindsay R, Ogden N. The accuracy of diagnostic tests for lyme disease in humans, a systematic review and meta-analysis of North American research. *PLoS One.* 2016;11(12):e0168613

43 Wormser GP, Dattwyler RJ, Shapiro ED, Halperin JJ, Steere AC, Klempner MS, et al. The clinical assessment, treatment, and prevention of Lyme disease, human granulocytic anaplasmosis, and babesiosis: clinical practice guidelines by the Infectious Diseases Society of America. *Clin Infect Dis.* 2006;43(9):1089–134

44 Dessau RB, van Dam AP, Fingerle V, Gray J, Hovius JW, Hunfeld KP, et al. To test or not to test? Laboratory support for the diagnosis of Lyme borreliosis: a position paper of ESGBOR, the ESCMID study group for Lyme borreliosis. *Clin Microbiol Infect.* 2018;24(2):118-124

45 Lipsker D, Lieber-Mbomeyo A, Hedelin G. How accurate is a clinical diagnosis of erythema chronicum migrans? Prospective study comparing the diagnostic accuracy of general practitioners and dermatologists in an area where Lyme borreliosis Is endemic. *Arch Dermatol.* 2004;140(5):620–1

46 Lieber-Mbomeyo A, Hedelin G, Lipsker D. The level of knowledge of general practitioners regarding the early phase of Lyme borreliosis. Survey conducted among 106 general practitioners. *Presse Med.* 2003;32:1734–6

47 Munro H, Mavin S, Duffy K, Evans R, Jarvis LM. Seroprevalence of Lyme borreliosis in Scottish blood donors. *Transfus Med.* 2015;25(4):284-6

48 Cerar D, Cerar T, Ružić-Sabljić E, Wormser GP, Strle F. Subjective symptoms after treatment of early Lyme disease. *Am J Med.* 2010;123(1):79–86

49 Cairns V, Godwin J. Post-Lyme borreliosis syndrome: a meta-analysis of reported symptoms. *Int J Epidemiol.* 2005;34(6):1340–5

50 Klempner MS, Baker PJ, Shapiro ED, Marques A, Dattwyler RJ, Halperin JJ, et al. Treatment trials for post-Lyme disease symptoms revisited. *Am J Med.* 2013;126(8):665–9

51 Krupp LB, Hyman LG, Grimson R, Coyle PK, Melville P, Ahnn S, et al. Study and treatment of post Lyme disease (STOP-LD): a randomized double masked clinical trial. *Neurology.* 2003;60(12):1923–30

52 Dersch R, Sommer H, Rauer S, Meerpohl JJ. Prevalence and spectrum of residual symptoms in Lyme neuroborreliosis after pharmacological treatment: a systematic review. *J Neurol.* 2016;263(1):17–24

53 Halperin JJ. Chronic Lyme disease: misconceptions and challenges for patient management. *Infect Drug Resist.* 2015;8:119–28

54 Lantos PM. Chronic Lyme disease: the controversies and the science. *Expert Rev Anti Infect Ther.* 2011;9(7):787–97

55 Rebman AW, Aucott JN, Weinstein ER, Bechtold KT, Smith KC, Leonard L. Living in Limbo: Contested narratives of patients with chronic symptoms following Lyme disease. *Qual Health Res.* 2015;27(4):534-546.

56 Berende A, ter Hofstede HJ, Vos FJ, van Middendorp H, Vogelaar ML, Tromp M, et al. Randomized trial of longer-term therapy for symptoms attributed to Lyme disease. *N Engl J Med.* 2016;374(13):1209–20

57 Melia MT, Auwaerter PG. Time for a different approach to Lyme disease and long-term symptoms. *N Engl J Med.* 2016;374(13):1277-80

58 Schneeberger PM, Wintenberger C, van der Hoek W, Stahl JP. Q fever in the Netherlands - 2007-2010: what we learned from the largest outbreak ever. *Med Mal Infect.* 2014;44(8):339–53

59 Kampschreur LM, Delsing CE, Groenwold RH, Wegdam-Blans MC, Bleeker-Rovers CP, de Jager-Leclercq MG, et al. Chronic Q fever in the Netherlands 5 years after the start of the Q fever epidemic: results from the Dutch chronic Q fever database. *J Clin Microbiol.* 2014;52(5):1637–43

60 Wegdam-Blans MCA, Tjhie HT, Korbeeck JM, Nabuurs-Franssen MN, Kampschreur LM, Sprong T, et al. Serology in chronic Q fever is still surrounded by question marks. *Eur J Clin Microbiol Infect Dis.* 2014;33(7):1089–94

61 Kampschreur LM, Wegdam-Blans MC, Wever PC, Renders NH, Delsing CE, Sprong T, et al. Chronic Q fever diagnosis - consensus guideline versus expert opinion. *Emerg Infect Dis.* 2015;21(7):1183–8

62 Dunmire SK, Hogquist KA, Balfour HH. Infectious mononucleosis. *Curr Top Microbiol Immunol.* 2015;390(1):211–40

63 Schotthoefer AM, Frost HM. Ecology and epidemiology of Lyme borreliosis. *Clin Lab Med.* 2016;35(4):723–43

64 Nelson CA, Saha S, Kugeler KJ, Delorey MJ, Shankar MB, Hinckley AF, et al. Incidence of clinician-diagnosed Lyme disease, United States, 2005-2010. 2015;21(9):1625-31

65 Centers for Disease Control and Prevention. Lyme disease. *CDC*. 2018. Available from: https://www.cdc.gov/lyme/stats/survfaq.html. [Accessed 1st January 2019]

66 Sykes RA, Makiello P. An estimate of Lyme borreliosis incidence in Western Europe. *J Public Health (Oxf).* 2017;39(1):74-81

67 Wilking H, Stark K. Trends in surveillance data of human Lyme borreliosis from six

federal states in eastern Germany, 2009-2012. *Ticks Tick Borne Dis.* 2014;5(3):219–24

68 www.parliament.uk. Publications & records. 2018. Available from: https://www.parliament.uk/business/publications/. [Accessed 1st January 2019]

69 UK Government and Parliament. Petitions. 2018. Available from: https://petition.parliament.uk/. [Accessed 1st January 2019]

70 Public Health England. Zoonoses Overview Report UK 2016. 2017. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/664448/UK_Zoonoses_report_2016.pdf. [Accessed 1st January 2019]

71 Google. Google trends. 2018. Available from: https://trends.google.com/trends/. [Accessed 1st January 2019]

72 Brunton G, Sutcliffe K, Hinds K, Khatwa M, Burchett H, Dickson K, et al. Stakeholder experience of the diagnosis of Lyme disease: a systematic review. 2017. London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University of London.

73 Sutcliffe K, Sowden A, Burchett H, Stansfield C, Harden M, Thomas J. Patient, clinician and researcher experiences of the treatment and management of Lyme diseas: a systematic review. 2017. London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University of London.

74 Richardson M, Khouja C, Walker R, Sutcliffe K, Stansfield C, Harden M, et al. Prevention interventions for Lyme disease: a systematic review. 2017. London: EPPI-Centre, Social Science Research Unit, UCL Institute of Education, University of London.

75 Evans R, Mavin S, Ho-Yen DO. Audit of the laboratory diagnosis of Lyme disease in Scotland. *J Med Microbiol.* 2005;54(2):1139–41

76 Mavin S, Watson EJ, Evans R. Distribution and presentation of Lyme borreliosis in Scotland – analysis of data from a national testing laboratory. *J R Coll Physicians Edinb.*2015;45(3)196–200

77 The Health Protection (Notification) Regulations 2010, No.659. 2010. Available from: http://www.legislation.gov.uk/uksi/2010/659/made. [Accessed 1st January 2019]

78 Cutler SJ, Ruzic-Sabljic E, Potkonjak A. Emerging borreliae – Expanding beyond Lyme borreliosis. *Mol Cell Probes.* 2017;31:22–7

79 Djennad A, Lo Iacono G, Sarran C, Fleming LE, Kessel A, Haines A, et al. A comparison of weather variables linked to infectious disease patterns using laboratory addresses and patient residence addresses. *BMC Infect Dis.* 2018;18(1):198

80 Elliot AJ, Bone A, Morbey R, Hughes HE, Harcourt S, Smith S, et al. Using real-time syndromic surveillance to assess the health impact of the 2013 heatwave in England. *Environ Res.* 2014;135:31–6

81 Health Protection Scotland. Reference laboratories. 2017. Available from: http://www.hps.scot.nhs.uk/reflab/STRL.aspx [Accessed 1st January 2019]

82 Health Protection Scotland. Lyme disease. 2017. Available from: http://www.hps.scot.nhs.uk/giz/lymedisease.aspx. [Accessed 1st January 2019]

83 Newman EN, Johnstone P, Bridge H, Wright D, Jameson L, Bosworth A, et al. Seroconversion for infectious pathogens among UK military personnel deployed to Afghanistan, 2008–2011. *Emerg Infect Dis.* 2014;20(12):2015–22

84 Adams K, Jameson L, Meigh R, Brooks T. Hantavirus: An infectious cause of acute kidney injury in the UK. *BMJ Case Rep.* 2014;pii:bcr2014205529

85 HSC Public Health Agency. Zoonoses (Infections acquired from animals). 2017. Available from: http://www.publichealth.hscni.net/directorate-public-health/health-protection/zoonoses-infections-acquired-animals. [Accessed 1st January 2019]

86 World Health Organisation. ICD-10 Version:2010. 2010. Available from: http://apps.who.int/classifications/icd10/browse/2010/en. [Accessed 1st January 2019]

87 NHS Digital. Hospital Episode Statistics. 2018. Available from: http://content.digital.nhs.uk/hes. [Accessed 1st January 2019]

88 Sinha S, Peach G, Poloniecki JD, Thompson MM, Holt PJ. Studies using english administrative data (Hospital Episode Statistics) to assess health-care outcomes-systematic review and recommendations for reporting. *Eur J Public Health.* 2013;23(1):86–92

89 Hobbelen PH, Stowe J, Amirthalingam G, Miller L, van Hoek AJ. The burden of hospitalisation for varicella and herpes zoster in England from 2004 to 2013. *J Infect.* 2016;73(3):241-53

90 Bowering K. Analysis of routine hospital administrative data (including hospital episode statistics) to assess variation in process and outcomes in gastroenterology [dissertation on the Internet].University of Liverpool; 2014. Available from: https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.617497. [Accessed 1st January 2019]

91 Spencer SA, Davies MP. Hospital episode statistics: improving the quality and value of hospital data: a national internet e-survey of hospital consultants. *BMJ Open.* 2012;2(6):e001651

92 Public Health Wales Observatory. Patient Episode Database for Wales (PEDW). 2017. Available from: http://www.wales.nhs.uk/sitesplus/922/page/50308. [Accessed 1st January 2019]

93 NHS Wales. Information and Statistics - PEDW Data Online. 2015. Available from: http://www.infoandstats.wales.nhs.uk/page.cfm?orgid=869&pid=40977. [Accessed 1st January 2019]

94 ISD Scotland. Acute Hospital Activity & NHS Beds Data Release. 2018. Available from: https://www.isdscotland.org/Health-Topics/Hospital-Care/Publications/Acute-Hospital-Publication/. [Accessed 1st January 2019]

95 Department of Health. Hospital Activity Statistics. 2017. Available from: https://www.health-ni.gov.uk/topics/dhssps-statistics-and-research/hospital-activity-statistics. [Accessed 1st January 2019]

96 Department of Health. Hospital statistics: emergency care activity 2017/2018. 2018. Available from: https://www.health-ni.gov.uk/publications/hospital-statistics-emergency-care-activity-201718. [Accessed 1st January 2019]

97 Wu TY, Jen MH, Bottle A, Molokhia M, Aylin P, Bell D, et al. Ten-year trends in hospital admissions for adverse drug reactions in England 1999-2009. *J R Soc Med.* 2010;103(6):239–50

98 Moxey PW, Hofman D, Hinchliffe RJ, Jones K, Thompson MM, Holt PJ.

Epidemiological study of lower limb amputation in England between 2003 and 2008. *Br J Surg.* 2010;97(9):1348–53

99 Long SJ, Fone D, Gartner A, Bellis MA. Demographic and socioeconomic inequalities in the risk of emergency hospital admission for violence: cross-sectional analysis of a national database in Wales. *BMJ Open.* 2016;6(8):e011169

100 Ismail SIMF, Puyk B. The rise of obstetric anal sphincter injuries (OASIS): 11-year trend analysis using Patient Episode Data base for Wales (PEDW) data. *J Obstet Gynaecol.* 2014;34(6):495–8

101 Benson T. The history of the Read Codes: The inaugural James Read memorial lecture 2011. *Inform Prim Care.* 2011;19(3):173–82

102 Kontopantelis E, Stevens RJ, Helms PJ, Edwards D, Doran T, Ashcroft DM. Spatial distribution of clinical computer systems in primary care in England in 2016 and implications for primary care electronic medical record databases: a cross sectional population study. *BMJ Open.* 2018;8(2):e020738

103 Kontopantelis E, Buchan I, Reeves D, Checkland K, Doran T. Relationship between quality of care and choice of clinical computing system: retrospective analysis of family practice performance under the UK's quality and outcomes framework. *BMJ Open.* 2013;3(8):e003190

104 CPRD. CPRD. 2018. Available from: https://www.cprd.com/. [Accessed 1st January 2019]

105 Ratib S, Fleming KM, Crooks CJ, Aithal GP, West J. 1 and 5 year survival estimates for people with cirrhosis of the liver in England, 1998-2009: A large population study. *J Hepatol.* 2014;60(2):282–9

106 Quint JK, Müllerova H, DiSantostefano RL, Forbes H, Eaton S, Hurst JR, et al. Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). *BMJ Open.* 2014;4(7):e005540

107 Thomas KH, Davies N, Metcalfe C, Windmeijer F, Martin RM, Gunnell D. Validation of suicide and self-harm records in the Clinical Practice Research Datalink. *Br J Clin Pharmacol.* 2013;76(1):145–57

108 In Practice Systems Ltd. The Health Improvement Network (THIN). 2018. Available

from: https://www.visionhealth.co.uk/portfolio-items/the-health-improvement-network-thin/. [Accessed 1st January 2019]

109 Wang Y, Hunt K, Nazareth I, Freemantle N, Petersen I. Do men consult less than women? An analysis of routinely collected UK general practice data. *BMJ Open.* 2013;3(8):e003320

110 Sharma M, Nazareth I, Petersen I. Trends in incidence, prevalence and prescribing in type 2 diabetes mellitus between 2000 and 2013 in primary care: a retrospective cohort study. *BMJ Open.* 2016;6(1):e010210

111 University of Nottingham. QResearch. 2018. Available from: http://www.qresearch.org/. [Accessed 1st January 2019]

112 Vezyridis P, Timmons S. Evolution of primary care databases in UK: A scientometric analysis of research output. *BMJ Open.* 2016;6(10):e012785

113 Coupland C, Hill T, Morriss R, Moore M, Arthur A, Hippisley-Cox J. Antidepressant use and risk of adverse outcomes in people aged 20-64 years: cohort study using a primary care database. *BMC Med.* 2018;16(1):36

114 Vinogradova Y, Coupland C, Hippisley-Cox J. Risk of pneumonia in patients taking statins: population-based nested case-control study. *Br J Gen Pract.* 2011;61(592):e742-8

115 ResearchOne. ResearchOne. 2018. Available from: http://www.researchone.org/. [Accessed 1st January 2019]

116 Burr NE, Smith C, West R, Hull MA, Subramanian V. Increasing prescription of opiates and mortality in patients with Inflammatory Bowel Diseases in England. *Clin Gastroenterol Hepatol.* 2018;16(4):534–41

117 Harcourt S, Morbey RA, Bates C, Carter H, Ladhini SN, de Lusignan S, et al. Estimating primary care attendance rates for fever in infants after meningococcal B vaccination in England using national syndromic surveillance data. *Vaccine.* 2018;36(4):565–71

118 Blak BT, Thompson M, Dattani H, Bourke A. Generalisability of The Health Improvement Network (THIN) database: Demographics, chronic disease prevalence and mortality rates. *Inform Prim Care.* 2011;19(4):251–5

119 Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). *Int J Epidemiol.* 2015;44(3):827–36

120 Hippisley-Cox J, Vingradova Y, Coupland C, Pringle M. Comparison of key practice characteristics between general practices in England and Wales and general practices in the QRESEARCH database. 2005. Available from: https://www.nhsemployers.org/~/media/Employers/Documents/Primary%20care%20contracts/GMS/GMS%20Finance/Global%20Sum/fr72_characteristics_qresearch_practices_database_cd_201105.pdf. [Accessed 1st January 2019]

121 Tam CC, Rodrigues LC, Viviani L, Dodds JP, Evans MR, Hunter PR, et al. Longitudinal study of infectious intestinal disease in the UK (IID2 study): incidence in the community and presenting to general practice. *Gut.* 2012;61(1):69–77

122 Stockwell MS, Reed C, Vargas CY, Camargo S, Garretson AF, Alba LR, et al. MoSAIC: Mobile surveillance for acute respiratory infections and influenza-like illness in the community. *Am J Epidemiol.* 2014;180(12):1196–1201

123 Nuti SV, Wayda B, Ranasinghe I, Wang S, Dreyer RP, Chen SI, et al. The use of Google trends in health care research: a systematic review. *PLoS One.* 2014;9(10):e109583

124 Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. *Science.* 2014;343(6176):1203–5

125 Velasco E, Aghenza T, Denecke K, Kirchner G, Eckmanns T. Social media and internet-based data in global systems for public health surveillance: a systematic review. *Milbank Q.* 2014;92(1):7–33

126 Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *J Med Internet Res.* 2013;15(7):e147

127 Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EH, Olsen JM, et al. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PLoS One.* 2015;10(10):e0139701

128 Gu H, Chen B, Zhu H, Jiang T, Wang X, Chen L, et al. Importance of internet surveillance in public health emergency control and prevention: evidence from a

digital epidemiologic study during avian influenza A H7N9 outbreaks. *J Med Internet Res.* 2014;16(1):e20

129 Denecke K, Krieck M, Otrusina L, Smrz P, Dolog P, Nejdl W, et al. How to exploit twitter for public health monitoring? *Methods Inf Med.* 2013;52(4):326–39

130 socialmedia. Most popular social networks in the UK. 2018. Available from: https://social-media.co.uk/list-popular-social-networking-websites. [Accessed 1st January 2019]

131 Pandey A, Patni N, Singh M, Sood A, Singh G. YouTube as a source of information on the H1N1 influenza pandemic. *Am J Prev Med.* 2010;38(3):e1–3

132 Dubey D, Amritphale A, Sawhney A, Dubey D, Srivastav N. Analysis of YouTube as a source of information for West Nile Virus infection. *Clin Med Res.* 2014;12(3-4):129–32

133 Velardi P, Stilo G, Tozzi AE, Gesualdo F. Twitter mining for fine-grained syndromic surveillance. *Artif Intell Med.* 2014;61(3):153–63

134 Nagar R, Yuan Q, Freifeld CC, Santillana M, Nojima A, Chunara R, et al. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. *J Med Internet Res.* 2014;16(10):e236

135 Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. *PLoS One.* 2011;6(5):e19467

136 Salathé M, Vu DQ, Khandelwal S, Hunter DR. The dynamics of health behavior sentiments on a large online social network. *EPJ Data Sci.* 2013;2(4):1–12

137 Abdullah S, Helps C, Tasker S, Newbury H, Wall R. Ticks infesting domestic dogs in the UK: a large-scale surveillance programme. *Parasit Vectors.* 2016;9(1):391

138 Goossens HA, van den Bogaard AE, Nohlmans MK. Dogs as sentinels for human Lyme borreliosis in The Netherlands. *J Clin Microbiol.* 2001;39(3):844–8

139 Smith FD, Ballantyne R, Morgan ER, Wall R. Estimating Lyme disease risk using pet dogs as sentinels. *Comp Immunol Microbiol Infect Dis.* 2012;35(2):163–7

140 Herrmann JA; Dahm NM, Ruiz MO, Brown WM. Temporal and spatial distribution of tick-borne disease cases among humans and canines in Illinois (2000-2009). *Environ Health Insights.* 2009;8(2):15-27

141 Jennett AL, Smith FD, Wall R. Tick infestation risk for dogs in a peri-urban park. *Parasit Vectors.* 2013;6:358

142 Tulloch JSP, McGinley L, Sánchez-Vizcaíno F, Medlock JM, Radford AD. The passive surveillance of ticks using companion animal electronic health records. *Epidemiol Infect.* 2017;145(10):2020-2029

143 University of Liverpool. Small Animal Veterinary Surveillance Network (SAVSNET). 2018. Available from: https://www.liverpool.ac.uk/savsnet/. [Accessed 1st January 2019]

144 Arsevska E, Singleton D, Sánchez-Vizcaíno F, Williams N, Jones PH, Smyth S, et al. Small animal disease surveillance: GI disease and salmonellosis. *Vet Rec.* 2017;181(9):228–32

145 Sánchez-Vizcaíno F, Wardeh M, Heayns B, Singleton DA, Tulloch JSP, McGinley L, et al. Canine babesiosis and tick activity monitored using companion animal electronic health records in the UK. *Vet Rec.* 2016;179(14):358

146 Health Research Authority. Defining research. 2013. Available from: https://researchsupport.admin.ox.ac.uk/sites/default/files/researchsupport/documents/media/defining-research.pdf. [Accessed 1st January 2019]

147 Health Resarch Authority. Defining research table. 2017. Available from: http://www.hra-decisiontools.org.uk/research/docs/DefiningResearchTable_Oct2017-1.pdf. [Accessed 1st January 2019]

148 The Health Protection (Notification) Regulations 2010, No.659. 2010. Available from: http://www.legislation.gov.uk/uksi/2010/659/made. [Accessed 1st January 2019]

149 University of Liverpool. Research Support Office. Available from: https://www.liverpool.ac.uk/research-support-office/. [Accessed 1st January 2019]

150 Public Health England. Laboratory reporting to Public Health England. 2016. Available from:

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/739854/PHE_Laboratory_Reporting_Guidelines.pdf. [Accessed 1st January 2019]

152 Sanchez E, Vannier E, Wormser GP, Hu LT. Diagnosis, treatment, and prevention of Lyme disease, human granulocytic anaplasmosis, and babesiosis: a review. *JAMA.* 2016;315(16):1767–77

153 gov.uk. Zoonoses: UK annual reports. 2018. Available from: https://www.gov.uk/government/publications/zoonoses-uk-annual-reports [Accessed 1st January 2019]

154 Smith R, O'Connell S, Palmer S. Lyme disease surveillance in England and Wales, 1986 1998. *Emerg Infect Dis.* 2000;6(4):404–7

155 Ratnapradipa D, McDaniel JT, Barger A. Social vulnerability and Lyme disease incidence: A regional analysis of the United States, 2000-2014. *Epidemiol Biostat Public Heal* 2017;14(2):1–12

156 Jackson LE, Hilborn ED, Thomas JC. Towards landscape design guidelines for reducing Lyme disease risk. *Int J Epidemiol.* 2006;35(2):315–22

157 Jackson LE, Levine JF, Hilborn ED. A comparison of analysis units for associating Lyme disease with forest-edge habitat. *Community Ecol.* 2006;7(2):189–97

158 gov.uk. Lyme disease test request form. 2018. Available from: https://www.gov.uk/government/publications/lyme-disease-test-request-form [Accessed 1st January 2019]

159 Office for National Statistics. People, population and community. 2018. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity. [Accessed 1st January 2019]

160 gov.uk. Public Health England regions, local centres and emergency contacts. 2015. Available from: https://www.gov.uk/government/collections/contacts-public-health-england-regions-local-centres-and-emergency [Accessed 1st January 2019]

161 Dudani S. The distance-weighted k-nearest-neighbor rule. *IEEE Trans Syst Man Cybern.* 1976;6(4):325–7

162 Enas GG, Choi SC. Choice of the smoothing parameter and efficiency of k-nearest neighbor classification. *Comput Math with Appl.* 1986;12(2):235–44

163 Cuzick J, Edwards R. Methods for investigating localized clustering of disease. Clustering methods based on k nearest neighbour distributions. *IARC Sci Publ.* 1996;(135):53–67

164 Anselin L. Local Indicators of Spatial Association—LISA. *Geogr Anal.* 1995;27(2):93–115

165 Anselin L, Syabri I, Smirnov O. Visualizing multivariate spatial correlation with dynamically linked windows. In Anselin L(ed). *New Tools for Spatial Data Analysis.* (2002) Santa Barbara: University of California, 21p.

166 Lloyd C. *Spatial Data Analysis: An Introduction for GIS Users*. (2010) Oxford: Oxford University Press.

167 gov.uk. English Indices of Deprivation 2015. 2015. Available from: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015 [Accessed 14th July 2017]

168 Welsh Government. Welsh Index of Multiple Deprivation. 2015. Available from: http://gov.wales/statistics-and-research/welsh-index-multiple-deprivation/?lang=en [Accessed 14th Jul 2017]

169 gov.uk. 2011 Rural Urban Classification. 2017. Available from: https://www.gov.uk/government/statistics/2011-rural-urban-classification [Accessed 1st January 2019]

170 Public Health England. Infection Report Volume 11 Number 6. 2017. Available from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/593004/hpr0617_zoos.pdf [Accessed 1st January 2019]

171 Halsby K. *Lyme data capture for RIPL.* Public Health England: Internal Report. 2016

172 The Doctors Laboratory. The Doctors Laboratory. 2019. Available from: https://tdlpathology.com/ [Accessed 1st Jan 2019]

173 Nelson C, Banks S, Jeffries CL, Walker T, Logan JG. Tick abundances in South London parks and the potential risk for Lyme borreliosis to the general public. *Med Vet*

Entomol. 2015;29(4):448–52

174 Public Health England. Lyme borreliosis epidemiology and surveillance. 2018. Available from: https://www.gov.uk/government/publications/lyme-borreliosis-epidemiology/lyme-borreliosis-epidemiology-and-surveillance. [Accessed 1st January 2019]

175 Medlock JM, Leach SA. Effect of climate change on vector-borne disease risk in the UK. *Lancet Infect Dis.* 2015;15(6):721–30

176 Sajanti E, Virtanen M, Helve O, Kuusi M, Lyytikäinen O, Hytönen J, et al. Lyme borreliosis in Finland in 1995-2014. *Emerg Infect Dis.* 2017;23(8):128–1288

177 Eliassen KE, Berild D, Reiso H, Grude N, Christophersen KS, Finckenhagen C, et al. Incidence and antibiotic treatment of erythema migrans in Norway 2005–2009. *Ticks Tick Borne Dis.* 2017;8(1):1-8

178 Rizzoli A, Hauffe H, Carpi G, Vourc HG, Neteler M, Rosa R. Lyme borreliosis in Europe. *Euro Surveill.* 2011;16(27):pii19906

179 Cooper L, Branagan-Harris M, Tuson R, Nduka C. Lyme disease and Bell's palsy: an epidemiological study of diagnosis and risk in England. *Br J Gen Pract.* 2017;67(658):e329-e335

180 Vandenesch A, Turbelin C, Couturier E, Arena C, Jaulhac B, Ferquel E, et al. Incidence and hospitalisation rates of Lyme borreliosis, France, 2004 to 2012. *Euro Surveill.* 2014;19(34):pii:20883

181 Sorouri R, Ramazani A, Karami A, Ranjbar R, Guy EC. Isolation and characterization of Borrelia burgdorferi strains from Ixodes ricinus ticks in the southern England. *Bioimpacts*. 2015;5(2):71–8

182 Bettridge J, Renard M, Zhao F, Bown KJ, Birtles RJ. Distribution of Borrelia burgdorferi sensu lato in Ixodes ricinus populations across central Britain. *Vector Borne Zoonotic Dis.* 2013;13(3):139-46

183 Hosseinpoor AR, Stewart Williams JA, Itani L, Chatterji S. Socioeconomic inequality in domains of health: results from the World Health Surveys. *BMC Public Health.* 2012;12:198

184 Pasqualini M, Lanari D, Minelli L, Pieroni L, Salmasi L. Health and income inequalities in Europe: What is the role of circumstances? *Econ Hum Biol.* 2017;26:164–73

185 Marmot M, Friel S, Bell R, Houweling TA, Taylor S. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet.* 2008;372(9650):1661–9

186 Curry N, Ravenscroft N. Countryside recreation provision in England: exploring a demand-led approach. *Land Use Policy. 2001*;18(3):281–91

187 Horner JS, Horner JW. Do doctors read forms? A one-year audit of medical certificates submitted to a crematorium. *J R Soc Med.* 1998;91(7):371–6

188 Lovett JK, Evans PH, O'Connell S, Gutowski NJ. Neuroborreliosis in the South West of England. *Epidemiol Infect.* 2008;136(12):1707–11

189 Dryden MS, Saeed K, Ogborn S, Swales P. Lyme borreliosis in southern United Kingdom and a case for a new syndrome, chronic arthropod-borne neuropathy. *Epidemiol Infect.* 2015;143(3):561-72

190 Marcu A, Barnett J, Uzzell D, Vasileiou K, O'Connell S. Experience of Lyme disease and preferences for precautions: a cross-sectional survey of UK patients. *BMC Public Health.* 2013;13:481

191 Dillon R, O'Connell S, Wright S. Lyme disease in the UK: clinical and laboratory features and response to treatment. *Clin Med (Lond).* 2010;10(5):454–7

192 Cottle LE, Mekonnen E, Beadsworth MB, Miller AR, Beeching NJ. Lyme disease in a British referral clinic. *QJM.* 2012;105(6):537–43

193 Office for National Statistics. Ethnicity and national identity in England and Wales: 2011. 2012. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/ethnicityandnationalidentityinenglandandwales/2012-12-11 [Accessed 1st January 2019]

194 Lohr B, Müller I, Mai M, Norris DE, Schöffski O, Hunfeld KP. Epidemiology and cost of hospital care for Lyme borreliosis in Germany: lessons from a health care utilization database analysis. *Ticks Tick Borne Dis.* 2015;6(1):56–62

195 Jivraj S, Khan O. Ethnicity and deprivation in England: How likely are ethnic minorities to live in deprived neighbourhoods? *Dyn Divers Evid from 2011 Census*. 2013. Available from: http://hummedia.manchester.ac.uk/institutes/code/briefingsupdated/ethnicity-and-deprivation-in-england-how-likely-are-ethnic-minorities-to-live-in-deprived-neighbourhoods%20(1).pdf. [Accessed 1st January 2019]

196 Mowbray F, Amlôt R, Rubin GJ. Predictors of protective behaviour against ticks in the UK: A mixed methods study. *Ticks Tick Borne Dis.* 2014;5(4):392–400

197 Cowling TE, Soljak MA, Bell D, Majeed A. Emergency hospital admissions via accident and emergency departments in England: time trend, conceptual framework and policy implications. *J R Soc Med.* 2014;107(11):432–8

198 Marcu A, Uzzell D, Barnett J. Making sense of unfamiliar risks in the countryside: The case of Lyme disease. *Heal Place.* 2011;17(3):843–50

199 Cecil E, Bottle A, Cowling TE, Majeed A, Wolfe I, Saxena S. Primary care access, emergency department visits, and unplanned short hospitalizations in the UK. *Pediatrics.* 2016;137(2):e20151492

200 Zhou Y, Abel G, Warren F, Roland M, Campbell J, Lyratzopoulos G, et al. Do difficulties in accessing in-hours primary care predict higher use of out-of-hours GP services? Evidence from an English National Patient Survey. *Emerg Med J.* 2015;32(3):373-8

201 O'Keeffe C, Mason S, Jacques R, Nicholl J. Characterising non-urgent users of the emergency department (ED): A retrospective analysis of routine ED data. *PLoS One.* 2018;13(2):e0192855

202 Rosenberg-Wohl S, Greenfield G, Majeed A, Hayhoe B. Seven-day access to NHS primary care: how does England compare with Europe? *J R Soc Med.* 2018;111(3):88-91

203 Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, et al. Systematic review of discharge coding accuracy. *J Public Health (Oxf).* 2012;34(1):138–48

204 NHS. Accident and Emergency Diagnosis Tables. 2018. Available from: http://www.datadictionary.nhs.uk/web_site_content/supporting_information/clinic

al_coding/accident_and_emergency_diagnosis_tables.asp?shownav=1. [Accessed 1st January 2019]

205 Newitt S, Ellliot AJ, Morbey RA, Durnall H, Pietzsch M, Medlock JM, et al. The use of syndromic surveillance to monitor the incidence of arthropod bites requiring healthcare in England, 2000–2013: a retrospective ecological study. *Epidemiol Infect.* 2016;144(11):2251–9

206 Baker PJ. Straight talk about chronic Lyme disease. *Am J Med.* 2018;131(6):592-594

207 van den Wijngaard CC, Hofhuis A, Wong A, Harms MG, de Wit GA, Lugner AK, et al. The cost of Lyme borreliosis. *Eur J Public Health.* 2017;27(3):538–47

208 Pham T, Morris TP, Petersen I. Ethnicity recording in primary care: multiple imputation of missing data in ethnicity recording using The Health Improvement Network (THIN) database. *United Kingdom Stats Users' Group Meetings 2015.* 2015. Available from: http://repec.org/usug2015/pham_uksug15.pdf

209 Office for National Statistics. UK censuses data. 2018. Available from:https://www.ons.gov.uk/census/2011census/2011ukcensuses/ukcensusesdata [Accessed 1st January 2019]

210 Townsend P. Deprivation. *Journal of Social Science.* 1987;16:125–46

211 INPS. Read Dictionary Changes Q3 2014. 2014. Available from: http://www.inps.co.uk/sites/default/files/Read Q3 2014 Changes.pdf. [Accessed 1st January 2019]

212 Altpeter E, Zimmermann H, Oberreich J, Peter O, Dvorak C, Swiss Sentinel Surveillance Network. Tick related diseases in Switzerland, 2008 to 2011. *Swiss Med Wkly.* 2013;143:w13725

213 Hofhuis A, Harms M, Bennema S, van den Wijngaard CC, van Pelt W. Physician reported incidence of early and late Lyme borreliosis. *Parasit Vectors.* 2015;8:161

214 Hansford KM, Fonville M, Jahfari S, Sprong H, Medlock JM. Borrelia miyamotoi in host-seeking Ixodes ricinus ticks in England. *Epidemiol Infect.* 2015;143(5):1079–87

215 Rauter C, Hartung T. Prevalence of Borrelia burgdorferi sensu lato genospecies in Ixodes ricinus ticks in Europe: a metaanalysis. *Appl Environ Microbiol.*

2005;71(11):7203–16

216 Jore S, Vanwambeke SO, Viljugrein H, Isaksen K, Kristoffersen AB, Woldehiwet Z, et al. Climate and environmental change drives Ixodes ricinus geographical expansion at the northern range margin. *Parasites and Vectors.* 2014;7:11

217 Natural England. *Monitor of Engagement with the Natural Environment Spatial Report 2009 - 2011*. 2011. Available from: http://publications.naturalengland.org.uk/file/209963 [Accessed 1st January 2019]

218 Scottish Natural Heritage. Scotland's People and Nature Survey 2013/2014 Special Interest Report No. 1. 2014. Available from: https://www.nature.scot/sites/default/files/2017-07/A1480507%20-%20SPANS%20-%20reporting%20-%20201314%20-%20special%20interest%20report%20series%20-%20Outdoor%20recreation.pdf. [Accessed 1st January 2019]

219 Nelson CA, Starr JA, Kugeler KJ, Mead PS. Lyme disease in Hispanics, United States, 2000-2013. *Emerg Infect Dis.* 2016;22(3):522–5

220 Fix AD, Peña CA, Strickland GT. Racial differences in reported Lyme disease incidence. *Am J Epidemiol.* 2000;152(8):756–9

221 Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol.* 2010;69(1):4–14

222 Marques A. Chronic Lyme disease: a review. *Infect Dis Clin North Am.* 2008;22(2):341–60

223 Khan NF, Harrison SE, Rose PW. Validity of diagnostic coding within the General Practice Research Database: a systematic review. *Br J Gen Pract.* 2010;60(572):e128-36

224 British Medical Association. *2014/15 General Medical Services (GMS) contract Quality and Outcomes Framework (QOF).* 2014. Available from: http://bma.org.uk/-/media/files/pdfs/practical advice at work/contracts/gpqofguidance2014-15.pdf [Accesed 1st January 2019]

225 Russell P, Banerjee S, Watt J, Adleman R, Agoe B, Burnie N, et al. Improving the

identification of people with dementia in primary care: evaluation of the impact of primary care dementia coding guidance on identified prevalence. *BMJ Open.* 2013;3(12):e004023

226 Herrett E, Shah AD, Boggon R, Denaxas S, Smeeth L, van Staa T, et al. Completeness and diagnostic validity of recording acute myocardial infarction events in primary care, hospital care, disease registry, and national mortality records: cohort study. *BMJ.* 2013;346:f2350

227 Hammad TA, Margulis AV, Ding Y, Strazzeri MM, Epperly H. Determining the predictive value of Read codes to identify congenital cardiac malformations in the UK Clinical Practice Research Datalink. *Pharmacoepidemiol Drug Saf.* 2013;22(11):1233–8

228 Kang EM, Pinheiro SP, Hammad TA, Abou-Ali A. Evaluating the validity of clinical codes to identify cataract and glaucoma in the UK Clinical Practice Research Datalink. *Pharmacoepidemiol Drug Saf.* 2015;24(1):38–44

229 Harkness EF, Grant L, O'Brien SJ, Chew-Graham CA, Thompson DG. Using read codes to identify patients with irritable bowel syndrome in general practice: a database study. *BMC Fam Pract.* 2013;14:183

230 Harkness EF, Harrington V, Hinder S, O'Brien SJ, Thompson DG, Beech P, et al. GP perspectives of irritable bowel syndrome - an accepted illness, but management deviates from guidelines: a qualitative study. *BMC Fam Pract.* 2013;14:92

231 de Lusignan S. The barriers to clinical coding in general practice : A literature review. *Med Inform Internet Med*. 2005;30(2):89-97

232 de Lusignan S, Wells SE, Hague NJ, Thiru K. Managers see the problems associated with coding clinical data as a technical issue whilst clinicians also see cultural barriers. *Methods Inf Med*. 2003;42(4):416–22

233 Cresswell K, Morrison Z, Kalra D, Sheikh A. "There are too many, but never enough": qualitative case study investigating routine coding of clinical information in depression. *PloS One*. 2012;7(8):e43831

234 Michiels-Corsten M, Donner-Banzhoff N. Beyond accuracy: hidden motives in diagnostic testing. *Fam Pract*. 2018;35(2):222-227

235 Vandererven C, Bellanger AP, Faucher JF, Marguet P. Primary care physician management of tick bites in the Franche-Comté region (Eastern France, 2013). *Med Mal Infect.* 2017;47(4):261–5

236 Vanthomme K, Bossuyt N, Boffin N, Van Casteren V. Incidence and management of presumption of Lyme borreliosis in Belgium: recent data from the sentinel network of general practitioners. *Eur J Clin Microbiol Infect Dis.* 2012;31(9):2385–90

237 Coumou J, Hovius JW, van Dam AP. Borrelia burgdorferi sensu lato serology in the Netherlands: guidelines versus daily practice. *Eur J Clin Microbiol Infect Dis.* 2014;33(10):1803–8

238 Eppes SC, Klein JD, Caputo GM, Rose CD. Physician beliefs, attitudes, and approaches toward Lyme disease in an endemic area. *Clin Pediatr (Phila).* 1994;33(3):130–4

239 Brett ME, Hinckley AF, Zielinski-Gutierrez EC, Mead PS. U.S. healthcare providers' experience with Lyme and other tickborne diseases. *Ticks Tick Borne Dis.* 2014;5(4):404–8

240 Ferrouillet C, Milord F, Lambert L, Vibien A, Ravel A. Lyme disease: Knowledge and practices of family practitioners in southern Quebec. *Can J Infect Dis Med Microbiol.* 2015;26(3):151–6

241 Gasmi S, Ogden NH, Leighton PA, Adam-Poupart A, Milord F, Lindsay LR, et al. Practices of Lyme disease diagnosis and treatment by general practitioners in Quebec, 2008-2015. *BMC Fam Pract.* 2017;18(1):65

242 Stricker RB, Johnson L. 'Lyme literacy' and physicians in Connecticut. *J Pediatr.* 2011;158(3):518–9

243 Johnson M, Feder HM. Chronic Lyme disease: a survey of Connecticut primary care physicians. *J Pediatr.* 2010;157(6):1025–1029

244 Magri JM, Johnson MT, Herring TA, Greenblatt JF. Lyme disease knowledge, beliefs, and practices of New Hampshire primary care physicians. *J Am Board Fam Pract.* 2002;15(4):277–84

245 Henry B, Crabtree A, Roth D, Blackman D, Morshed M. Lyme disease: knowledge, beliefs, and practices of physicians in a low-endemic area. *Can Fam Physician.*

2012;58(5):e289–95

246 Silverman D. *Interpreting Qualitative Data: Methods for Analyzing Talk, Text and Interaction*. 3rd ed. London: SAGE Publications Ltd; 2006.

247 Guest G, Bunce A, Johnson L. How many interviews are enough? An experiment with data saturation and variability. *Field Methods.* 2006;18(1):59–82

248 Fusch PI, Ness LR. Are we there yet? Data saturation in qualitative research. *Qual Rep.* 2015;20(9):1408–16

249 Glasziou P, Burls A, Gilbert R. Evidence based medicine and the medical curriculum. *BMJ.* 2008;337:a1253

250 Breakspear Medical. Lyme disease/chronic borreliosis. 2018. Available from: https://breakspearmedical.com/treatments/lyme-disease/ [Accessed 1st January 2019]

251 Public Health England. Mental health data and analysis: a guide for health professionals. 2018. Available from: https://www.gov.uk/guidance/mental-health-data-and-analysis-a-guide-for-health-professionals [Accessed 1st January 2019]

252 Parkin E. *Mental health policy in England*. Briefing Paper; Number CBP 07547, 4th September 2018

253 Atherton M. Lyme disease: 59p spice reduces symptoms in woman's battle with tick bite condition. *Sunday Express*. 2018. Available from: https://www.express.co.uk/life-style/health/951083/lyme-disease-symptoms-uk-supplement-turmeric-tick-bite. [Accessed 1st January 2019]

254 Auwaerter PG, Bakken JS, Dattwyler RJ, Dumler JS, Halperin JJ, McSweegan E, et al. Antiscience and ethical concerns associated with advocacy of Lyme disease. *Lancet Infect Dis.* 2011;11(9):713–9

255 Medical Defence Union. NICE issues guidance on Lyme disease. 2018.Available from: https://www.themdu.com/guidance-and-advice/latest-updates-and-advice/nice-issues-guidance-on-lyme-disease [Accessed 1st January 2019]

256 McGarrol S. The emotional challenges of conducting in-depth research into significant health issues in health geography: reflections on emotional labour,

fieldwork and life course. *Area* (Oxf). 2017;49(4):436–42

257 Ipsos. Tech Tracker Quarterly Release: Q4 2017. 2017. Available from: https://www.ipsos.com/sites/default/files/ct/publication/documents/2018-01/ipsos-tech_tracker_q4_2017.pdf [Accessed 1st January 2019]

258 Twitter. Twitter. 2018. Available from: https://twitter.com/ [Accessed 1st January 2019]

259 Sloan L. Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015. *Soc Media + Society.* 2017;3

260 1psos. Social Networking - Aug 2017. 2017. Available from: https://www.ipsos.com/en-ie/social-networking-aug-2017 [Accessed 1st January 2019]

261 Antheunis ML, Tates K, Nieboer TE. Patients' and health professionals' use of social media in health care: motives, barriers and expectations. *Patient Educ Couns.* 2013;92(3):426–31

262 Twitter. Search Tweets. 2018. Available from: https://developer.twitter.com/en/docs/tweets/search/guides [Accessed 1st January 2019]

263 Wang Y, Callan J, Zheng B. Should we use the sample? Analyzing datasets sampled from Twitter's stream API. *ACM Trans Web.* 2015;9(3)

264 Allen C, Tsou MH, Aslam A, Nagel A, Gawron JM. Applying GIS and machine learning methods to Twitter data for multiscale surveillance of influenza. *PLoS One.* 2016;11(7):e0157734

265 Nagel AC, Tsou MH, Spitzberg BH, An L, Gawron JM, Gupta DK, et al. The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. *J Med Internet Res.* 2013;15(10):e237

266 Health Protection Surveillance Centre. Lyme disease (Neuroborreliosis) (Borrelia burgdorferi). 2018. Available from: http://www.hpsc.ie/a-z/vectorborne/lymedisease/casedefinition/ [Accessed 1st January 2019]

267 Health Protection Surveillance Centre. HPSC Annual Epidemiological Report 2016.

2016. Available from: http://www.hpsc.ie/a-z/vectorborne/lymedisease/epidemiologicaldata/Other Vectorborne Diseases.pdf [Accessed 1st January 2019]

268 Health Protection Surveillance Centre. Cumulative crude incidence rate of Lyme neuroborreliosis, 2012-2018. 2018. http://www.hpsc.ie/a-z/vectorborne/lymedisease/epidemiologicaldata/LymeNeuro_2012-2018Cumulative_Map_WebContent_v1.1_20190102.pdf [Accessed 1st January 2019]

269 Yiannakoulias N, Tooby R, Sturrock SL. Celebrity over science? An analysis of Lyme disease video content on YouTube. *Soc Sci Med.* 2017;191:57–60

270 Basch CH, Mullican LA, Boone KD, Yin J, Berdnik A, Eremeeva ME, et al. Lyme disease and YouTube: a cross-sectional study of video contents. *Osong Public Health Res Perspect.* 2017;8(4):289–92

271 Seifter A, Schwarzwalder A, Geis K, Aucott J. The utility of 'Google Trends' for epidemiological research: Lyme disease as an example. *Geospat Health.* 2010;4(2):135–7

272 Pollett S, Althouse BM, Forshey B, Rutherford GW, Jarman RG. Internet-based biosurveillance methods for vector-borne diseases: are they novel public health tools or just novelties? *PLoS Negl Trop Dis.* 2017;11(11):e0005871

273 Pesälä S, Virtanen MJ, Sane J, Jousimaa J, Lyytikäinen O, Murtopuro S, et al. Health care professionals' evidence-based medicine internet searches closely mimic the known seasonal variation of Lyme borreliosis: a register-based study. *JMIR Public Health Surveill.* 2017;3(2):e19

274 Gentry J. Package 'twitteR'. *Cran.r-project*. 2016. Available from: https://cran.r-project.org/web/packages/twitteR/twitteR.pdf [Accessed 1st January 2019]

275 Central Statistics Office. Population. 2018. Available from: https://www.cso.ie/en/statistics/population/ [Accessed 1st January 2019]

276 Merck Animal Health. Bravecto (Fluralaner). 2019. Available from: https://us.bravecto.com/for-dogs [Accessed 1st January 2019]

277 Press Association. Rugby star Matt Dawson warns of tick dangers after heart scare.

The Guardian. 2017. Avaiable from: https://www.theguardian.com/society/2017/aug/21/rugby-star-matt-dawson-warns-of-tick-dangers-after-heart-scare [Accessed 1st January 2019]

278 Silver K. Matt Dawson: I had to have heart surgery after a tick bite. *BBC News*. 2017. Available from: https://www.bbc.co.uk/news/health-40973709 [Accessed 1st January 2019]

279 MSD Animal Health. The Big Tick Project. 2018. Available from: http://www.bigtickproject.co.uk [Accessed 1st January 2019]

280 Tulloch JSP. What is the risk of tickborne diseases to UK pets? *Vet Rec.* 2018;182(18):511–513

281 Vellinga A, Kilkelly H, Cullinan J, Hanahoe B, Cormican M. Geographic distribution and incidence of Lyme borreliosis in the west of Ireland. *Ir J Med Sci.* 2018;187(2):435–440

282 Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: Implications for infectious disease dynamics and control. *PLoS Comput Biol.* 2011;7(10):e1002199

283 Lavorgna L, Lanzillo R, Brescia Morra V, Abbadessa G, Tedeschi G, Bonavita S. Social media and multiple sclerosis in the posttruth age. *Interact J Med Res.* 2017;6(2):e18

284 Hancock JT, Toma C, Ellison N. The truth about lying in online dating profiles. In: *CHI 2007 Proceedings- Online Representation of Self*. 2007:449–52

285 Norman RA, Worton AJ, Gilbert L. Past and future perspectives on mathematical models of tick-borne pathogens. *Parasitology.* 2016;143(7):850-9

286 Public Health England. Tick awareness and the tick surveillance scheme. 2018. Available from: https://www.gov.uk/guidance/tick-surveillance-scheme [Accessed 1st January 2019]

287 Smith FD, Ballantyne R, Morgan ER, Wall R. Prevalence, distribution and risk associated with tick infestation of dogs in Great Britain. *Med Vet Entomol.* 2011;25(4):377–84

288 Wageningen. tekenradar.nl. 2012. Available from: https://www.tekenradar.nl/

[Accessed 25 July 2016]

289 Harms MG, Fonville M, van Vliet AJH, Bennema A, Hofhuis D, Beaujean H, et al. Tekenradar.nl, een webplatform over tekenbeten en de ziekte van Lyme. *Infect Bull.* 2014;7:204–6

290 Murray JK, Gruffydd-Jones TJ, Roberts MA, Browne WJ. Assessing changes in the UK pet cat and dog populations: numbers and household ownership. *Vet Rec.* 2015;177(10):259

291 Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas.* 1960;20(1):37–46

292 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977;33(1):159–74

293 Cleveland WS. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc.* 1979;74(368):829–36

294 Jacoby WG. Loess: a nonparametric, graphical tool for depicting relationships between variables. *Elect Stud.* 2000;19(4):577–613

295 Office for National Statistics. Postal Geography. 2016. Available from: http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/postal/index.html [Accessed 9 September 2016]

296 Pietzsch ME, Medlock JM, Jones L, Avenell D, Abbott J, Harding P, et al. Distribution of Ixodes ricinus in the British Isles: investigation of historical records. *Med Vet Entomol.* 2005;19(3):306–14

297 Randolph SE, Green RM, Hoodless AN, Peacey MF. An empirical quantitative framework for the seasonal population dynamics of the tick Ixodes ricinus. *Int J Parasitol.* 2002;32(8):979–89

298 Dobson AD, Taylor JL, Randolph SE. Tick (Ixodes ricinus) abundance and seasonality at recreational sites in the UK: hazards in relation to fine-scale habitat types revealed by complementary sampling methods. *Ticks Tick Borne Dis.* 2011;2(2):67–74

299 Ogden NH, Cripps P, Davison CC, Owen G, Parry JM, Timms BJ, et al. The ixodid tick species attaching to domestic dogs and cats in Great Britain and Ireland. *Med Vet Entomol.* 2000;14(3):332–8

300 Qviller L, Grøva L, Viljugrein H, Klingen I, Mysterud A. Temporal pattern of questing tick Ixodes ricinus density at differing elevations in the coastal region of western Norway. *Parasit Vectors.* 2014;7:179

301 Gilbert L, Aungier J, Tomkins JL. Climate of origin affects tick (Ixodes ricinus) host-seeking behavior in response to temperature: implications for resilience to climate change? *Ecol Evol.* 2014;4(7):1186–98

302 Robinson NJ, Dean RS, Cobb M, Brennan ML. Investigating common clinical presentations in first opinion small animal consultations using direct observation. *Vet Rec.* 2015;176(18):463

303 Radford AD, Noble PJ, Coyne KP, Gaskell RM, Jones PH, Bryan JG, et al. Antibacterial prescribing patterns in small animal veterinary practice identified via SAVSNET: the small animal veterinary surveillance network. *Vet Rec.* 2011;169(12):310

304 Pfäffle M, Petney T, Skuballa J, Taraschewski H. Comparative population dynamics of a generalist (Ixodes ricinus) and specialist tick (I. hexagonus) species from European hedgehogs. *Exp Appl Acarol.* 2011;54(2):151–64

305 Dziemian S, Sikora B, Piłacińska B, Michalik J, Zwolak R. Ectoparasite loads in sympatric urban populations of the northern white-breasted and the European hedgehog. *Parasitol Res.* 2015;114(6):2317–23

306 Haigh AJ. The Ecology of the European hedgehog (Erinaceus europaeus) in rural Ireland. *PhD Thesis, Univ Coll Cork.* 2011.

307 Gray JS, Dautel H, Estrada-Peña A, Kahl O, Lindgren E. Effects of climate change on ticks and tick-borne diseases in Europe. *Interdiscip Perspect Infect Dis.* 2009;593232

308 Phipps LP, Del Mar Fernandez De Marco M, Hernández-Triana LM, Johnson N, Swainsbury C, Medlock JM, et al. Babesia canis detected in dogs and associated ticks from Essex. *Vet Rec.* 2016;178(10):243-4

309 Hansford KM, Pietzsch ME, Cull B, Medlock JM. Importation of R sanguineus into the UK via dogs: tickborne diseases. *Vet Rec.* 2014;175(15):385–6

310 The Microchipping of Dogs (England) Regulations 2015, No.108.. Available from: http://www.legislation.gov.uk/uksi/2015/108/contents/made. [Accessed 1st January 2019]

311 Lindgren E, Talleklint L, Polfeldt T. Impact of climatic change on the northern latitude limit and population density of the disease-transmitting European tick Ixodes ricinus. *Environ Health Perspect.* 2000;108(2):119–23

312 Cleveland RB, Cleveland WS, McRae JE, Terpenning I. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *J Off Stat.* 1990;6(1):3–33

313 Serrà J, Arcos JL. An empirical evaluation of similarity measures for time series classification. *Knowledge-Based Syst.* 2014;67:305–14

314 Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans Acoust, Speech, Signal Process.* 1978;26(1):43-49

315 Said E, Dickey DA. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika.* 1984;71(3):599-607

316 Ratanamahatana CA, Keogh E. Making time-series classification more accurate using learned constraints. *Proceedings of the 2004 SIAM International Conference on Data Mining.* 2004;11-22

317 Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. *J Biomed Inform.* 2013;46(5):830–6

318 Sebastian-Coleman L. *Measuring data quality for ongoing improvement - a data quality assessment framework.* Elesevier, 2013. Available from: https://www.sciencedirect.com/book/9780123970336/measuring-data-quality-for-ongoing-improvement [Accessed 1st January 2019]

319 Chen H, Hailey D, Wang N, Yu P. A review of data quality assessment methods for public health information systems. *Int J Environ Res Public Health.* 2014;11(5):5170–207

320 Mathur R, Bhaskaran K, Chaturvedi N, Leon DA, van Staa T, Grundy E, et al. Completeness and usability of ethnicity data in UK-based primary care and hospital databases. *J Public Health (Oxf).* 2014;36(4):684–92

321 Health Protection Scotland. Zoonotic disease in Scotland. 2018. Available from: https://www.hps.scot.nhs.uk/pubs/detail.aspx?id=3598 [Accessed 1st January 2019]

322 Public Health England. Lyme disease: sample testing advice. 2018. Available from: https://www.gov.uk/guidance/lyme-disease-sample-testing-advice [Accessed 1st January 2019]

323 The Public Health (Infectious Diseases) Regulations 1988, No 1546. 1988. Available from: http://www.legislation.gov.uk/uksi/1988/1546/contents/made [Accessed 1st January 2019]

324 Bijkerk P, Fanoy EB, Kardamanidis K, van der Plas SM, Te Wierik MJ, Kretzschmar ME et al. To notify or not to notify: decision aid for policy makers on whether to make an infectious disease mandatorily notifiable. *Euro Surveill.* 2015;20(34):30003

325 Aenishaenslin C, Michel P, Ravel A, Gern L, Milord F, Waaub JP, et al. Factors associated with preventive behaviors regarding Lyme disease in Canada and Switzerland: a comparative study. *BMC Public Health.* 2015;15:185

326 Beaujean DJ, Bults M, van Steenbergen JE, Voeten HA. Study on public perceptions and protective behaviors regarding Lyme disease among the general public in the Netherlands: implications for prevention programs. *BMC Public Health.* 2013;13:225

327 Herbert S, Leong G, Hewitt K, Cassell J. Do genitourinary physicians report notifiable diseases? A survey in South East England. *Int J STD AIDS.* 2015;26(3):173–80

328 Vraukó K, Jancsó Z, Kalabay L, Lukacs A, Maraczi G, Mester L, et al. An appraisal: how notifiable infectious diseases are reported by Hungarian family physicians. *BMC Infect Dis.* 2018;18(1):45

329 Van Hest NA, Story A, Grant AD, Antoine D, Crofts JP, Watson JM. Record-linkage and capture-recapture analysis to estimate the incidence and completeness of reporting of tuberculosis in England 1999-2002. *Epidemiol Infect.* 2008;136(12):1606–16

330 Pillaye J, Clarke A. An evaluation of completeness of tuberculosis notification in the United Kingdom. *BMC Public Health.* 2003;3:31

331 Keramarou M, Evans MR. Completeness of infectious disease notification in the United Kingdom: a systematic review. *J Infect.* 2012;64(6):555–64

332 Squires SG, Aronson KJ, Remis RS, Hoey JR. Improved disease reporting: a randomized trial of physicians. *Can J Public Heal.* 1998;89(1):66–9

333 European Commission. Communicable diseases. 2018. Available from: https://ec.europa.eu/health/communicable_diseases/overview_en [Accessed 1st January 2019]

334 European Centre for Disease Prevention and Control. EU Case definitions. 2018. Available from: https://ecdc.europa.eu/en/all-topics-z/surveillance-and-disease-data/eu-case-definitions [Accessed 1st January 2019]

335 Public Health England. Syndromic surveillance: systems and analyses. 2015. Available from: https://www.gov.uk/government/collections/syndromic-surveillance-systems-and-analyses [Accessed 1st January 2019]

336 RCGP. RCGP Research and Surveillance Centre. 2018. Available from: http://www.rcgp.org.uk/clinical-and-research/our-programmes/research-and-surveillance-centre.aspx [Accessed 1st January 2019]

337 RCGP. RCGP Research and Surveillance Centre Annual Report 2015-2016. 2016. Available from: http://www.rcgp.org.uk/clinical-and-research/our-programmes/research-and-surveillance-centre.aspx [Accessed 1st January 2019]

Appendix I GP Validation Study: Participant Information Sheet

Understanding the decision making process of general practitioners (GPs) presented with non-specific conditions – A Pilot Study

We invite you to take part in a research study.

- Before you decide whether to take part, it is important for you to understand why this research is being done and what it will involve.
- Please take time to read the following information carefully.
- You are free to decide whether or not to take part in this study and may withdraw from it at any time.
- Ask us if there is anything that is not clear or if you would like more information.

Important things that you need to know

- We want to understand how GPs translate a variety of non-specific conditions into clinical codes (Read codes).
- Through a semi-structured interview you will be asked about a series of randomly presented case studies.
- These questions will be about how you diagnose, treat and refer these case patients, and ultimately how you code them within their electronic health records.
- This study will help inform a national internet-based questionnaire, utilising the same format and questions you will experience.
- We would like to receive as much feedback from yourself about the questionnaire, as this will inform any amendments for the national study.
- After the interview, which we anticipate to take 30 minutes, you will receive a debrief and be given a further opportunity for feedback and questions.

How to contact us

If you have any questions about this study, please email John Tulloch on jtulloch@liverpool.ac.uk

What's involved?

- GPs have been utilising Read codes as a way of recording a clinical diagnosis for over thirty years.
- They have become routinely used for auditing and surveillance purposes.
- However, for diseases or conditions with non-specific presentations, it can be extremely difficult to validate a Read code diagnosis, especially if the patient is only managed in the primary care setting.
- We have designed a questionnaire that will provide insight in to the decision making process around choosing a Read code and making further diagnostic, treatment and referral choices.
- We will be performing a semi-structured interview with 10 anonymous GPs in the North-West Coast area to assess how well our proposed questionnaire works.
- This study will help inform a larger nationwide on-line GP questionnaire.

What would taking part involve?

- You would meet with a researcher from the study at your practice and be introduced to the study.
- When you are ready and have signed the consent form, we will proceed with a semi-qualitative interview lasting between 30-45 minutes.
- This interview will be based around a series of fictitious patient cases provided by the researcher.
- On completion of the interview a 5 minute debrief will occur. After the debrief you will be free to withdraw from the study and all your results will be destroyed.
- At any point in the study you may withdraw your consent, and all your results will be destroyed.
- Your details will be anonymised and you will be referred to as GP X, Y etc. on all records.
- You will be given the opportunity to receive updates as the study progresses.
- Your personal data will be stored for 3 months while any anonymous study data may be kept for 5 years.

Possible benefits of taking part

- Help improve the use of Read codes used by GPs nationally, ensuring that surveillance based on them is as useful as possible for a variety of conditions.

Disadvantage and risks

- None that we are aware of.

Version 1.3 11/07/17

IRAS Number - 208815

Appendix II GP Validation Study: Participant Consent Form

The Royal Liverpool and Broadgreen University Hospitals NHS Trust

I

Anonymised Participant Identification Number for this interview:

CONSENT FORM

Title of Project: Understanding the decision making process of general practitioners (GPs) presented with non-specific conditions.

Name of Researcher: John Tulloch

Please initial box

1. I confirm that I have read the information sheet dated 11th July 2017 (version 1.3) for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily. ☐
2. I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason, without my legal rights being affected. ☐
3. I agree to take part in the above study. ☐

Name of Participant	Date	Signature
Name of Person taking consent	Date	Signature

Version 1.2 11/07/17

IRAS Number - 208815

When completed: 1 for participant; 1 for researcher site file; 1 (original) to be kept in medical notes.

Appendix III GP Validation Study: Questionnaire

Version 1.2 07/06/17 IRAS Number - 208815

Understanding the decision making process of general practitioners (GPs) presented with non-specific conditions

Participant Number:

Researcher, please note start time: ……..

Gender

- ❍ Male
- ❍ Female

Year of Birth

What is the first half of your practice postcode? (For example; ST15)

How many years have you been in general practice?

Are you a GP with a special interest?

- ❍ Yes
- ❍ No

What is your special interest?

In the following cases we will use the term Practice Management System (PMS). This refers to the computer based system that houses your patients' electronic health records.

Case 1

29 year old female

Presentation: A headache and fever for the last three days. On examination a single skin lesion is presented.

(Courtesy of CDC/James Gathany)

What are your three differential diagnoses for this case?

You discover that she has recently been bitten by a tick. What are your differentials now?

How would you code your top diagnosis on your PMS?

Would you perform further diagnostics for this case?

- ❍ Yes
- ❍ No

If yes, what would they be?

Would you prescribe a treatment for this case?

- ❍ Yes
- ❍ No

If yes, what would you prescribe, and what would the course of treatment be?

Would you refer this patient?

- ❍ Yes
- ❍ No

If yes, where/to whom would you refer this patient?

Any further comments? (For example, personal reflections about decision making or further information on case management.)

Case 2

10 Year Old Female

Presentation: Over the last few weeks has developed this painless nodule on her left ear lobe, no other abnormal clinical signs were seen.

(Courtesy of: http://www.meduni09.edis.at/eucalb/cms_15/index.php?option=com_content&task=view&id=98&Itemid=109)

What are your three differential diagnoses for this case?

How would you code your top diagnosis on your PMS?

Would you perform further diagnostics for this case?

- ❍ Yes
- ❍ No

If yes, what would they be?

Would you prescribe a treatment for this case?

- ❍ Yes
- ❍ No

If yes, what would you prescribe, and what would the course of treatment be?

Would you refer this patient?

- ❍ Yes
- ❍ No

If yes, where/to whom would you refer this patient?

Any further comments? (For example, personal reflections about decision making or further information on case management.)

Case 3

71 Year Old Male

Presentation: Over the last six months the extensor surface of his hands have gradually developed blueish red discolouration, with thinning and wrinkling of the skin. He has also developed a sensory peripheral neuropathy of the extensor surface of his right hand and forearm. He is showing strong signs of allodynia, an exaggerated nociceptive pain reaction to minimal trauma.

(Courtesy of http://www.dermis.net)

What are your three differential diagnoses for this case?

How would you code your top diagnosis on your PMS?

Would you perform further diagnostics for this case?

- ❍ Yes
- ❍ No

If yes, what would they be?

Would you prescribe a treatment for this case?

- ❍ Yes
- ❍ No

If yes, what would you prescribe, and what would the course of treatment be?

Would you refer this patient?

- ❍ Yes
- ❍ No

If yes, where/to whom would you refer this patient?

Any further comments? (For example, personal reflections about decision making or further information on case management.)

Case 4

10 year old female

Presentation: Acute unilateral facial palsy and headaches.

(Courtesy of Wellcome Images)

What are your three differential diagnoses for this case?

You discover an insect bite on her scalp.

What are your differentials now?

How would you code your top diagnosis on your PMS?

Would you perform further diagnostics for this case?

- ❍ Yes
- ❍ No

If yes, what would they be?

Would you prescribe a treatment for this case?

- ❍ Yes
- ❍ No

If yes, what would you prescribe, and what would the course of treatment be?

Would you refer this patient?

- ❍ Yes
- ❍ No

If yes, where/to whom would you refer this patient?

Any further comments? (For example, personal reflections about decision making or further information on case management.)

Case 5

44Year Old Male

Presentation: Recurrent episodes of synovitis in both knees, which has been persisting for the last 4 months.

What are your three differential diagnoses for this case?

How would you code your top diagnosis on your PMS?

Would you perform further diagnostics for this case?

- ❍ Yes
- ❍ No

If yes, what would they be?

Would you prescribe a treatment for this case?

- ❍ Yes
- ❍ No

If yes, what would you prescribe, and what would the course of treatment be?

Would you refer this patient?

- ❍ Yes
- ❍ No

If yes, where/to whom would you refer this patient?

Any further comments? (For example, personal reflections about decision making or further information on case management.)

Case 6

25 Year Old Female

Presentation: Presents with three identical lesions (one shown in picture) across her torso. No other clinical findings, and on discussion spent the previous day walking on Dartmoor.

(Courtesy of http://www.dermis.net)

What are your three differential diagnoses for this case?

How would you code your top diagnosis on your PMS?

Would you perform further diagnostics for this case?

- ❍ Yes
- ❍ No

If yes, what would they be?

Would you prescribe a treatment for this case?

- ❍ Yes
- ❍ No

If yes, what would you prescribe, and what would the course of treatment be?

Would you refer this patient?

- ❍ Yes
- ❍ No

If yes, where/to whom would you refer this patient?

Any further comments? (For example, personal reflections about decision making or further information on case management.)

Case 7

32 Year Old Male

Presentation: Last week took part in a 10km fun run, which he competes in on a regular basis. He has now noticed missed heart beats and a fluttering in his chest.

What are your three differential diagnoses for this case?

How would you code your top diagnosis on your PMS?

Would you perform further diagnostics for this case?

- ❍ Yes
- ❍ No

If yes, what would they be?

Would you prescribe a treatment for this case?

- ❍ Yes
- ❍ No

If yes, what would you prescribe, and what would the course of treatment be?

Would you refer this patient?

- ❍ Yes
- ❍ No

If yes, where/to whom would you refer this patient?

Any further comments? (For example, personal reflections about decision making or further information on case management.)

Case 8

53 year old female

Presentation: Presents with fatigue and post-exertional malaise, anxiety, headaches and self-reported memory issues.

What are your three differential diagnoses for this case?

After examination she presents you with a report from an international lab stating that she is positive for Lyme disease. She requests that you treat her with long term antibiotics.

What are your differentials now?

How would you code your top diagnosis on your PMS?

Would you perform further diagnostics for this case?

- ❍ Yes
- ❍ No

If yes, what would they be?

Would you prescribe a treatment for this case?

- ❍ Yes
- ❍ No

If yes, what would you prescribe, and what would the course of treatment be?

Would you refer this patient?

- ❍ Yes
- ❍ No

If yes, where/to whom would you refer this patient?

Any further comments? (For example, personal reflections about decision making or further information on case management.)

Case 9

62 Year Old Male

Presentation: Fatigue, aching of his knees and ankles, poor ability to concentrate, general muscle pain and mood swings.

What are your three differential diagnoses for this case?

How would you code your top diagnosis on your PMS?

Would you perform further diagnostics for this case?

- ❍ Yes
- ❍ No

If yes, what would they be?

Would you prescribe a treatment for this case?

- ❍ Yes
- ❍ No

If yes, what would you prescribe, and what would the course of treatment be?

Would you refer this patient?

- ❍ Yes
- ❍ No

If yes, where/to whom would you refer this patient?

Any further comments? (For example, personal reflections about decision making or further information on case management.)

Case 10

15 Year Old Male

Presentation: Only complaint is seen in the below image

(Courtesy of Public Health England; contains public sector information licensed under the Open Government Licence v3.0)

What are your three differential diagnoses for this case?

How would you code your top diagnosis on your PMS?

Would you perform further diagnostics for this case?

- ❍ Yes
- ❍ No

If yes, what would they be?

Would you prescribe a treatment for this case?

- ❍ Yes
- ❍ No

If yes, what would you prescribe, and what would the course of treatment be?

Would you refer this patient?

- ❍ Yes
- ❍ No

If yes, where/to whom would you refer this patient?

Any further comments? (For example, personal reflections about decision making or further information on case management.)

Case 11

38 Year Old Male

Presentation: Presents feeling low, sleeping badly, says he has noticed change in the character of his writing and difficulty with fine movements of his hands e.g. turning keys in locks. He had been fine until 1 month before. 2 months before he helped out on a scout camp and he came to see the GP shortly after that with a rash on his shoulder and the GP gave him erythromycin.

What are your three differential diagnoses for this case?

How would you code your top diagnosis on your PMS?

Would you perform further diagnostics for this case?

- ❍ Yes
- ❍ No

If yes, what would they be?

Would you prescribe a treatment for this case?

- ❍ Yes
- ❍ No

If yes, what would you prescribe, and what would the course of treatment be?

Would you refer this patient?

- ❍ Yes
- ❍ No

If yes, where/to whom would you refer this patient?

Any further comments? (For example, personal reflections about decision making or further information on case management.)

Post Interview Questions

Lyme Knowledge Assessment

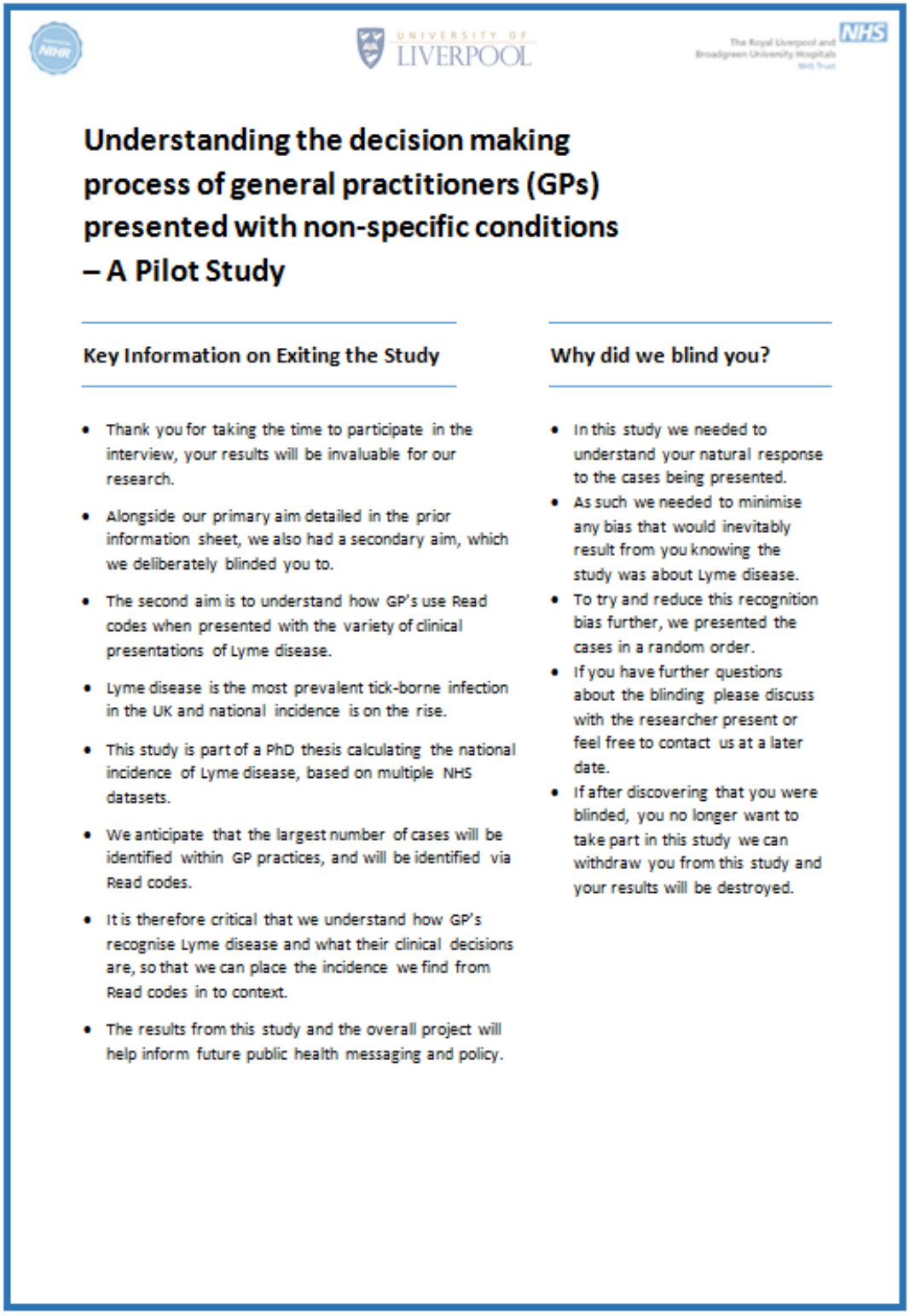
1) Prior to this questionnaire had you heard of Lyme disease? (Y/N)
2) What causes it?
3) Do you know where/who the reference lab is?
4) Do you think UK testing is reliable (Y/N)
5) Do you think Chronic Lyme disease exists? (Y/N)
6) Do you think Chronic Fatigue syndrome exists? (Y/N)
7) Who would you seek for advice about a Lyme disease patient?

Assessment of the Questionnaire

1) How does being initially blinded to the research make you feel?
2) Are you still happy to consent? Why?
3) Do you understand why the blinding was performed? (If no, researcher to explain methodology more)
4) Did you recognise that the study was about Lyme disease? (If yes, how and at what stage?)
5) How would you improve the blinding?
6) Are there any questions that you didn't understand or need improving?
7) How do you feel about the format and structure of the questionnaire? (Try to elicit from the participant positives and negatives)
8) How well do you think it would work in an on-line format, especially in regards to the questions around coding?
9) How would you distribute the survey to GPs on-line? Any websites that they use in particular?

Researcher, please note end time: ……

Appendix IV GP Validation Study – Participant exit information



Understanding the decision making process of general practitioners (GPs) presented with non-specific conditions – A Pilot Study

Key Information on Exiting the Study

- Thank you for taking the time to participate in the interview, your results will be invaluable for our research.
- Alongside our primary aim detailed in the prior information sheet, we also had a secondary aim, which we deliberately blinded you to.
- The second aim is to understand how GP's use Read codes when presented with the variety of clinical presentations of Lyme disease.
- Lyme disease is the most prevalent tick-borne infection in the UK and national incidence is on the rise.
- This study is part of a PhD thesis calculating the national incidence of Lyme disease, based on multiple NHS datasets.
- We anticipate that the largest number of cases will be identified within GP practices, and will be identified via Read codes.
- It is therefore critical that we understand how GP's recognise Lyme disease and what their clinical decisions are, so that we can place the incidence we find from Read codes in to context.
- The results from this study and the overall project will help inform future public health messaging and policy.

Why did we blind you?

- In this study we needed to understand your natural response to the cases being presented.
- As such we needed to minimise any bias that would inevitably result from you knowing the study was about Lyme disease.
- To try and reduce this recognition bias further, we presented the cases in a random order.
- If you have further questions about the blinding please discuss with the researcher present or feel free to contact us at a later date.
- If after discovering that you were blinded, you no longer want to take part in this study we can withdraw you from this study and your results will be destroyed.

More Information on Lyme disease

- If you'd like to more information about Lyme disease please follow this link to Public Health England's website (https://www.gov.uk/government/collections/lyme-disease-guidance-data-and-analysis)
- The Royal College of General Practitioners also produces a **free** e-learning module on the RCGP website, worth **0.5 CPD points**. (http://elearning.rcgp.org.uk/course/info.php?id=164)
- The British Infection Association also has a position statement on Lyme disease that can be found here: (http://www.aldf.com/pdf/BIA%202011statement%20on%20Lyme%20disease.pdf).

How to contact us

If you have any questions about this study, please email John Tulloch on jtulloch@liverpool.ac.uk

This study has been supported by the Health Protection Research Unit in Emerging and Zoonotic Infections. (http://www.hpruezi.nihr.ac.uk)

Version 1.1 13/01/17

IRAS Number - 208815

If you no longer wish to take part in the study please tick here ☐

Anonymised GP number (Researcher to fill in): .....

Appendix V GP Validation Study – Case 1 Answers

Appendix Table 1 A summary of GP responses to a blinded case presentation of a classic erythema migrans (EM) rash; Case 1.

Differential and coding questions. 7 (87.5%) GPs answered questions on this case					
Initial Differentials	Percentage of GPs with this differential (n)	Differential after tick bite reveal	Percentage of GPs with this differential (n)	Read Codes chosen	Percentage of GPs with this code (n)
Lyme disease	**71.4% (5)**	**Lyme disease**	**85.7% (6)**	Rash	42.6% (3)
Bite	28.6% (2)	Rash	14.3% (1)	**Lyme disease**	**28.6% (2)**
Drug reaction	14.3% (1)	Target lesion	14.3% (1)	Target lesion	28.6% (2)
Erythema multiform	14.3% (1)			**Erythema migrans**	**14.3% (1)**
Erythema nodosum	14.3% (1)			**Suspected Lyme disease**	**14.3% (1)**
Herald lesion	14.3% (1)				
Localised skin dermatitis	14.3% (1)				
Non-specific eruption	14.3% (1)				
Rash	14.3% (1)				
Target Lesion	14.3% (1)				
Would you perform diagnostics? (% Yes)			71.4% (5)		
			Bloods – 100% (5)		
Would you prescribe anything? (% Yes)			71.4% (5)		
			Check guidelines first – 80% (4)		
			Antibiotics – 20% (1)		
Would you refer this case? (% Yes)			57.1% (4)		
			Admit straight to hospital – 50% (2)		
			Infectious disease department – 25% (1)		
			Look up where to refer to – 25% (1)		

Appendix VI GP Validation Study – Case 2 Answers

Appendix Table 2 A summary of GP responses to a blinded case presentation of Borrelial lymphocytoma; Case 2.

Differential and coding questions. 7 (87.5%) GPs answered questions on this case			
Initial Differentials	Percentage of GPs with this differential (n)	Read Codes chosen	Percentage of GPs with this code (n)
Allergy	14.3% (1)	Ear swelling	28.6% (2)
Bite	14.3% (1)	Cellulitis	14.3% (1)
Cellulitis	14.3% (1)	Erythema of pinna	14.3% (1)
Chondritis	14.3% (1)	Inflammation of pinna	14.3% (1)
Cyst	14.3% (1)	No code	14.3% (1)
Ear symptom	14.3% (1)	Skin nodule	14.3% (1)
Eczema	14.3% (1)		
Infection	14.3% (1)		
Inflammation of pinna	14.3% (1)		
Irritation	14.3% (1)		
Malignant lump	14.3% (1)		
No idea	14.3% (1)		
Painless abscess	14.3% (1)		
Would you perform diagnostics? (% Yes)		28.6% (2)	
		Bloods – 50% (1) Look up what diagnostics – 50% (1)	
Would you prescribe anything? (% Yes)		42.9% (3)	
		Flucoxacillin and steroid cream – 66.7% (2) Ibuprofen and anti-histamines – 33.3% (1)	
Would you refer this case? (% Yes)		57.1% (4)	
		Ear, nose and throat consultant – 50% (2) Dermatology consultant – 25% (1) Look up where to refer to – 25% (1)	

Appendix VII GP Validation Study – Case 3 Answers

Appendix Table 3 A summary of GP responses to a blinded case presentation of acrodermatitis chronica atrophicans (ACA); Case 3.

Differential and coding questions. 8 (100%) GPs answered questions on this case			
Initial Differentials	Percentage of GPs with this differential (n)	Read Codes chosen	Percentage of GPs with this code (n)
Alcohol related issue	50% (4)	Peripheral neuropathy	37.5% (3)
Vitamin B12 deficiency	37.5% (3)	Sensory loss	25% (2)
Diabetes	25% (2)	No code	12.5% (1)
Natural ageing	25% (2)	Neuropathy	12.5% (1)
No idea	25% (2)	Numbness	12.5% (1)
Rheumatological cause	25% (2)		
Arthritis	12.5% (1)		
Cervical neck problems	12.5% (1)		
Iatrogenic cause	12.5% (1)		
Liver disease	12.5% (1)		
Multiple sclerosis	12.5% (1)		
Neurovascular issue	12.5% (1)		
New systemic condition	12.5% (1)		
Nutritional condition	12.5% (1)		
Peripheral neuropathy	12.5% (1)		
Radicular nerve root palsy	12.5% (1)		
Reynaud's phenomenon	12.5% (1)		
Thoracic outlet tumour	12.5% (1)		
Thyroid issue	12.5% (1)		
Would you perform diagnostics? (% Yes)		100% (8)	
		Bloods – 87.5% (7) Nerve conduction study – 37.5% (3) Diagnostic imaging – 25% (2) Electomyography – 12.5% (1) Neurological examination – 12.5% (1)	
Would you prescribe anything? (% Yes)		25% (2)	
		Painkillers – 100% (2) Gabapentins – 50% (1)	
Would you refer this case? (% Yes)		50% (4)	
		Neurology consultant – 75% (3) Dermatology consultant – 25% (1) Musculoskeletal consultant – 25% (1) Rheumatological consultant – 25% (1)	

Appendix VIII GP Validation Study – Case 4 Answers

Appendix Table 4 A summary of GP responses to a blinded case presentation of Bell's palsy caused by Lyme disease; Case 4.

Differential and coding questions. 7 (87.5%) GPs answered questions on this case					
Initial Differentials	Percentage of GPs with this differential (n)	Differential after tick bite reveal	Percentage of GPs with this differential (n)	Read Codes chosen	Percentage of GPs with this code (n)
Bell's palsy	100% (7)	Bell's palsy	100% (7)	Bell's palsy	57.1% (4)
Brain tumour	42.9% (3)	Brain tumour	28.6% (2)	Nerve palsy	28.6% (2)
Space occupying lesion	28.6% (2)	AV Malformations	14.3% (1)	Facial palsy	14.3% (1)
Arteriovenous malformation	14.3% (1)	Ear problems	14.3% (1)	Suspected zoster	14.3% (1)
Ear problems	14.3% (1)	Intracellabellar event	14.3% (1)	Unilateral facial palsy	14.3% (1)
Intracellabellar event	14.3% (1)	**Lyme disease**	**14.3% (1)**		
Nerve condition	14.3% (1)	Tick-borne virus	14.3% (1)		
Viral infection	14.3% (1)				
Zoster virus infection	14.3% (1)				
Would you perform diagnostics? (% Yes)			71.4% (5)		
			CT scan – 80% (4) Cranial Nerve exam – 40% (2) Bloods – 20% (1) Ophthalmic exam – 20% (1)		
Would you prescribe anything? (% Yes)			14.3% (1)		
			Prednisolone and anti-viral – 100% (1)		
Would you refer this case? (% Yes)			85.6% (6)		
			Paediatric consultant – 66.7% (4) Ear, nose and throat consultant – 16.7% (1) Emergency ophthalmology clinic – 16.7% (1) Neurology consultant – 16.7% (1)		

Appendix IX GP Validation Study – Case 5 Answers

Appendix Table 5 A summary of GP responses to a blinded case presentation of recurrent synovitis of knees; Case 5.

Differential and coding questions. 7 (87.5%) GPs answered questions on this case			
Initial Differentials	Percentage of GPs with this differential (n)	Read Codes chosen	Percentage of GPs with this code (n)
Osteoarthritis	71.4% (5)	Knee pain	42.9% (3)
Rheumatoid arthritis	57.1% (4)	Synovitis	28.6% (2)
Gout	42.9% (3)	Polyarthropathy	14.3% (1)
Inflammatory arthritis	28.6% (2)	Inflammatory arthritis	14.3% (1)
Bursitis	14.3% (1)		
Knee infection	14.3% (1)		
Lupus	14.3% (1)		
Polyarthritis	14.3% (1)		
Post-infectious arthritis	14.3% (1)		
Would you perform diagnostics? (% Yes)		100% (7)	
		Bloods – 85.7% (6) Diagnostic imaging – 71.4% (5)	
Would you prescribe anything? (% Yes)		100% (7)	
		Non-steroidal anti-inflammatories – 100% (7) Painkillers – 14.3% (1)	
Would you refer this case? (% Yes)		42.9% (3)	
		Musculoskeletal consultant – 66.7% (2) Rheumatological consultant – 66.7% (2)	

Appendix X GP Validation Study – Case 6 Answers

Appendix Table 6 A summary of GP responses to a blinded case presentation of multiple EM rashes; Case 6.

Differential and coding questions. 7 (87.5%) GPs answered questions on this case			
Initial Differentials	Percentage of GPs with this differential (n)	Read Codes chosen	Percentage of GPs with this code (n)
Insect bite	85.7% (6)	Non-specific rash	28.6% (2)
Fungal infection	42.9% (3)	**Erythema migrans**	**14.3% (1)**
Lyme disease	**42.9% (3)**	Insect bite	14.3% (1)
Contact dermatitis	28.6% (2)	No code	14.3% (1)
Allergy	14.3% (1)	Non-specific dermatitis	14.3% (1)
Eczema	14.3% (1)	Rash	14.3% (1)
Rash	14.3% (1)	Skin lesion	14.3% (1)
Ringworm	14.3% (1)		
Skin eruption	14.3% (1)		
Tick bite	14.3% (1)		
Viral infection	14.3% (1)		
Would you perform diagnostics? (% Yes)		42.9% (3)	
		Bloods – 66.7% (2) Skin scrape – 33.3% (1) Trial of medication – 33.3% (1)	
Would you prescribe anything? (% Yes)		85.7% (6)	
		Steroid and fungal cream – 50% (3) Antibiotics – 16.7% (1) Eyrthromycin – 16.7% (1) Hydrocortisone cream – 16.7% (1)	
Would you refer this case? (% Yes)		0% (0)	

Appendix XI GP Validation Study – Case 7 Answers

Appendix Table 7 A summary of GP responses to a blinded case presentation of heart rhythm abnormalities; Case 7.

Differential and coding questions. 7 (87.5%) GPs answered questions on this case			
Initial Differentials	Percentage of GPs with this differential (n)	Read Codes chosen	Percentage of GPs with this code (n)
Arrhythmia	57.1% (4)	Palpitations	71.4% (5)
Palpitations	42.9% (3)	Arrhythmia	14.3% (1)
Anxiety	28.6% (2)	No code	14.3% (1)
Atrial fibrillation	28.6% (2)		
Ectopic beats	28.6% (2)		
Anaemia	14.3% (1)		
Heart defect	14.3% (1)		
Performance enhancing drugs	14.3% (1)		
Recreational drug use	14.3% (1)		
Stress	14.3% (1)		
Supraventricular tachycardia	14.3% (1)		
Would you perform diagnostics? (% Yes)		85.7% (6)	
		Electrocardiogram – 100% (6) Bloods – 83.3% (5) Test for drug use – 16.7% (1) Urinalysis – 16.7% (1)	
Would you prescribe anything? (% Yes)		0% (0)	
Would you refer this case? (% Yes)		28.6% (2)	
		Cardiology consultant – 50% (1) Refer to emergency department – 50% (1)	

Appendix XII GP Validation Study – Case 8 Answers

Appendix Table 8 A summary of GP responses to a blinded case presentation of multiple symptoms; Case 8.

Differential and coding questions. 5 (62.5%) GPs answered questions on this case					
Initial Differentials	Percentage of GPs with this differential (n)	Differential after results reveal	Percentage of GPs with this differential (n)	Read Codes chosen	Percentage of GPs with this code (n)
Thyroid problems	80% (5)	Thyroid problems	60% (3)	Tired all the time	60% (3)
Chronic fatigue syndrome (CFS)	60% (3)	CFS	40% (2)	Anxiety with depression	20% (1)
Depression	60% (3)	Depression	40% (2)	Fatigue	20% (1)
Anaemia	40% (2)	Anaemia	20% (1)	No code	20% (1)
Anxiety	40% (2)	Anxiety	20% (1)		
Menopausal	20% (1)	**Lyme disease**	**20% (1)**		
Psychosomatic cause	20% (1)	Menopausal	20% (1)		
Tired all the time	20% (1)	Tired all the time			
Would you perform diagnostics? (% Yes)			100% (5)		
			Bloods – 60% (3) **Lyme disease serology – 40% (2)** PHQ-9 – 20% (1)		
Would you prescribe anything? (% Yes)			20% (1)		
			Counselling – 100% (1)		
Would you refer this case? (% Yes)			60% (3)		
			Infectious disease consultant – 66.7% (2) Microbiology consultant– 33.3% (1) Neurology consultant – 33.3% (1)		

Appendix XIII GP Validation Study – Case 9 Answers

Appendix Table 9 A summary of GP responses to a blinded case presentation of multiple symptoms; Case 9.

Differential and coding questions. 7 (87.5%) GPs answered questions on this case			
Initial Differentials	Percentage of GPs with this differential (n)	Read Codes chosen	Percentage of GPs with this code (n)
Depression	71.4% (5)	Fatigue	42.9% (3)
Anaemia	42.9% (3)	Polyarthralgia	28.6% (2)
Thyroid problems	42.9% (3)	Low mood	14.3% (1)
Chronic fatigue syndrome	28.6% (2)	No code	14.3% (1)
Fibromyalgia	28.6% (2)	Tired all the time	14.3% (1)
Osteoarthritis	28.6% (2)		
Anxiety	14.3% (1)		
Arthralgia	14.3% (1)		
Iatrogenic	14.3% (1)		
Multiple joint pain	14.3% (1)		
Polymyalgia	14.3% (1)		
Psychosomatic cause	14.3% (1)		
Rheumatological cause	14.3% (1)		
Would you perform diagnostics? (% Yes)		100% (7)	
		Bloods – 100% (7)	
		PHQ-9 – 28.6% (2)	
		Prostate specific antigen test – 14.3% (1)	
		Stool analysis – 14.3% (1)	
		Urinalysis – 14.3% (1)	
		X-ray – 14.3% (1)	
Would you prescribe anything? (% Yes)		71.4% (5)	
		Painkiller – 40% (2)	
		Treat symptomatically - 40% (2)	
		Non-steroidal anti-inflammatories – 20% (1)	
Would you refer this case? (% Yes)		14.3% (1)	
		Radiology department – 100% (1)	

Appendix XIV GP Validation Study – Case 10 Answers

Appendix Table 10 A summary of GP responses to a blinded case presentation of a tick; Case 10.

Differential and coding questions. 6 (75%) GPs answered questions on this case			
Initial Differentials	Percentage of GPs with this differential (n)	Read Codes chosen	Percentage of GPs with this code (n)
Cutaneous horn	33.3% (2)	Skin lesion	33.3% (2)
Foreign body	33.3% (2)	**Tick bite**	**33.3% (2)**
Skin lesion	33.3% (2)	Pustule	16.7% (1)
Tick	**33.3% (2)**	Scalp lesion	16.7% (1)
Cyst	16.7% (1)		
Granulomatous lesion	16.7% (1)		
Infected hair follicle	16.7% (1)		
Insect	16.7% (1)		
Lump	16.7% (1)		
Pustular lesion	16.7% (1)		
Would you perform diagnostics? (% Yes)		66.7% (4)	
Would you prescribe anything? (% Yes)		Remove and send to pathology – 100% (4) 16.7% (1)	
Would you refer this case? (% Yes)		Antibiotics – 100% (1) 0% (0)	

Appendix XV GP Validation Study – Case 11 Answers

Appendix Table 11 A summary of GP responses to a blinded case presentation of motor movement issues and rashes; Case 11.

Differential and coding questions. 8 (100%) GPs answered questions on this case			
Initial Differentials	Percentage of GPs with this differential (n)	Read Codes chosen	Percentage of GPs with this code (n)
Depression	50% (4)	Feeling low	25% (2)
Neurological condition	50% (4)	Low mood	25% (2)
Anxiety	25% (2)	Neurological symptoms	25% (2)
Motor neurone disease	25% (2)	No code	12.5% (1)
Parkinson's disease	25% (2)	Parasitic infection	12.5% (1)
Alcohol intake problems	12.5% (1)		
Brain tumour	12.5% (1)		
Dementia	12.5% (1)		
Drug induced	12.5% (1)		
Fungal lesion	12.5% (1)		
Herald lesion	12.5% (1)		
Infectious disease issue	12.5% (1)		
Lyme disease	**12.5% (1)**		
Mental health issue	12.5% (1)		
Parasitic infection	12.5% (1)		
Post-viral arthralgia	12.5% (1)		
Thyroid function	12.5% (1)		
Tick-borne virus	12.5% (1)		
Would you perform diagnostics? (% Yes)		87.5% (7)	
		Bloods – 71.4% (5) Lyme disease serology – 14.3% (1) Neurological exam – 14.3% (1)	
Would you prescribe anything? (% Yes)		0% (0)	
Would you refer this case? (% Yes)		37.5% (3)	
		Refer to emergency department – 33.3% (1) Infectious disease consultant – 33.3% (1) Neurology consultant – 33.3% (1)	