USING LARGE-SCALE SYNDROMIC DATASETS TO SUPPORT EPIDEMIOLOGY AND SURVEILLANCE

By Dr Daniel Todkill, MBChB, BSc (hons), MPH, FFPH

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF PHD BY PUBLISHED WORKS November 2021

WORD COUNT * 10,210

* not including Abstract, Tables, References or Figures

UNIVERSITY OF WARWICK, WARWICK MEDICAL SCHOOL

“Surveillance serves as the eyes of Public Health”

Fairchild et.al [1]

CONTENTS

1. ACKNOWLEDGEMENTS

I would like to acknowledge the exceptional teaching and mentorship that have made this PhD possible, particularly by Prof Gillian Smith, Dr Obaghe Edeghere, Prof Jeremy Hawker, Prof Alex Elliot, Prof Aileen Clarke and Prof Noel McCarthy.

I would like to acknowledge all my co-authors on the indexed publications, and the unnamed individuals and organisations responsible for collecting, collating and sharing their data.

I would like to acknowledge the support from my parents Jean and Patrick William Todkill in supporting my family through especially difficult times and whose constant support enabled both this PhD and my career in its entirety, and my wife Kerry and two children Sophia and Arthur who provide much of the fun in life.

2. INDEX OF PUBLISHED WORK FOR CONSIDERATION

- **PAPER 1: Todkill D**, Loveridge P, Elliot AJ, Morbey R, Rayment Bishop T, Rayment Bishop C, Edeghere O, Smith GE Utility of ambulance dispatch data for real-time syndromic surveillance. A pilot in the West Midlands region, United Kingdom. Journal of Prehospital and Disaster Medicine, 2017;32:667-672. [2]
- **PAPER 2: Todkill D**, Loveridge P, Elliot AJ, Morbey R, De Lusignan S, Edeghere O, Smith GE. Socio-economic and geographical variation in general practitioner consultations for allergic rhinitis in England, 2003 to 2014: an observational study. BMJ Open 2017;7:e017038. doi: 10.1136/bmjopen-2017-017038 [3]
- **PAPER 3: Todkill D**, Colón-González FJ, Morbey R, Charlett A, Hajat S, Kovats S, Osborne NJ, McInnes RN, Vardoulakis S, Exley K, Edeghere O, Elliot AJ. Key environmental predictors of general practitioner consultations for allergic rhinitis in London, England: a retrospective time series analysis. BMJ Open, 2020; ;10(12):e036724 doi: 10.1136/bmjopen-2019-036724 [4]
- **PAPER 4: Todkill D**, Elliot A J, Morbey R, Harris J, Hawker J, Edeghere O, Smith G. What is the utility of syndromic surveillance systems during large subnational infectious gastrointestinal disease outbreaks? An observational study using case studies from the past five years in England. Epidemiology and Infection. 2016;144:2241-2250. 10.1017/S0950268816000480 [5]
- **PAPER 5: Todkill D**, Hughes H, Elliot A, Morbey R, Edeghere O, Harcourt S, Hughes T, Endericks T, McCloskey B, Catchpole M, Ibbotson S, Smith G. An Observational Study using English syndromic surveillance data collected during the 2012 London Olympics - what did syndromic surveillance show and what can we learn for future mass gathering events? Prehospital and Disaster Medicine. 2016;31:628-634. [6]

3. STATEMENT OF ETHICAL CONSIDERATIONS

All data used in the papers indexed in this PhD were fully anonymised, collated and stored on secure servers within PHE, and their use was for Public Health purposes. All datasets used in the indexed papers comply with PHE data protection standards and comply with Caldicott principles.

4. DECLARATION

This work has not previously submitted or used before at either the University of Warwick or another University for another degree. This is the candidates own work and is based on collaborative research. The individual contribution is outlined in Section Six.

5. SUMMARY OF ABBREVIATIONS

ADSSS	Ambulance Dispatch Syndromic Surveillance System
AR	Allergic Rhinitis
CDC	Centre for Disease Control
CPC	Chief Presenting Complaint
CUSUM	Cumulative Sum
EARS	Early Aberration Reporting System
ED	Emergency Department
EHR	Electronic Health Records
GI	Gastro-Intestinal
GP	General Practitioner
IMD	Indices of Multiple Deprivation
MG	Mass Gathering
ML	Machine Learning
NASSS	National Ambulance Dispatch Data Syndromic Surveillance System
NHS	National Health Service
PHE	Public Health England
RCGP	Royal College of General Practitioners
ReSST	Real Time Syndromic Surveillance Team
SS	Syndromic Surveillance
WHO	World Health Organisation
WMAS	West Midlands Ambulance Service

6. STATEMENT OF CANDIDATE'S CONTRIBUTION TO THE PUBLISHED WORK

Paper 1	Daniel Todkill (DT) co-ordinated the project and co-developed the idea for the study. DT completed analysis of the data, Roger Morbey (RM) provided statistical support and Tracey Rayment Bishop provided support with Ambulance data. DT drafted and the original manuscript. All authors contributed to, and commented on the protocol, methodology and final manuscript, and DT submitted.
Paper 2	DT co-ordinated the project and co-developed the idea for the study. DT and Paul Loveridge completed analysis of the data, RM provided statistical support and Simon De Lusignan provided support with RCGP data. DT drafted and the original manuscript. All authors contributed to, and commented on the protocol, methodology and final manuscript, and DT submitted.
Paper 3	DT co-ordinated the project, provided initial analysis and drafted the manuscript. Felipe de Jesus Colon Gonzalez completed the statistical analysis. RM and Andre Charlett also provided statistical support. Shakoor Hajat, Sari Kovats (SK), Nicholas J Osbourne, Rachel McInnes (RMc), Sotiris Vardoulakis and Karen Exley provided expertise in environmental factors. RMc organised data acquisition for environmental variables. SK, Gillian Smith, DT and Alex J Elliot were involved in the project initiation. DT drafted and submitted the original manuscript. All authors contributed to, and commented on the protocol, methodology and final manuscript.
Paper 4	DT co-ordinated the project and co-developed the idea for the study. DT completed the analysis of the data. RM provided statistical support and John Harris and Jeremy Hawker provided expertise in gastrointestinal illness. DT drafted and submitted the original manuscript. All authors contributed to, and commented on the protocol, methodology and final manuscript.
Paper 5	DT co-ordinated the project and co-developed the idea for the study. DT and Helen Hughes (HH) completed the analysis of the data. RM

	provided statistical support and Tina Endericks and Brian McCloskey provided expertise in Mass Gatherings. DT drafted the original manuscript. All authors contributed to, and commented on the protocol, methodology and final manuscript, and DT submitted. DT and HH are joint first authors on this paper.

7. ABSTRACT

Healthcare and the healthcare industry have traditionally produced huge amounts of data and information; patient care necessitates accurate record keeping, records of attendances and often details of the reason for contact with healthcare and outcomes.[7] During the past decade, there has been a dramatic shift to digitize healthcare related information, with a view to both increasing efficiencies in these areas, and to generate new insights.[8] These rich, but often unstructured data sources can present both opportunities and challenges to data scientists and epidemiologists. Syndromic surveillance (SS) is the real-time (or near real-time) collection, analysis, interpretation, and dissemination of health-related data to enable the early identification of the impact (or absence of impact) of potential human or veterinary public-health threats which require effective public-health action.[9] In England, Public Health England (PHE) coordinates a suite of national real-time syndromic surveillance systems. Underpinning their operation is the collation, analysis and interpretation of large-scale datasets ("big data").

This PhD by Published Works describes work which has evaluated, developed or utilised a number of these large healthcare datasets for both surveillance and epidemiology of public health events. The thesis is divided into four themes covering critical aspects of SS. Firstly, developing SS systems using novel data sources; something which is currently under-reported in the literature. Secondly, using syndromic data systems for non-infectious disease epidemiology; understanding how these systems can inform public health insight and action outside of their original remit. Thirdly, determining the utility in identifying outbreaks which was one of the original envisioned purposes of SS, using gastrointestinal illness (GI) as a case-study. The final theme is understanding how SS is used in the context of mass gatherings; again, a key original aspect of syndromic surveillance.

The thesis collates a portfolio of indexed works, all of which use (combined with other data sources) large, health-related data collated and operated by the PHE Real-Time Syndromic Surveillance Team (ReSST) and employ a range of different methodologies to translate data into public health action. These include describing the development of a novel system, observational studies and time series analysis.

Key findings from the papers include; learning how to develop these systems, demonstration of their utility in non-infectious disease epidemiology, leading to new insights into the socio-demographic distribution and causes of presentations to healthcare with Allergic Rhinitis, understanding the challenges and limitations of syndromic surveillance in identifying outbreaks of GI disease and how they can be used during mass gatherings.

Using diverse methodologies and data as a collective, the papers have led to significant public health impacts; both in terms of how these systems are used in England currently and how they have influenced global development of this small but growing speciality.

8. BACKGROUND

8.1 INTRODUCTION AND RESEARCH THEMES OF THIS PHD.

The World Health Organisation define public health surveillance as the continuous, systematic collection, analysis and interpretation of health related data.[10] This can serve as an early warning system for public health threats, enable monitoring of the impacts of public health interventions or track progress toward specific goals and support the epidemiology of public health problems; guiding priority setting, planning and strategy.[10] In England, Public Health England (PHE) is the Government agency responsible for protecting the Nation's health and wellbeing, and reducing health inequalities.[11] As part of this remit, PHE operates numerous surveillance systems to support public health intelligence; ranging from enhanced surveillance of Tuberculosis to non-communicable cancer clusters.[12]

To support the overall surveillance strategy, PHE coordinates a suite of national real-time syndromic surveillance systems. Syndromic surveillance (SS) is the real-time (or near real-time) collection, analysis, interpretation, and dissemination of health-related data to enable the early identification of the impact (or absence of impact) of potential human or veterinary public-health threats which require effective public-health action.[9] These systems include a National Health Service (NHS) 111 surveillance system that monitors daily calls made to the national NHS 111 telephone service, an emergency department (ED) surveillance system that monitors daily ED attendances [13] and general practitioner (GP) surveillance systems that monitor 'in-hours' consultations and 'out-of-hours' GP activity across England.[14] In addition, a national ambulance system has recently been launched (the development and evaluation of the utility of the pilot ambulance surveillance system is described in one of the papers included in the PhD [2]).

Underpinning the operation of these surveillance systems is the collation, analysis and interpretation of large-scale datasets ("big data"). The surveillance systems are passive; the data are collected by healthcare service providers for other purposes and transferred to PHE using passive automated systems. PHE uses these data primarily for routine surveillance purposes, however there is a spectrum of potential uses and applications in public health; a role in 'mass gatherings' (MG), infectious disease outbreak detection, monitoring changes in secular trends of key public health problems, supporting the evaluation of interventions (e.g. the impact of new vaccines) and using the data to support the epidemiology of non-infectious diseases.

This PhD by Published Works describes work that has evaluated, developed or utilised a number of large healthcare datasets for both surveillance and epidemiology of public health events. The PhD collates a portfolio of first-authored publications all of which use (combined with other data sources) large, health-related data collated and operated by the PHE ReSST and employ a range of different methodologies to translate data into public health action.

The work covered in this thesis was conducted in close collaboration with the ReSST, and the indexed papers and associated research questions were developed to address critical questions about the abilities of these systems pertinent to their operation in daily public health practice. These questions make up the four ‘themes’ discussed here.

The themes are illustrated by the indexed papers, and the questions which each theme was intended to explore are described in Table A.

Table A: The research questions underpinning each Theme in this thesis

Theme One Research Question: How can different data sources be used for the spectrum of syndromic surveillance purposes, and specifically what additional utility does ambulance data bring?

The first indexed paper describes the development of a pilot syndromic surveillance system and identifies lessons learnt using ambulance dispatch data, and specifically aimed to determine the feasibility and utility of using ambulance dispatch data to complement the suite of existing national SS systems. This theme explores how different data sources can be used for the spectrum of SS purposes, and using the indexed paper as a case study, specifically what additional utility ambulance data brings to SS.

Theme Two Research Question: What is the potential for UK SS systems to provide insight into non-infectious disease epidemiology?

There has been a recent call in the literature for enhanced use of syndromic data to better understand the relationships between the social determinants of health and non-communicable disease illness.[15] The aim of this theme is to determine the potential of PHE syndromic systems in providing insight into non-infectious disease epidemiology, both in terms of descriptive epidemiology and combining with other large scale data sources through modelling. This is illustrated using two indexed papers and allergic rhinitis as a case study; the aim of the first was to describe the epidemiology of GP consultations for allergic rhinitis in the UK, and the second linking syndromic data with meteorological and environmental data sets using retrospective time series analysis to determine key predictors of general practitioners' consultations for allergic rhinitis.

Theme Three Research Question: What is the utility of SS in detection of outbreaks of infectious Gastrointestinal disease?

Epidemiologists in the ReSST are routinely asked by stakeholders and incident directors if the suite of SS systems can detect outbreaks of gastrointestinal illness, and a role in infectious disease cluster identification and tracking has been posited. This theme explores the utility of SS systems for this purpose, including the use of aberration detection algorithms and the indexed paper specifically aimed to assess if existing PHE based systems were able to reliably detect large, sub-national GI outbreaks in the England.

Theme Four Research Question: What is the utility of SS during Mass Gatherings?

Internationally, many SS systems were developed in preparation for mass gathering events, including an expansion of systems in England in response to the London 2012 Olympic and Paralympic Games. With the return of the Olympic Games in 2024 to Europe, and locally the 2022 Commonwealth Games being held in Birmingham, learning from previous experience is necessary. The indexed paper specifically investigates the impact of a large mass gathering event as monitored in real time using two SS systems. More broadly for this PhD, this theme explores the role of SS in Mass Gatherings.

8.2 WHAT IS SYNDROMIC SURVEILLANCE?

There have been a number of definitions of SS, however for this thesis I have adopted the definition proposed by the European Triple S Project:[9] "SS is the real time (or near-real time) collection, analysis, interpretation and dissemination of health-related data to enable the early identification of the impact (or absence of impact) of potential human or veterinary public health threats which require effective public health action".

There is yet to be a fully accepted definition of what SS is, especially of what constitutes 'syndromic' data, and this may be in part be due to the constant evolution of what is considered 'SS'; and shifts over time as these surveillance systems adapt to differing public health needs. [16] During the early 2000's, what has eventually become known as SS was referred to by a number of definitions such as 'early warning systems', 'bio-surveillance' or 'health-indicator systems',[17] alluding to the diversity of what might be eventually considered syndromic. Traditionally, syndromic data was considered relevant to the monitoring of groups of clinical symptoms or features which relate to specific illnesses;[18] of course this does not however encompass non-clinical sources of data [19] which also have potential for surveillance purposes.

Broadly, SS involves the utilisation of data relating to the early symptoms of an illness (otherwise known as a prodrome), which happen during the period before laboratory confirmation of illness [17] with syndromes having been defined as a set of conditions or symptoms which occur together suggesting the presence of an increased chance of developing the disease.[19] The use of 'condition' within this definition encompasses non-clinical data sources. More recently authors have sought to refine definitions; in an attempt to incorporate machine learning (ML) into the analytics pathway, Kulessa et.al [20] have divided SS into 'specific' and 'non-specific'. They argue that 'specific' SS is the monitoring of the characteristics of a given or known disease, and 'non-specific' SS is the monitoring of the stream or streams of data for anomaly detection to identify public health threats.

The range of data sources with demonstrable potential for SS includes data from clinical encounters, such as ED attendances [21] , general practitioner consultations [14 22] remote health advice [23] or non-clinical data sources, such as school absenteeism [24-26], internet searches [27-29] or over-the-counter medicine purchases.[30-32] Somewhat surprisingly, there is little in the literature about understanding the purposes of different data sources, or the lessons learnt from developing systems. This critical aspect of SS is expanded on in Theme One.

The roles of SS are broad, but have been summarised as allowing the estimation of the magnitude of public health problems, describing the natural history of a disease and the distribution and spread of illness, the early warning of outbreaks and detection of changes in health practices [33], as well as reassurance to decision makers about changes in community morbidity.[6] Syndromic systems have been developed and deployed to identify health threats at earlier stages than traditional surveillance methods (such as those using laboratory or clinical disease notifications) which can both enable timely public health action alongside providing near-real time information on changes on disease activity. This can range from monitoring pandemic [34] or seasonal [35] influenza, monitoring the effectiveness of vaccinations [36] to describing changes in the secular trends of seasonal organisms such as norovirus.[21] Syndromic systems also have the potential to identify poorly understood phenomena, such as the relationships between thunderstorms and asthma.[37] During the recent and on-going COVID-19 pandemic, syndromic data are providing information to decision makers on the community activity of SARS-CoV-2.[38-40] Using the case study of gastrointestinal disease outbreaks, the abilities of SS systems to detect outbreaks is explored during Theme Three of this PhD.

Although the evidence-base for, and associated amount of literature is comparatively small (albeit growing), the concept of SS is not new. Originally conceptualised to identify biological terrorist attacks in the U.S.A in the wake of the 9/11 attacks.[18] Following the initial bioterrorist aims, it was used primarily for influenza surveillance before widening to a multi-hazard approach. It is now practised by public health agencies across the globe, with sophisticated SS systems in operations in many European [41 42] nations and North American States.[43]

In contrast to most other countries, in England a national programme of SS is co-ordinated centrally within Public Health England (PHE), the national Government organisation that coordinates the public health and health protection response for England. This is co-ordinated through a dedicated team of analysts, epidemiologists and scientists who specialise in SS systems, data and analyses (the ReSST). ReSST operate a suite of systems which includes monitoring of a remote health advice service, attendances at Eds [13], in and out of hours GP attendances [14] and national ambulance dispatch data.[44] Syndromic data are monitored by visualising daily counts of indicators over time to identify secular and non-seasonal trends, and unusual activity identified using automated alerting of higher-than-expected levels of indicators with exceedance algorithms.[45] These automated alerts highlight 'alarms', which are risk assessed according to pre-defined criteria.[46]

8.3 WHAT IS BIG DATA?

We are currently witnessing the confluence of several factors which hold the potential of understanding human behaviour and biology, improving health and generating new insight through the collection, use and analysis of large datasets.   The 21st century has seen a rapid growth in the collection, collation and analysis of these large-scale datasets; which have become popularly known both in the media, and increasingly academic literature as 'big data'.   Usage of both the colloquial term and these data themselves has become intrinsic in both society and our daily lives; from assisting the manufacturing of the foods we eat [47], managing the traffic on our daily commutes [48], how we search for and access information [49], to even influencing our political perspectives.[50]   The use of 'big data' has become pervasive in fields ranging from finance [51], advertising [52], sports [53] to sociology [54] , psychology [55] and medicine.[56]

Understanding what is meant by 'big data' is challenging. The term itself it thought to have arisen in Silicon Valley during the 1990s [57], and Favaretto et.al [58] link the exponential increase in use of the term with the 'explosion' in the quantity of potentially relevant data with advancements in data recording and storage technologies, and the five 'V's' being attributed to it; heterogeneous (**v**ariety), high-speed processing (**v**elocity), large amounts (**v**olume), authenticity (**v**eracity) and perhaps most importantly; turning the data into something of **v**alue.[59] The qualitative work by Faveretto et.al [58] to identify a consensus definition recognised the associated uncertainty and that it might be a culturally evolving concept. Numerous authors and studies have attempted definitions, including a 2014 systematic review [60] which concluded that big data should be defined as datasets with a logarithm of the product of the number of statistical individuals (n) and the number of variables (p) $(\text{Log}(n * p)) > 7$. This somewhat arbitrary definition doesn't take into account important and necessary features of big data; namely the form of analysis and purpose. This is described in the definition by De Mauro and colleagues [61]; 'Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value'. More comprehensively, in their well cited paper, Wu et.al [62] present a model based on a HACE theorem (table B) which describes the characteristics of big data. Through consideration of this concept of what big data is, it incorporates that these data sets are large in volume, complex, growing and utilise multiple autonomous sources; and intrinsic to this are the challenges associated with mining these data.

Table B: 'HACE' theorem of Big Data characteristics [62]

HACE theorem characteristic of Big Data	Explanatory Note
Huge Data with **H**eterogeneity and Diverse Dimensionality	Fundamental Characteristic of Big Data is a large volume of data represented by heterogeneous and diverse dimensionalities; necessary as different information gatherers often use this for varied purposes.
Autonomous sources with distributed and decentralized control	Each data source is able to generate and collect information without involving centralised control.
Complexity and **E**volving Relationships between data	As the volume of Big Data increases, so does the complexity and underlying relationships between the data.
Challenges with Mining the Data	Multiple challenges in understanding the data; ultimately the purpose is to transform the complex (non-linear) data relationships and evolving changes into consideration to discover useful patterns from Big Data collection and collation.

Although the definition of big or large data has been traditionally conceptually vague [58] the HACE concept encompasses much of how large-scale datasets are used for SS purposes. Fundamentally, syndromic data is from a variety of sources; while primary data collection is active, it's secondary collection for SS is invariably passive, with different organisations and individuals responsible for its collection and collation. This fundamental tenet of syndromic data leads to the argument that the data sources are autonomous. For example, a different organisation will run ambulance care to those receiving patients in an emergency department, and to a micro-degree, within an organisation, system or even at individual practitioner level there is invariably differentiation in how medical encounters are coded.

The complex temporal-spatial and exceedance analysis necessary for understanding syndromic data relates to a degree to 'data-mining'. Unlike in a traditional analytical study, there are not specific hypotheses being tested. Instead the use of 'exceedance' algorithms [63] are commonly employed to compare current trends against a baseline generated using historical data and accounting for factors such as seasonality, day of the week effects and changes in secular trend direction. This type of exceedance algorithm derives partially from the usual practice in SS of exploring the data using a form of temporal plot, which can be automated and translates well to analysis using alarm systems such as the multi-level regression approach [45], the quasi-Poisson regression based methods [63] used within PHE, or the Shewhart control chart based 'Early Aberration Reporting System' used in the U.S Centre for Disease Control.[64] These forms of analyses differ from the ML approaches typically associated with 'big data', and aberration detection algorithms are discussed during Theme Three.

In the context of SS, ML has been used for risk assessment and decision making support [65] and natural language processing of free-text [66], but in terms of analysis or insight generation, the application of analyses to syndromic data has been, perhaps surprisingly, limited. There are several challenges to the application of ML to SS. Firstly, syndromic data is rarely 'labelled' to allow an algorithm the opportunity to train as it might during supervised learning, depending on the purpose. This often stems from the lack of a 'gold standard'; for example in outbreak detection, distilling a true 'signal' associated with an outbreak beyond background cases is challenging (exemplified by Paper Five in this PhD [5]), and whilst studies have been conducted using simulated data [67-69], real-world examples in the literature which consider specificity and sensitivity, rather than simply reporting outbreaks found by SS are scarce.

Kulessa et.al [20] highlight that the use of Bayesian networks [70] and sum-product networks [71] which allow the underlying probability distributions of the data to be identified (and thus 'normal' behaviour captured) indicate that SS is a form of exceptional model mining.[72] As much of the data is unlabelled, unsupervised learning approaches are necessary. Unsupervised learning offers promise in terms of understanding which indicators most closely correspond to pathogens of interest through the use of convolutional neural networks [20] however again the usage for actual outbreak detection are rarer. Despite the limitations of using modern ML approaches to syndromic data itself, these concepts are becoming known as 'pre-syndromic' surveillance [73] as a public health 'safety-net' whereby techniques look to identify emerging patterns which public health practitioners hadn't previously thought to look for; with special potential utility for emerging or novel pathogens.[74]

An important absence from the HACE model in relation to SS is included in the DeMunro definition; *'transformation into Value'*. A central component of field epidemiology is that information is used for action [75] and in terms of SS, that action may be a risk assessment or communication to a stakeholder, but the over-arching aim is to provide the information to facilitate a public health response if necessary.

8.4 BIG DATA AND PUBLIC HEALTH

The concept of big data has huge implications for our understanding of human health, biology, disease and its epidemiology. There are now unprecedented levels of information about the individual; ranging from biological information on a micro scale at the level of the 'omics' sciences to information about the health or behaviour of that individual to the truly macro scale of the environment which individuals live and how they interact with others. The revolution in 'omics' and the increasing affordability of sequencing data [76] has led to the potential for vast amounts of information from an individuals' proteome [77], genome [78], transcriptome [79], and epigenome .[80] Alongside these vast data, is the potentially modifiable and unique individual data on the microbiome.[81] Ultimately, some of the greatest insights to health may come from the integration of these types of data [82], and although with current processing power incorporation into SS systems is at present unlikely, it is probable that in the future these may become increasingly relevant; changes in microbiomes may be detected and amenable to public health (or individual level clinical) intervention or those at particular risk of different diseases due to their genetic makeup may be sent differing public health warnings. Even today, sewage analysis systems are in operation for COVID-19 to monitor virus variance predominance.[83] [84]

Analogous to the wealth of individual level biological data, there is an unprecedented amount of data generated about the health of an individual, ranging from electronic health records (EHR) [85] and the digitisation of radiological images [86] and wearable technologies [87] to movement data. [88] EHR data are already in use as a source of data for SS purposes [89 90]; but there remains potential to improve the granularity of information through better coding and linkage with other systems. [91] This coincides with improvements to natural language processing; raising the possibility of automated reading and interpretation of both patient notes [92] and in addition the automated analysis of user generated social media posts about possible illness [93] or the rapid analysis of traditional media outlets [94] to provide insights. User generated information, such as internet searches has been described in the literature for a range of SS systems covering a wide variety of infectious diseases; and is described in detail in Theme One (Chapter 8).

The development of certain specific SS systems internationally has been expedited by MG events (such as the Olympics) where large scale datasets can support existing surveillance programmes to give indications about morbidity, or pertinently lack of excess morbidity in visiting populations at times of often intense political and societal scrutiny. There has been a recent 'call to arms' to incorporate new data sources and maximise SS's potential with the return of the Olympics to Europe in 2024 [95], and locally to the ReSST, the Commonwealth Games in Birmingham in 2022. As such, Theme Four of this PhD explores the role of SS during mass gatherings.

Beyond biological and behavioural factors, as the numerous reports such as the Black Report [96] and Marmot's 'Fair Society, Healthy Lives' Report [97] eloquently demonstrated, health inequalities and social determinants have huge implications for a population's health, and can impact the life course of individuals. Large scale datasets now allow us to quantify those determinants and the use of indices of socio-economic deprivation are now commonplace [98-100] in establishing links between illness and the socio-economic gradient and as described in Theme Two, data from syndromic systems can be used to support this type of epidemiological insight. Alongside socio-economic determinants of health, environmental data and how individuals interact with the environment can be used to provide insight and early warning of disease; meteorological data can be considered 'big data' in terms of volume, veracity and velocity and traditionally has combined with agriculture, transportation and the tourism industry data to generate insight.[101] However, there are clear relationships between health and the surrounding environment; the links between air pollution and respiratory, cardiovascular [102] and neurological diseases [103] are well established. Using these large health-related datasets or combining with non-health datasets to generate insight into the health of a population are the basis of SS systems globally, and Theme Two in this PhD explores how these rich data sources are being used to understand non-infectious disease epidemiology.

9. KEY FINDINGS FROM INDEXED PAPERS

The key findings from the indexed papers are described in Table C.

Table C: Key Findings from Indexed Papers

Paper One: Ambulance Dispatch Data Pilot.

The study demonstrated the development of a novel system pilot, and some of the challenges faced developing such a system. Specifically, that some, but not all ambulance dispatch data indicators had potential utility for SS purposes in England. Key lessons were identified in setting up the system which included the benefits of operation as part of a suite, ensuring stakeholder 'buy-in' and challenges of demonstrating value when extreme events are rare.

Abridged Abstract

Objective: To determine whether an Ambulance Data SS System (ADSSS) is feasible and of utility in enhancing the existing suite of PHE SS systems?

Methods: An ADSSS was designed, implemented, and a pilot conducted from September 01, 2015 through March 01, 2016. Surveillance cases were defined as calls to the West Midlands Ambulance Service (WMAS) regarding patients who were assigned any of 11 specified chief presenting complaints (CPCs) during the pilot period. The WMAS collected anonymized data on cases and transferred the dataset daily to ReSST, which contained anonymized information on patients' demographics, partial postcode of patients' location, and CPC. The 11 CPCs covered a broad range of syndromes. The dataset was analysed descriptively each week to determine trends and key epidemiological characteristics of patients, and an automated statistical algorithm was employed daily to detect higher than expected number of calls. A preliminary assessment was undertaken to assess the feasibility, utility (including quality of key indicators), and timeliness of the system for SS purposes. Lessons learned and challenges were identified and recorded during the design and implementation of the system.

Results: The pilot ADSSS collected 207,331 records of individual ambulance calls (daily mean = 1,133; range = 923-1,350). The ADSSS was found to be timely in detecting seasonal changes in patterns of respiratory infections and increases in case numbers during seasonal events.

Conclusions: Further validation is necessary; however, the findings from the assessment of the pilot ADSSS suggest that selected, but not all, ambulance indicators appear to have some utility for SS purposes in England. There are certain challenges that need to be addressed when designing and implementing similar systems.

Paper Two: Socioeconomic and Geographical Variation in GP Consultations for Allergic Rhinitis
The study demonstrated how syndromic data can be used to estimate the burden of disease and how that breaks down by different factors, specifically the burden which AR presents to GP practices, the consistent seasonal pattern of the disease and how it varied by demographic sub-groups, deprivation quintile and rural-urban characteristics, which needs to be considered during health care planning.

Abridged Abstract

Objective Allergic rhinitis (AR) is a global health problem, potentially impacting individuals' sleep, work and social life. We aimed to use a surveillance network of general practitioners (GPs) to describe the epidemiology of AR consultations in England.

Methods GP consultations for AR across England between 30 December 2002 and 31 December 2014 were analysed. Using more granular data available between 2 April 2012 and 31 December 2014 rates and rate ratios (RR) of AR were further analysed in different age groups, gender, rural-urban classification and index of multiple deprivation score quintile of location of GP.

Results The mean weekly rate for AR consultations was 19.8 consultations per 100 000 GP registered patients (range 1.13–207), with a regular peak occurring during June (weeks 24–26), and a smaller peak during April. Between 1 April 2012 and 31 December 2014, the highest mean daily rates of consultations per 1 00 000 were: in age group 5–14 years (rate=8.02, RR 6.65, 95% CI 6.38 to 6.93); females (rate=4.57, RR 1.12 95% CI 1.12 to 1.13); persons registered at a GP in the most socioeconomically deprived quintile local authority (rate=5.69, RR 1.48, 95% CI 1.47 to 1.49) or in an urban area with major conurbation (rate=5.91, RR 1.78, 95% CI 1.69 to 1.87).

Conclusions AR rates were higher in those aged 5–14 years, females and in urban and socioeconomically deprived areas. This needs to be viewed in the context of this study's limitations but should be considered in health promotion and service planning.

Paper Three: Environmental Factors Associated with GP Consultations for AR in London, England

This study demonstrated how syndromic data, when combined with other rich data sources, can be used to model the impact of different variables upon health care presentation. Specifically, how changes in pollen counts, temperature and pollutants were associated with presentations for AR.

Abridged Abstract

Objectives: To identify key predictors of general practitioner (GP) consultations for allergic rhinitis (AR) using meteorological and environmental data.

Methods: A retrospective, time series analysis of GP consultations for AR. The study population was all persons who presented to general practices in London that report to the Public Health England GP in-hours SS system during the study period (3 April 2012 to 11 August 2014). Primary measure was Consultations for AR (numbers of consultations).

Results: During the study period there were 186 401 GP consultations for AR. High grass and nettle pollen counts (combined) were associated with the highest increases in consultations (for the category 216-270 grains/m^3 , relative risk (RR) 3.33, 95%CI 2.69 to 4.12) followed by high tree (oak, birch and plane combined) pollen counts (for the category 260–325 grains/m^3 , RR 1.69, 95%CI 1.32 to 2.15) and average daily temperatures between 15°C and 20°C (RR 1.47, 95%CI 1.20 to 1.81). Higher levels of nitrogen dioxide (NO_2) appeared to be associated with increased consultations (for the category 70–85 µg/m^3, RR 1.33, 95%CI 1.03 to 1.71), but a significant effect was not found with ozone. Higher daily rainfall was associated with fewer consultations (15–20mm/day; RR 0.812, 95% CI 0.674 to 0.980).

Conclusions: Changes in grass, nettle or tree pollen counts, temperatures between 15°C and 20°C, and (to a lesser extent) NO_2 concentrations were found to be associated with increased consultations for AR. Rainfall has a negative effect. In the context of climate change and continued exposures to environmental air pollution, intelligent use of these data will aid targeting public health messages and plan healthcare demand.

Paper Four: The Utility of Syndromic Surveillance for large, subnational outbreaks of GI disease

This study found that syndromic surveillance systems in England did not detect examples of large, subnational GI outbreaks contemporaneously. Although automated statistical alarms highlighted potential changes in two of the outbreaks, these were assessed as low risk and public health action not initiated.

Abridged Abstract

Objectives: To determine if SS systems in England had identified large, retrospective sub-regional outbreaks of GI disease.

Methods: To investigate using SS for this purpose we retrospectively identified eight large GI outbreaks between 2009 and 2014 (four randomly and four purposively sampled). We then examined SS information prospectively collected by the Real-time SS team within Public Health England for evidence of possible outbreak-related changes.

Results: None of the outbreaks were identified contemporaneously and no alerts were made to relevant public health teams. Retrospectively, two of the outbreaks – which happened at similar times and in proximal geographical locations – demonstrated changes in the local trends of relevant syndromic indicators and exhibited a clustering of statistical alarms but did not warrant alerting local health protection teams.

Conclusions: Our suite of SS systems may not be suitable as means of detecting or monitoring localized, subnational GI outbreaks. This should, however, be considered in the context of this study's limitations; further prospective work is needed to fully explore the use of SS for this purpose. Provided geographical coverage is sufficient, SS systems could be able to provide reassurance of no or minor excess healthcare systems usage during localized GI incidents.

Paper Five: Syndromic Surveillance Findings During the 2012 London Olympics

This paper demonstrated the role of syndromic surveillance during the 2012 London Olympics, and how it was able to provide near-real time information to decision makers during the Events. Importantly, it clearly demonstrated the importance of the role SS can provide in providing reassurance to decision makers of discernible changes in community morbidity and provided a model for the use of SS in future mass gathering events.

Abridged Abstract:

Objectives: To investigate the impact of a large mass-gathering event on public health and health services as monitored in near real-time by SS of GP out-of-hours contacts and ED attendances, and to identify learning to aid the planning of future events.

Methods: ED and GP out-of-hours data for London and England from July 13 to August 26, 2012, and a similar period in 2013, were divided into three distinct time periods: pre-Olympic period (July 13-26, 2012); Olympic period (July 27 to August 12); and post-Olympic period (August 13-26, 2012). Time series of selected syndromic indicators in 2012 and 2013 were plotted, compared, and risk assessed by members of the ReSST. Student's t test was used to test any identified changes in pattern of attendance.

Results: Very few differences were found between years or between the weeks which preceded and followed the Olympics. One significant exception was noted: a statistically significant increase (P value = .0003) in attendances for "chemicals, poisons, and overdoses, including alcohol" and "acute alcohol intoxication" were observed in London EDs coinciding with the timing of the Olympic opening ceremony (9:00 PM July 27, 2012 to 01:00 AM July 28, 2012).

Conclusions: SS was able to provide near to real-time monitoring and could identify hourly changes in patterns of presentation during the London 2012 Olympic Games. Reassurance can be provided to planners of future mass-gathering events that there was no discernible impact in overall attendances to sentinel EDs or GP out-of-hours services in the host country. The increase in attendances for alcohol-related causes during the opening ceremony, however, may provide an opportunity for future public health interventions.

10. DISCUSSION, IMPACT AND CONCLUSION

10.1 THEME ONE: DEVELOPING SYNDROMIC SURVEILLANCE SYSTEMS, CASE STUDY USING AMBULANCE DISPATCH DATA

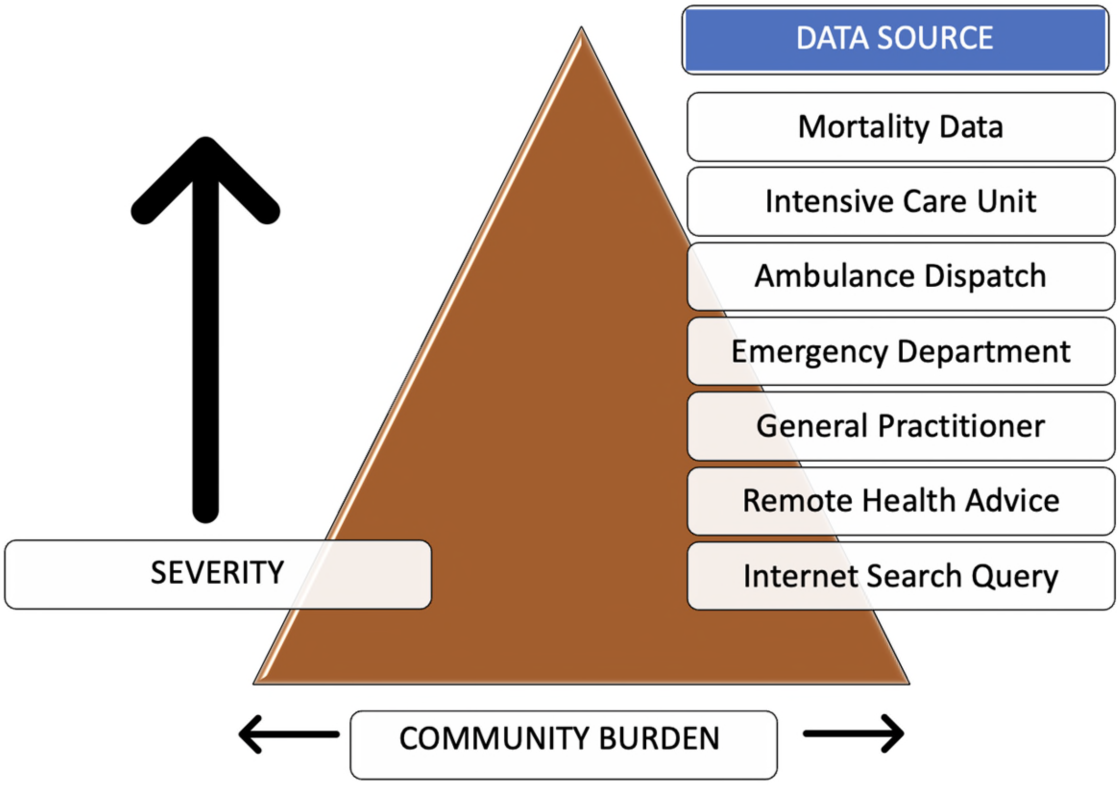
In the existing literature, the emphasis has been on the statistical methodologies which underpin syndromic surveillance systems; with far less information on the practical development and application of such systems.[104] Through the publication of the development of systems such as the first indexed paper [2] in this thesis, the practicalities of piloting and operating such systems are shared, which is critical for effective public health practice and supporting/enhancing existing epidemiology and surveillance programmes. The role of the relationships with stakeholders was clearly identified in the indexed paper One, but is rarely reported or discussed in detail in the literature.[104]

The first stage of development of a syndromic surveillance system is generating an understanding of what the public health purpose of the system is, and what added value that system creates to the stakeholders, decision makers and wider public health program of the operating country or geographical region. Without a clear benefit identified it is very difficult to justify the resource required to develop such systems.

Part of this understanding comes from the type of data used to inform the system. Currently in operation across the globe, there are numerous types of data employed in SS. These different data sources range from large scale health data such as GP consultations [14], ambulance dispatch data [2], telehealth [23] ED admissions [21] and 911 / emergency medical calls [105] to non-health related data sources, such as social media analysis [106], school absenteeism [25] or web searches.[27]

In terms of the roles of the different systems, these could be viewed as a ‘spectrum’ of severity. To date, in the literature there has not been a clear picture of the traits; abilities and facets of each individual data source and its relevance to individual systems. The concept of viewing data sources according to severity allows visualisation of why and how they might be used. For example, the patient pathway for less severe, or non-urgent health queries might involve large-scale internet searches or calls to a telehealth helpline (for example in England, NHS 111). Increasing severity of symptoms or disease might warrant an overall smaller numbers of presentations progressively through consultations to GP's, ED attendances or if the problem is acute and / or severe or life threatening, an ambulance call. Each part of this pathway represents an opportunity for information gathering for SS purposes; and provides potentially different intelligence about the pattern of morbidity in the community (Figure A). This is visually represented in Figure A. The analogy of classification of SS systems by severity would allow stakeholders and those responsible for understanding SS system's output better insight into the potential and limitations of each different data source.

Figure A: Syndromic Surveillance Systems in operation; Data Source by Severity



Those systems described in Figure A are in active operation globally. Beyond this there are a range of 'novel data sources' which aren't linked to the collection of 'traditional' health record data,[107] but which offer potential for use. In 2008, the first description of using trends in internet searches to quantify the behaviour of influenza (Google 'Flu Trends) was published [108], although this wasn't without controversy; especially in terms of potentially overestimating the community burden.[109] The use of Google Flu Trends has subsequently been reappraised using more sophisticated analytical methods which have demonstrated their potential utility.[110] The original study did however demonstrate the potential of these types of internet based data for syndromic surveillance use, and there has been an increase in publications such that a systematic review was recently published [111] which highlighted the increasing number of publications, and concluded that there was some utility in web retrieval for syndromic surveillance, in particular for the monitoring and early predictions of an epidemic. There has also been an increasing number of papers which describe the use of machine or deep learning techniques to mine Twitter data; with varying degrees of success.[112-115] Other authors [111] have highlighted the limitations of using social media or internet search data for SS, suggesting that it should at most act as an adjunct to other surveillance systems, and the lack of these systems reported in the literature which are in actual practice is notable, despite the duration of this data being available.

10.1.1 AMBULANCE DISPATCH DATA

In Paper One presented in this PhD [2], this was the first time in the UK that large-scale ambulance dispatch data had been used for SS purposes. Using ambulance dispatch data had been identified as a potential system to complement the existing PHE syndromic surveillance programme; especially at the 'severe' end of the disease spectrum. Although some information was already captured in the existing national ED SS system, at the time this was a sentinel ED system with limited geographical coverage across England and sub-regions.

Ambulance data have been used for SS purposes in other countries although the degree of usefulness reported has varied. In New York City (New York USA) [116] and Denmark,[117] ambulance data have been effectively used to monitor trends in influenza-like illness, but the experience in Melbourne, Australia highlighted difficulties in interpretation due to an observed high degree of background "noise" associated with ambulance data.[118] More recently, a scoping review has been published [119] which identified 20 peer reviewed publications and 24 publications in the 'grey' literature describing 44 studies and systems. Most of these studies described the outcomes of temporary systems, and – in common with our experience – noted the relationship between ambulance dispatch data and respiratory illness, and the potential benefit of timeliness of information above traditional surveillance systems.

After the publication of Paper One, and the continued operation of the pilot in the West Midlands, the programme has developed further to encompass all ten NHS ambulance service providers in England; becoming the first National Ambulance Dispatch Data SS System (NASSS) globally.[44]

In common with other authors, during the initial pilot we found limited use for other ambulance data indicators beyond respiratory during the initial pilot.[116 117 120] However, during the construction of the system we adopted a pragmatic approach using only limited information for coding. Eleven 'chief presenting complaints' (CPCs) were used, which were the codes assigned to individual calls by the call handler and provided to the ambulance during dispatch e.g, "breathing problems". These CPCs were based on the emergency call hander's clinical assessment of the caller and their judgement of the most likely cause of the call. The calls themselves, however, were based on a more granular, algorithmic assessment during the call. This more granular assessment data is used in the UK in the NHS 111 syndromic system [23], and is an opportunity to further study the utility of ambulance data.

A key finding from Paper One was the benefit of using the other existing systems in the suite operated by ReSST for triangulation and sense-checking of findings. At the time of this observation this triangulation was performed through epidemiologist interpretation and done manually through the examination of individual systems; this process is still in practice today. In the literature, however, methods are beginning to emerge which combine multiple data sources and the detection of anomalies (such as an outbreak) through the use of ML techniques [20 106]; in essence using statistical approaches to address the triangulation of data discussed in the paper presented in this PhD. Through training algorithms using syndromic data from multiple systems combined with historical temporal trends of confirmed laboratory cases, these types of techniques may offer the opportunity to use codes from multiple systems to produce 'machine-selected' indicators which match epidemiological trends of different infectious agents better than existing human-designed indicators. There may be additional benefits from this in the context of differentiating between pathogens; of particular importance during future winters where a syndromic system that could provide differentiation between influenza, COVID-19 and other respiratory viruses would be of high value to decision makers.

10.2 THEME TWO: THE APPLICATIONS OF SYNDROMIC SURVEILLANCE IN NON-INFECTIOUS DISEASE EPIDEMIOLOGY

As described earlier during this Thesis, one of the original and primary roles of syndromic surveillance was the detection and quantification of outbreaks of infectious diseases. [33] However, these rich data combined with epidemiological expertise have proved to have utility far beyond this original scope; providing insight into the epidemiology of diseases, the burden on healthcare systems, the effectiveness of public health interventions and morbidity associated with natural phenomena. During this theme, how syndromic data has been used to address some of these issues is described, alongside some of the types of analysis frequently used to understand syndromic data and how this may be developed in the future.

Data from syndromic surveillance systems is of particular use for those diseases or syndromes where there isn't laboratory confirmation, nor other surveillance systems in place to monitor the epidemiology; and to illustrate this, papers three and four presented in this PhD focus on using syndromic data for such a disease; allergic rhinitis. The first of these uses standard descriptive epidemiology, and the second time series modelling to understand the aetiology of this disease and its relationship with other factors such as socio-economic status, meteorological conditions and air pollution.

10.2.1 SYNDROMIC DATA SETS FOR NON-INFECTIOUS DISEASE EPIDEMIOLOGY

There is a growing body of literature using syndromic datasets for furthering the understanding of non-infectious disease epidemiology, and Paper Three and Paper Four support this. In this thesis an argument is presented whereby SS for non-infectious disease epidemiology appears most useful in those instances where the disease, risk factor or condition presents – at a population level – in a similar manner to an infectious disease agent. This is either in terms of defined 'outbreaks' (i.e. linked in time, space or person) often in response to an external agent such as disaster or environmental phenomenon, or otherwise with a regular seasonal pattern, like is observed with seasonal influenza, or for non-infectious disease; like allergic rhinitis. This has become particularly important during the COVID-19 pandemic, as one of the key non-COVID related impacts has been the impact on population-level mental health.[121] Syndromic-data has been used in the US to quantify increases in mental health-related visits for children aged 5-11 and 12-17 years of 24% and 31% respectively between 2019 and 2020, leading to calls for public health promotion of coping and resilience strategies.[122] There is a demonstrated role in SS providing information after terrorist attacks on population mental health; increases in mental health related visits to ED's were observed subsequent to the Boston Marathon bombings in 2013 [89] and SS of twitter data was applied following the Paris terrorist attacks during 2015 to detect clusters of 'fear' and 'sadness' both geographically and demographically.[123]

At the severe end of the mental health spectrum, ED's are a vital source of information about health-care seeking behaviour amongst individuals with risk factors for suicide or suicidal attempts.[124] Despite this, in the UK (and most other countries) suicide data is difficult to obtain due, in part, to delays in legal processes (such as the time lag between an unexpected death report and coroner's conclusion at inquest), leading to the Royal College of Psychiatrists developing a support pack to local areas to develop real-time surveillance programmes.[125] The analogies in the literature between suicidal ideation and infectious disease outbreaks have been made with the controversial use of the term 'contagion' [126] used for clusters of suicides linked by time, place or person which perhaps fits with the syndromic concept of identifying changes in population level morbidity in near-real time to lead to geographically and demographically targeted and timely public health intervention. A number of attempts to use syndromic systems for this purpose have been published, with varying success.[127 128]

Although there are challenges to the identification of clusters of poor mental health using SS; there are multiple risk factors which can readily be identified by SS systems. The most frequently cited of these is alcohol or substance abuse; in particular as this is readily coded in ED departments, and numerous publications report using SS to monitor harm due to alcohol [129 130], poisonings or drug overdoses.[131 132] In Paper One in this PhD, the Ambulance Dispatch Data syndromic data identified unexpected local temporal-spatial and demographic clusters of alcohol use potentially amenable to public health intervention.[2]

Syndromic data is frequently used to study the correlation between events and the impact that has on morbidity. Flexibility has been highlighted as a strength of SS, [133] and adaptability to different events is evidenced though numerous examples. These can be comparatively acute, such as determining the impact of football matches on cardiovascular events [134], the impact of terrorist events [135] or conflict.[136] There is a clear role for SS in the study of natural disasters or phenomenon; this has involved surveillance of morbidity associated with flooding [137], earthquakes [138] or the impact of volcanic plumes.[139]

The concept of SS data being of particular use for those diseases which are 'outbreak-like' in their presentations at population-level is evident with some non-infectious respiratory diseases. Presentations of asthma are associated with environmental factors [140], and syndromic data has been used to quantify the burden of asthma on ED attendances [141], and to justify the development of early warning public health alerts to asthmatics.[142]

Elucidating the relationship between health, meteorological and environmental factors demands large-scale data, and is necessarily observational, which makes syndromic data one of the most widely used data sources for this. SS has been used to quantify the impact of air pollution 'events', and how these impact across national borders.[143] The relationship between extreme weather such as heat-waves [144] or cold-weather [145] [146] and illness has been well described using syndromic data. The benefits of SS data for these purposes are that firstly it allows the relationship between hazards and illness to be quantified; providing decision makers with estimates on the expected burden on healthcare services, and secondly informs decisions about the need for public health interventions such as preventative messaging.

The second indexed paper in this PhD uses syndromic data but combines it with census-level geographical data to provide a description of allergic rhinitis by geography, urban-rural classification and indices of multiple deprivation score (IMD) to provide a geographical level estimate of socioeconomic deprivation. Socio-economic deprivation is a key driver of many diseases which exhibit a gradient whereby often people from more socio-economically deprived areas suffer greater morbidity and poorer health outcomes.[96] [97] Syndromic data offers potential for understanding these relationships but is also associated with limitations. As discussed in the paper, people in socio-economically deprived, and urban areas were more likely to be found to attend their GP for allergic rhinitis (AR). There were a range of possible reasons for this; including differing presentation patterns between socio-economically deprived and affluent areas, different behaviour due to the need to pay for prescriptions, and the possibility that a socio-economic gradient is evident for this disease and / or differing levels on antigenicity due to differences in exposure in urban versus rural environments. Whilst syndromic data can provide large scale, population wide snapshots; it remains critical to interpret findings in the context of the data collected. It is imperative to explore inequalities and highlight their existence when and where possible as evidence such as that provided in the paper is necessary to generate hypotheses that might eventually support public health programmes to reduce inequalities; and this paper is one of a growing number of examples of how syndromic data can be used to both develop the evidence base and detect changing patterns of inequalities.[15]

The third indexed paper in this PhD demonstrated how syndromic data, when combined with environmental and meteorological data can provide powerful insight into the drivers behind illness. The paper demonstrated that change in grass, nettle or tree pollen counts, temperatures between 15°C and 20°C, and (to a lesser extent) NO_2 concentrations were found to be associated with increased consultations for AR. Rainfall has a negative effect. Intelligent use of these data will aid targeting public health messages and planning healthcare demand, particularly in the context of a changing environment and the challenges of global warming.   The challenging time-series modelling used in the paper demonstrated how effective data used for syndromic purposes can be for providing insight into non-infectious disease epidemiology.

10.3 THEME THREE: SYNDROMIC SURVEILLANCE SYSTEMS IN THE IDENTIFICATION OF LOCALISED OUTBREAKS OF GASTROINTESTINAL INFECTIOUS DISEASES

The original intended purposes of SS were of detection of either the deliberate release of an agent capable of causing disease or monitoring of outbreaks of infectious disease. The initial rationale for developing these systems was that they had the potential for provision of situational awareness to decision makers in a timelier fashion than non-SS systems which relied upon the processing of samples in a laboratory.[33]

As early as 2006, strengths and limitations of SS systems were being described. A systematic review by Hope et.al [133] identified the following strengths; i) the ability to detect community wide seasonal influenza outbreaks ii) timeliness of data availability iii) completeness of data iv) alleviation of community concern when outbreaks are elsewhere v) case-finding when outbreaks identified and vi) flexibility in being able to conduct surveillance for new and emerging issues. Despite this, they also identified that there were significant limitations to localised outbreak identification; an inability to distinguish between background 'noise' and true signals, the burden of false alarms associated with low specificity and positive predictive value, and a major constraint of unavailability of denominator data at lower-level geographies.

In Paper Four, the limitations of SS for outbreak detection of large, sub-regional GI outbreaks were clear. None of the five identified outbreaks were identified contemporaneously, nor on retrospective review of the syndromic data would have warranted further alerting. The dual constraints of 'noise' and distinguishing 'true' outbreaks were clearly evidenced from this paper.

Concluding this work, a clear distinction should be made in terms of the definition of what should be considered a localised outbreak and what is a community wide change in the secular trends or seasonality of a pathogen. Hale et.al [147] helpfully define a localised GI outbreak as an unexplained, spatially and temporally localised increase in the fraction of GI consultations amongst all consultations. Although this definition was devised in relation to veterinary SS; the concept is applicable to humans, and also non-GI communicable disease.

There is distinct evidence for the role of SS in monitoring epidemic-level respiratory pathogens such as influenza [148] or respiratory syncytial virus [149] which happen most winters in temperate climates. Near-real time surveillance allows community onset and burden to be estimated, and is a key objective for winter surveillance systems.[150] At the time of writing, publications describing the ability of SS to adapt to the challenges of COVID-19 are few, but despite initial challenges, early papers suggest utility in monitoring the pandemic.[151] [152]

Despite the utility in widescale epidemics, in relation to localised outbreaks the limitations of SS described by Hope et.al [133] and in Paper Four still remain today, and there remains a lack of published, peer reviewed and real-world evidence to suggest that localised outbreak detection has improved with more granular data sources or better detection techniques.

Understanding utility in SS can be challenging. Within the published literature there are few examples of either how to pilot, set up, operate on a routine basis or evaluate a SS system.

The need for evaluation of systems was recognised as early as 2004, where 35 detection and decision support systems were evaluated, but found to be deficient in a number of domains [153] leading the Centre for Disease Control (CDC), USA to develop a framework for the evaluation of surveillance systems for outbreak detection.[154] The focus of the framework is on the evaluation of timeliness of outbreak detection, and understanding the balance between sensitivity, predictive positivity value and predictive negative value. This consensus framework was preceded slightly by a framework for the evaluation of SS systems [33] developed in 2003 (and still widely used today), which composed of five main components; description of the system (e.g. stakeholders, how the system operates), outbreak detection capacity (e.g. sensitivity, timeliness of detection), data quality (e.g. representativeness, data completeness, reliability), conclusions and recommendations.

Both of these frameworks allude to sensitivity and specificity; however, this is often very difficult to determine in relation to localised outbreaks for a number of reasons. Firstly there is there is the problem of 'noise', or high background rates of sporadic cases which mean spotting genuine outbreaks is difficult. In Paper Four, one posited reason for the lack of identifying outbreaks, was that GI disease may be largely mild in the population, with only a small number seeking healthcare. This is supported by the findings from the Infectious Intestinal Disease study in the UK [155] which demonstrated that the community burden of infectious intestinal disease far outweighs the burden on presentation to healthcare; estimating that for every 10 GP consultations for infectious intestinal illness, there were 147 community cases. The implication for SS using healthcare data sources – and Paper Four - is that even large-scale outbreaks would have a low proportion of cases seeking healthcare, and identifying a signal becomes increasingly difficult, as demonstrated in the paper presented for this PhD. Secondly, there is often a lack of a 'gold standard' to compare against; as many localised outbreaks go undetected, they will not be recorded by public health authorities, making determining sensitivity and specificity of syndromic systems for this purpose difficult. The lack of data which might be considered accurately 'labelled' (outbreak related vs sporadic cases) precludes the use of supervised machine learning techniques for outbreak detection.

These difficulties in identifying GI outbreaks from syndromic data have been reported elsewhere [156], and the utility of SS for identifying localised GI outbreaks is not clear, with mixed results reported in the literature using both simulated and real-world data.[31 157-159] The findings from Paper Four had a large impact on how SS is operated in England; in terms of a reduction in confidence in the suite of systems to accurately identify such outbreaks, and clear caveating of messages to stakeholders when dealing with GI outbreaks.

Theoretically, SS may be useful for the detection of waterborne outbreaks (e.g. cryptosporidiosis), where a contaminated water source could expose a high number of individuals in a short time period. As such, this is a relatively well-studied area and has been the subject of two systematic reviews; the first in 2006 [160] and more recently in 2020.[161] Hyllestad et.al [161] identified six simulation studies and ten retrospective studies; there conclusion was that there was no conclusive evidence on the effectiveness of SS for the detection of waterborne outbreaks.

10.3.1 STATISTICAL ABERRATION DETECTION

A key aspect of SS is the ability to analyse and interpret vast quantities of data, and this is usually done through the automated generation of statistical alarms; it is this type of statistical alarm which informs Paper Four in this PhD. Globally there are a number of different algorithms which have been or are in operation. Cumulative Sum (CUSUM) methodologies were used in the early literature to monitor disease counts in discrete geographic areas [162] and applied to SS but adapted for small counts [163] and more recently, regression methodologies such as quasi-Poisson regression [164] or multi-level regression approaches.[45] Commonly used in the U.S.A is the Shewart control chart based Early Aberration Reporting System (EARS).[64] Bayesian methods have also been used for anomaly detection.[165]

The intended outcome of these sophisticated, automated algorithms is to alert public health teams to potential problems (whilst maintaining a manageable level of sensitivity), and there is a developing body of literature comparing the merits of individual approaches.[63 166 167]

Beyond algorithm detection; and something that is often described less in the literature, is the role of skilled epidemiologists in maintaining surveillance systems, risk assessment and stakeholder relationships.[104] In Paper Four, the ultimate decision making around if the systems had correctly identified the outbreaks was ultimately based on epidemiological risk assessment of the data. In England, a risk assessment process was developed in advance of the 2012 London Olympic Games [46] and remains in daily operation, incorporating other available epidemiological data and triangulation with data from other systems, and clear lines of communication between the surveillance team and decision makers with auditable decision trails. This aspect of SS is critical to effective operation of these systems for public health intervention but is far less well documented in the published literature.

10.4 THEME FOUR: UTILITY IN AND DURING MASS GATHERINGS

The World Health Organization (WHO; Geneva, Switzerland) defines a MG as "any occasion, either organized or spontaneous, that attracts sufficient numbers of people to strain the planning and resources of the community, city, or nation hosting the event. [168] A recent systematic review [169] investigated the public health threats associated with mass gatherings; the authors classified MGs primarily into three groups; religious (such as the Roman Catholic World Youth Day or the Hajj) festival (e.g. Glastonbury) and sporting events (e.g. the Olympics or the Commonwealth Games) with different public health concerns associated with each. Religious MG's were most associated with infectious diseases, road traffic accidents and environmental health problems. At MG sporting events infectious diseases, alcohol and drug related problems were most frequently reported and at festival MGs, alcohol and drug related problems most reported.

These type of mass gatherings and associated public health threats are suitable for monitoring using SS, and the WHO provides have made explicit recommendations on the use of surveillance during such events [169] and there are numerous examples of SS having been deployed for sporting events [170-173], religious pilgrimages [174 175] and festivals or political conventions.[176 177]

One of the key applications of SS is the provision of reassurance to decision makers of the absence or lack of unusual morbidity and mortality amongst the population as was the case during the 2012 Olympics. To provide that reassurance, however, confidence in that system is necessary. Success of SS systems, and that confidence in their capabilities and limitations is dependent on sustainability, and integration into existing surveillance pathways and again the relationships between the surveillance team and decision makers were key.[104] The system described in the 2012 Olympics [6] was well rehearsed prior to the games, and was integrated within existing surveillance and stakeholder networks. It has subsequently been described as a 'legacy' of the Olympic Games in England.[178] In Paper Five, the key finding was that of the role of reassurance to decision makers which SS systems played, and again in the relationships between operators and key stakeholders.

Despite the clear importance of professional relationships and networks during mass gatherings, SS of MGs is another area which will benefit from advances in data capture, rapid analysis and technologies. Baselines are key to effective operation of systems, however Paper Five identified the importance of denominator populations; and how they may change as result of an MG. MG's necessarily cause changes in populations (and differences in utilisation of healthcare); be it influx of athletes, spectators or pilgrims or the efflux of local populations. Population level movements have been used to predict the spread of emerging infectious disease [179] and during the current COVID-19 pandemic, tracking human movement behaviours have allowed monitoring of the impact of societal-level interventions.[180] During a mass gathering event, there are a range of data sources which are potentially available to understand human movement data; mobile phone data records have been used to track the mobility of over 15 million Kenyans and correlate spatially-explicit phone data with malaria incidence.[181] Commuter data has been used to model local movements [182] and at an even more local level, the collection of real-world data from wearable devices merged with environmental and location data via deep learning approaches can predict emotional state.[183] Alongside understanding of the population movements during mass gatherings, the real-time monitoring of mass gatherings through the use of geo-tagged social media feeds such as Twitter can provide situational awareness [184] both in terms of sentiment analysis [185] and disease detection; although it is clear that further work is necessary for validation of such systems.[186]

These types of data gathered through the 'internet of things' offer great potential for insight during mass gatherings. However, these very granular – and highly personal – data streams, of course raise privacy and ethical issues.[187] [188] The ethics surrounding SS have been considered since close to its outset; with an expert meeting convened during 2007 which recognised the dichotomy between individual rights around privacy, confidentiality, the lack of implicit consent for data sharing and the governmental needs of protecting the population.[189] A recent paper [190] describing the ethics of communicable disease surveillance highlight some pertinent risks specific to SS; firstly it points to a perceived lack of transparency of methods and indicators which can lead to errors, such as the Google 'Flu Trends initial failures.[109] Secondly; a risk is that some data sources do not capture basic demographics which may lead to a failure to determine if under-represented groups are appropriately captured in the systems, become disadvantaged as a result and conventional epidemiological research tools are difficult to apply to syndromic data. Perhaps prophetically, lastly the argument is made that the most important risk is that of public fear, ignorance and mistrust and that "*ill-informed media scrutiny and political risk aversion could prevent or delay the incorporation of de-identified personal health data into Big Data – based public health surveillance, despite the benefits".*[190] At the time of writing this thesis, the date for the 'general practice data for planning and research' programme of work unifying general practitioner records into one pseudo-anonymised database for research has been delayed from 01 July 2021 to 01 September 2021 largely due to public concerns, and the government facing legal challenge.[191] These ethical issues are particularly acute around mass-gathering surveillance, where necessary enhanced surveillance (which can often become routine business) is introduced and may be at odds with the reasons the public attend the gathering. The argument that a framework to guide policy, minimise risk and build public trust is necessary becomes especially pertinent during a MG setting.

10.5 IMPACT OF INDEXED WORKS

The field of SS in academic literature is currently relatively small and does not attract a huge number of citations. The work submitted in this Thesis has been completed as a collaboration with the ReSST within Public Health England and the major impact of the work has been the application of the indexed works in public health practice. The work which has been submitted in this thesis has a significant impact on the operation of ReSST and in acknowledgement won a National PHE Prize for translational research.

The extensive observational study that investigated the ability of systems to identify outbreaks of GI disease found that systems would not reliably identify GI outbreaks. This often previously cited possible function of SS was demonstrated to be ineffective; the result has been that information on GI related syndromes is provided with evidence-based caveats and allowing ReSST to provide a more accurate picture to decision makers on systems' abilities.

The work on describing the experience of SS during the 2012 Olympics has been used when both the ReSST and PHE Global Health teams have advised other countries on the development of similar systems during Mass Gatherings and will directly inform planning of use of these systems during the Commonwealth Games, planned for Birmingham 2022. As a result of this work, I was invited to Rio de Janeiro, Brazil as an expert in SS to help in the translation of the research in developing new, participatory surveillance systems in advance of the 2016 Rio Olympics.

The paper on allergic rhinitis, using SS data combined with other data sources have not yet had chance to accrue citations, but the methods which were employed during its development have influenced future research approaches within the ReSST team. The same methodologies have been employed, combining unique datasets to provide new insights into important conditions such as asthma and the link with health inequalities. Syndromic data had rarely been used to describe such links before.

The work developing a pilot ambulance SS System in the West Midlands directly informed and led to the development of the world's first Nationwide Ambulance SS System (NASS) which has recently started operation. The same platform, development codes, indicators and template outputs developed during the pilot study were adopted in the National system, which is currently in operation and providing novel insights into activity at the severe end of the disease spectrum.

From the body of work presented here, there is a significant role for large, healthcare related data to support both surveillance and epidemiology. There were a series of recommendations based on the findings which have been adopted by the ReSST team and have had direct policy implications. These broad recommendations, and the impact of these are listed in Table D.

Table D: Adopted Policy and Key Strategy Recommendations from Indexed works in this thesis, and where available applicable evidence

• There is Public Health utility in an Ambulance Dispatch Data SS System, and this would bolster the current suite of SS systems operated in England. *This paper and work led to the CEO's of the ten Ambulance Trusts in England agreeing to data sharing to create the National Ambulance Dispatch Data SS System, which is the*

first of its kind to be operated on a National basis. This work has also opened up discussions with CEOs of Scottish, Welsh and Northern Ireland ambulance Trusts to develop a UK-wide ambulance surveillance system.

- There is added utility in operating a ‘suite’ of SS systems across the spectrum of illness type and severity. *This supported the expansion of the ReSST to include the NADSS from the severe end of the spectrum and enhance capabilities at the less severe end of the spectrum, leading to the development of a supplementary SS system monitoring NHS online assessment data to support COVID-19 and ‘business as usual’ needs.*
- The utility of SS systems in England to detect outbreaks of sub-regional GI outbreaks is limited for this purpose. *The supporting literature on this topic is equivocal as to the utility of SS systems. However, Paper Four provided a real-world assessment of the systems in England. As such, this is considered the most reliable assessment of ReSST systems and as such, messages to stakeholders in relation to GI are caveated that systems may not reliably identify GI outbreaks at the local level.*
- Data used by SS can be effectively combined with other data sources to provide significant insight into healthcare behaviours associated with non-infectious disease epidemiology, in terms of socio-economic, demographic and geographic distribution, and modelling to understand the relationship with other variables. *The indexed papers on Allergic Rhinitis were able to describe multiple novel findings relevant to healthcare planning, and in the context of a changing climate. However, they were also able to demonstrate the effectiveness of these data for this purpose; subsequent work has built on the methodologies described for other non-infectious diseases such as Asthma.*

10.6 CONCLUDING REMARKS

This PhD is written at an opportune time. The ongoing COVID-19 pandemic has highlighted the importance of 'data' and 'surveillance' and coincides with the confluence of generalised increased data collection, processing power and new computing techniques during the last two decades, which have the potential to revolutionise insight into the health state of the population.

The title of this thesis is 'using large scale datasets to support epidemiology and surveillance', and it clear from the examples provided how this can be achieved in the context of SS delivery, both in real-time, and using these collated data to generate new epidemiological insight. Through the indexed works, and how they are contextualised in the current literature, it has been possible to explore some of the key issues and applications of SS; from developing systems using novel data sources, understanding how effective these systems are (and their limitations) in their original remit of outbreak detection and in supporting mass gatherings to the emerging and critical role they are beginning to play in elucidating information in non-infectious epidemiology.

At the outset of the thesis four specific research questions were outlined, which have subsequently been discussed extensively and evidenced by the indexed papers and supporting literature. In summary;

Research Question One: How can different data sources be used for the spectrum of SS purposes, and specifically what additional utility does ambulance data bring?

There is a large, and growing range of data sources which can have utility for SS purposes. Importantly, matching that data source to the research question which is being asked is critical, and we argue that there is a 'spectrum' from which different data sources can add public health insight which broadly matches the spectrum of disease severity in which individuals might use to present to healthcare. Ambulance data represents the more 'severe' end of that spectrum and offers potential utility in enhancing the current suite of systems operated in the England.

Research Question Two: What is potential of UK SS systems to provide insight into non-infectious disease epidemiology?

By using two case studies demonstrating the usage of syndromic data, combined with other large-scale datasets, we were able to demonstrate the role that data used for SS purposes can play in providing non-real time insights into non-infectious disease epidemiology. These multiple, rich data sources offer the opportunity to understand socio-economic, geographic and temporal distribution of healthcare presentations, and how they associate with other possibly explanatory variables.

Research Question Three: What is the utility of SS in detection of outbreaks of infectious Gastrointestinal disease?

We were not able to fully answer this question but based on the pragmatic results found in the indexed paper, we were able to determine that SS systems in England at the time would not have reliably identified large, sub-regional outbreaks of infectious Gastrointestinal disease, and as such, we anticipate the utility is limited.

Research Question Four: What is the utility of Syndromic Surviellance during Mass Gatherings?

Using the case study of the London 2012 Olympic and Paralympic Games, which was one of the largest Mass Gathering events in England during recent history, we were able to demonstrate the applicability and effectiveness of SS systems in England of providing near-real time information on community morbidity to stakeholders and decision-makers. Importantly, the role of reassurance of the lack of increased pressure on healthcare services due to any cause, was a key finding and role of SS during the 2012 Olympics.

Much of the literature surrounding SS is focussed on the data science aspects, perhaps naturally due to the large data sets and complex analytics that form its basis. However, far less reported is the human aspects; the stakeholder relationships and epidemiological risk assessments which turn the data and statistical aberration alarms into public health action. The papers presented in this Thesis on the Olympics, developing the ambulance system and identifying GI outbreaks focus on the daily operation of SS; and the importance of people is evident throughout and should be a key part of the growing evidence base for SS going forward.

These operation-focussed papers also highlight the importance that the data sources form part of a “suite” of systems, and a ‘service’, rather than stand-alone data sources. This provides multiple benefits, allowing observed events to be triangulated with information from other systems (and vice versa), the sharing of development and operational expertise, and additional resilience in the event of problems with a single system; and as demonstrated can cover the spectrum of community disease states and severity.

As demonstrated through contextualising these papers, there are increasing opportunities to use novel and developing technologies to improve SS, and techniques such as machine learning will undoubtedly play a role. This confluence of increasingly large and available datasets also allows the potential for greater insight into non-infectious diseases, and understanding socio-economic and differences in presentation, as evidenced during Theme Two.

The seismic shifts in analytics capability, and the global COVID-19 pandemic have highlighted the importance of the small but growing speciality of SS.

11. REFERENCES

1. Fairchild A, Bayer R, Colgrove J. Privacy and public health surveillance: the enduring tension. *AMA Journal of Ethics* 2007;9(12):838-41.

2. Todkill D, Loveridge P, Elliot AJ, et al. Utility of ambulance data for real-time syndromic surveillance: a pilot in the West Midlands region, United Kingdom. *Prehospital and disaster medicine* 2017;32(6):667-72.

3. Todkill D, Loveridge P, Elliot AJ, et al. Socioeconomic and geographical variation in general practitioner consultations for allergic rhinitis in England, 2003–2014: an observational study. *BMJ open* 2017;7(8):e017038.

4. Todkill D, Gonzalez FdJC, Morbey R, et al. Environmental factors associated with general practitioner consultations for allergic rhinitis in London, England: a retrospective time series analysis. *BMJ open* 2020;10(12):e036724.

5. Todkill D, Elliot A, Morbey R, et al. What is the utility of using syndromic surveillance systems during large subnational infectious gastrointestinal disease outbreaks? An observational study using case studies from the past 5 years in England. *Epidemiology & Infection* 2016;144(11):2241-50.

6. Todkill D, Hughes HE, Elliot AJ, et al. An observational study using English syndromic surveillance data collected during the 2012 London Olympics–what did syndromic surveillance show and what can we learn for future mass-gathering events? *Prehospital and disaster medicine* 2016;31(6):628-34.

7. Raghupathi W. Data mining in health care. *Healthcare informatics: improving efficiency and productivity* 2010;211:223.

8. Groves P KB, Knott D, Kuiken SV. The 'big data' revolution in healthcare; Accelerating value and innovation. Available online at: http://repositorio.colciencias.gov.co/bitstream/handle/11146/465/1661-The_big_data_revolution_in_healthcare.pdf?sequence=1&isAllowed=y. : Centre for US Health System Reform Business Technology Office, McKinsey & Company,; 2016 [accessed 22/01/2020].

9. Ziemann A. Assessment of syndromic surveillance in Europe. *The Lancet* 2011;378(9806):1833-4.

10. World Health Organisation. Surveillance https://www.who.int/emergencies/surveillance2020 [accessed 05 September 2021 2021].

11. Public Health England. Public Health England, About Us https://www.gov.uk/government/organisations/public-health-england/about2021 [accessed 06 September 2021 2021].

12. Public Health England. Public Health England: Approach to Surveillance https://www.gov.uk/government/publications/public-health-england-approach-to-surveillance/public-health-england-approach-to-surveillance2017 [accessed 06 September 2021 2021].

13. Elliot AJ, Hughes HE, Hughes TC, et al. Establishing an emergency department syndromic surveillance system to support the London 2012 Olympic and Paralympic Games. *Emergency Medicine Journal* 2012;29(12):954-60.

14. Harcourt S, Fletcher J, Loveridge P, et al. Developing a new syndromic surveillance system for the London 2012 Olympic and Paralympic Games. *Epidemiology & Infection* 2012;140(12):2152-56.

15. Scales D. Opportunities and Challenges for Developing Syndromic Surveillance Systems for the Detection of Social Epidemics. *Online Journal of Public Health Informatics* 2020;12(1)

16. Paterson BJ, Durrheim DN. The remarkable adaptability of syndromic surveillance to meet public health needs. *Journal of epidemiology and global health* 2013;3(1):41-47.

17. Henning KJ. What is syndromic surveillance? *Morbidity and mortality weekly report* 2004:7-11.

18. Reingold A. If syndromic surveillance is the answer, what is the question? *Biosecurity and Bioterrorism: Biodefense strategy, practice, and science* 2003;1(2):77-81.

19. Fricker Jr RD. Syndromic surveillance. Encyclopedia for Quantitative Risk Analysis and Assessment: John Wiley & Sons Ltd 2008.

20. Kulessa M, Loza Mencía E, Fürnkranz J. A Unifying Framework and Comparative Evaluation of Statistical and Machine Learning Approaches to Non-Specific Syndromic Surveillance. *Computers* 2021;10(3):32.

21. Hughes HE, Edeghere O, O'Brien SJ, et al. Emergency department syndromic surveillance systems: a systematic review. *BMC public health* 2020;20(1):1-15.

22. Jones NF, Marshall R. Evaluation of an Electronic General-Practitioner—Based Syndromic Surveillance System—Auckland, New Zealand, 2000–2001. *Morbidity and Mortality Weekly Report* 2004:173-78.

23. Harcourt S, Morbey R, Loveridge P, et al. Developing and validating a new national remote health advice syndromic surveillance system in England. *Journal of Public Health* 2017;39(1):184-92.

24. Lawpoolsri S, Khamsiriwatchara A, Liulark W, et al. Real-time monitoring of school absenteeism to enhance disease surveillance: a pilot study of a mobile electronic reporting system. *JMIR mHealth and uHealth* 2014;2(2):e3114.

25. Rodriguez D, Zhang G, Leguen F, et al. Using public school absentee data to enhance syndromic surveillance in Miami-Dade County, 2007. *Advances in Disease Surveillance* 2007;4:188.

26. Besculides M, Heffernan R, Mostashari F, et al. Evaluation of school absenteeism data for early outbreak detection, New York City. *BMC public health* 2005;5(1):1-7.

27. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. AMIA annual symposium proceedings; 2006. American Medical Informatics Association.

28. Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance. *PloS one* 2009;4(2):e4378.

29. Samaras L, García-Barriocanal E, Sicilia M-A. Syndromic surveillance using web data: a systematic review. *Innovation in Health Informatics* 2020:39-77.

30. Edge VL, Pollari F, King L, et al. Syndromic surveillance of norovirus using over the counter sales of medications related to gastrointestinal illness. *Canadian Journal of Infectious Diseases and Medical Microbiology* 2006;17(4):235-41.

31. Edge VL, Pollari F, Lim G, et al. Syndromic surveillance of gastrointestinal illness using pharmacy over-the-counter sales. *Canadian Journal of Public Health* 2004;95(6):446-50.

32. Magruder SF, Lewis SH, Najmi A, et al. Progress in understanding and using over-the-counter pharmaceuticals for syndromic surveillance. *Morbidity and Mortality Weekly Report* 2004:117-22.

33. Sosin DM. Draft framework for evaluating syndromic surveillance systems. *Journal of urban health* 2003;80(1):i8-i13.

34. Baker M, Wilson N, Huang Q, et al. Pandemic influenza A (H1N1) v in New Zealand: the experience from April to August 2009. *Eurosurveillance* 2009;14(34):19319.

35. Olson DR, Konty KJ, Paladini M, et al. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS computational biology* 2013;9(10):e1003256.

36. Bawa Z, Elliot AJ, Morbey RA, et al. Assessing the likely impact of a rotavirus vaccination program in England: the contribution of syndromic surveillance. *Clinical Infectious Diseases* 2015;61(1):77-85.

37. Elliot A, Hughes H, Hughes T, et al. The impact of thunderstorm asthma on emergency department attendances across London during July 2013. *Emergency Medicine Journal* 2014;31(8):675-78.

38. Güemes A, Ray S, Aboumerhi K, et al. A syndromic surveillance tool to detect anomalous clusters of COVID-19 symptoms in the United States. *Scientific reports* 2021;11(1):1-11.

39. Lapointe-Shaw L, Rader B, Astley CM, et al. Syndromic surveillance for COVID-19 in Canada. *medrxiv* 2020

40. Maharaj AS, Parker J, Hopkins JP, et al. The effect of seasonal respiratory virus transmission on syndromic surveillance for COVID-19 in Ontario, Canada. *The Lancet Infectious Diseases* 2021;21(5):593-94.

41. Caserio-Schönemann C, Meynard J. Ten years experience of syndromic surveillance for civil and military public health, France, 2004-2014. *Eurosurveillance* 2015;20(19):21126.

42. Dupuy C, Bronner A, Watson E, et al. Inventory of veterinary syndromic surveillance initiatives in Europe (Triple-S project): Current situation and perspectives. *Preventive veterinary medicine* 2013;111(3-4):220-29.

43. O'Connell EK, Zhang G, Leguen F, et al. Innovative uses for syndromic surveillance. *Emerging infectious diseases* 2010;16(4):669.

44. Public Health England. National Ambulance Syndromic Surveillance: Weekly Bulletins 2021 https://www.gov.uk/government/publications/national-ambulance-syndromic-surveillance-weekly-bulletins-20212021 [accessed 06 September 2021 2021].

45. Morbey RA, Elliot AJ, Charlett A, et al. The application of a novel 'rising activity, multi-level mixed effects, indicator emphasis'(RAMMIE) method for syndromic surveillance in England. *Bioinformatics* 2015;31(22):3660-65.

46. Smith GE, Elliot AJ, Ibbotson S, et al. Novel public health risk assessment process developed to support syndromic surveillance for the 2012 Olympic and Paralympic Games. *Journal of Public Health* 2017;39(3):e111-e17.

47. Kamilaris A, Kartakoullis A, Prenafeta-Boldú FX. A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture* 2017;143:23-37.

48. Big data analytics architecture for real-time traffic control. 2017 5th IEEE international conference on models and technologies for intelligent transportation systems (MT-ITS); 2017. IEEE.

49. Drivas IC, Sakas DP, Giannakopoulos GA, et al. Big data analytics for search engine optimization. *Big Data and Cognitive Computing* 2020;4(2):5.

50. Thorson K, Cotter K, Medeiros M, et al. Algorithmic inference, political interest, and exposure to news and politics on Facebook. *Information, Communication & Society* 2021;24(2):183-200.

51. Hasan MM, Popp J, Oláh J. Current landscape and influence of big data on finance. *Journal of Big Data* 2020;7(1):1-17.

52. Lee H, Cho C-H. Digital advertising: present and future prospects. *International Journal of Advertising* 2020;39(3):332-41.

53. Goes F, Meerhoff L, Bueno M, et al. Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. *European journal of sport science* 2021;21(4):481-96.

54. Fussey P, Roth S. Digitizing sociology: Continuity and change in the internet era. *Sociology* 2020;54(4):659-74.

55. Woo SEE, Tay LE, Proctor RW. Big data in psychological research: American Psychological Association 2020.

56. Razzak MI, Imran M, Xu G. Big data analytics for preventive medicine. *Neural Computing and Applications* 2020;32(9):4417-51.

57. Diebold FX. On the Origin (s) and Development of the Term'Big Data'. 2012

58. Favaretto M, De Clercq E, Schneble CO, et al. What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade. *PloS one* 2020;15(2):e0228987.

59. Anuradha J. A brief introduction on Big Data 5Vs characteristics and Hadoop technology. *Procedia computer science* 2015;48:319-24.

60. Baro E, Degoul S, Beuscart R, et al. Toward a literature-driven definition of big data in healthcare. *BioMed research international* 2015;2015

61. De Mauro A, Greco M, Grimaldi M. A formal definition of Big Data based on its essential features. *Library Review* 2016

62. Wu X, Zhu X, Wu G-Q, et al. Data mining with big data. *IEEE transactions on knowledge and data engineering* 2013;26(1):97-107.

63. Noufaily A, Morbey RA, Colón-González FJ, et al. Comparison of statistical algorithms for daily syndromic surveillance aberration detection. *Bioinformatics* 2019;35(17):3110-18.

64. Hutwagner L, Thompson W, Seeman GM, et al. The bioterrorism preparedness and response early aberration reporting system (EARS). *Journal of Urban Health* 2003;80(1):i89-i96.

65. Lake IR, Colon-Gonzalez FJ, Barker GC, et al. Machine learning to refine decision making within a syndromic surveillance service. *BMC Public Health* 2019;19(1):1-12.

66. Bollig N, Clarke L, Elsmo E, et al. Machine learning for syndromic surveillance using veterinary necropsy reports. *PloS one* 2020;15(2):e0228105.

67. Jackson ML, Baer A, Painter I, et al. A simulation study comparing aberration detection algorithms for syndromic surveillance. *BMC medical informatics and decision making* 2007;7(1):1-11.

68. Reis BY, Pagano M, Mandl KD. Using temporal context to improve biosurveillance. *Proceedings of the National Academy of Sciences* 2003;100(4):1961-65.

69. Reis BY, Mandl KD. Syndromic surveillance: the effects of syndrome grouping on model accuracy and outbreak detection. *Annals of emergency medicine* 2004;44(3):235-41.

70. Jensen FV. An introduction to Bayesian networks: UCL press London 1996.

71. Sum-product networks: A new deep architecture. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops); 2011. IEEE.

72. Duivesteijn W, Feelders AJ, Knobbe A. Exceptional model mining. *Data Mining and Knowledge Discovery* 2016;30(1):47-98.

73. Zeng D, Cao Z, Neill DB. Artificial intelligence–enabled public health surveillance—from local detection to global epidemic monitoring and control. Artificial Intelligence in Medicine: Elsevier 2021:437-53.

74. Nobles M, Lall R, Mathes R, et al. Multidimensional semantic scan for pre-syndromic disease surveillance. *Online Journal of Public Health Informatics* 2019;11(1)

75. Goodman RA, Buehler JW, Mott JA, et al. Defining field epidemiology: Oxford University Press New York, 2019:3-20.

76. National Human Genome Research Institute CfDC. The Cost of Sequencing a Human Genome ttps://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost2020 [accessed 28/07/2021 2021].

77. Marx V. A dream of single-cell proteomics. *Nature Methods* 2019;16(9):809-12.

78. Zou J, Huss M, Abid A, et al. A primer on deep learning in genomics. *Nature genetics* 2019;51(1):12-18.

79. Wainberg M, Sinnott-Armstrong N, Mancuso N, et al. Opportunities and challenges for transcriptome-wide association studies. *Nature genetics* 2019;51(4):592-99.

80. Tost J. 10 years of Epigenomics: a journey with the epigenetic community through exciting times: Future Medicine, 2020.

81. Turnbaugh PJ, Ley RE, Hamady M, et al. The human microbiome project. *Nature* 2007;449(7164):804-10.

82. Subramanian I, Verma S, Kumar S, et al. Multi-omics data integration, interpretation, and its application. *Bioinformatics and biology insights* 2020;14:1177932219899051.

83. Martin J, Klapsa D, Wilton T, et al. Tracking SARS-CoV-2 in sewage: evidence of changes in virus variant predominance during COVID-19 pandemic. *Viruses* 2020;12(10):1144.

84. Wilton T, Bujaki E, Klapsa D, et al. Rapid increase of SARS-CoV-2 variant B. 1.1. 7 detected in sewage samples from England between October 2020 and January 2021. *Msystems* 2021;6(3):e00353-21.

85. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* 2018;25(10):1419-28.

86. Morris MA, Saboury B, Burkett B, et al. Reinventing radiology: big data and the future of medical imaging. *Journal of thoracic imaging* 2018;33(1):4-16.

87. Ali F, El-Sappagh S, Islam SR, et al. An intelligent healthcare monitoring framework using wearable sensors and social networking data. *Future Generation Computer Systems* 2021;114:23-43.

88. Hoang T, Coletti P, Melegaro A, et al. A systematic review of social contact surveys to inform transmission models of close-contact infections. *Epidemiology (Cambridge, Mass)* 2019;30(5):723.

89. Thomas MJ, Yoon PW, Collins JM, et al. Evaluation of syndromic surveillance systems in 6 US state and local health departments. *Journal of public health management and practice: JPHMP* 2018;24(3):235.

90. Sim JXY, Conceicao EP, Wee LE, et al. Utilizing the electronic health records to create a syndromic staff surveillance system during the COVID-19 outbreak. *American journal of infection control* 2021;49(6):685-89.

91. de Lusignan S, Jones N, Dorward J, et al. Oxford Royal College of General Practitioners Clinical Informatics Digital Hub: Rapid innovation to deliver extended COVID-19 surveillance and trial platforms. *JMIR Public Health and Surveillance* 2020

92. Koleck TA, Dreisbach C, Bourne PE, et al. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association* 2019;26(4):364-79.

93. Coppersmith G, Leary R, Crutchley P, et al. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights* 2018;10:1178222618792860.

94. Abbood A, Ullrich A, Busche R, et al. EventEpi—A natural language processing framework for event-based surveillance. *PLoS computational biology* 2020;16(11):e1008277.

95. Berry A. Syndromic surveillance and its utilisation for mass gatherings. *Epidemiology & Infection* 2019;147

96. Black SD. Inequalities in health: the Black report. 1982

97. Marmot M, Bell R. Fair society, healthy lives. *Public health* 2012;126:S4-S10.

98. Hillman S, Shantikumar S, Ridha A, et al. Socioeconomic status and HRT prescribing: a study of practice-level data in England. *British Journal of General Practice* 2020;70(700):e772-e77.

99. Mooney J, Yau R, Moiz H, et al. Associations between socioeconomic deprivation and pharmaceutical prescribing in primary care in England. *Postgraduate Medical Journal* 2020

100. Soyombo S, Stanbrook R, Aujla H, et al. Socioeconomic status and benzodiazepine and Z-drug prescribing: a cross-sectional study of practice-level data in England. *Family practice* 2020;37(2):194-99.

101. Application of meteorological big data. 2016 16th international Symposium on Communications and information technologies (ISCIT); 2016. IEEE.

102. Dominici F, Peng RD, Bell ML, et al. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *Jama* 2006;295(10):1127-34.

103. Jeremy W. Air pollution and brain health: an emerging issue. *Lancet* 2017;390:1345-422.

104. Smith GE, Elliot AJ, Lake I, et al. Syndromic surveillance: two decades experience of sustainable systems–its people not just data! *Epidemiology & Infection* 2019;147

105. Buehler JW, Sonricker A, Paladini M, et al. Syndromic surveillance practice in the United States: findings from a survey of state, territorial, and selected local health departments. *Advances in Disease Surveillance* 2008;6(3):1-20.

106. Șerban O, Thapen N, Maginnis B, et al. Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. *Information Processing & Management* 2019;56(3):1166-84.

107. Althouse BM, Scarpino SV, Meyers LA, et al. Enhancing disease surveillance with novel data streams: challenges and opportunities. *EPJ Data Science* 2015;4(1):1-8.

108. Ginsberg J, Mohebbi MH, Patel RS, et al. Detecting influenza epidemics using search engine query data. *Nature* 2009;457(7232):1012-14.

109. Lazer D, Kennedy R, King G, et al. The parable of Google Flu: traps in big data analysis. *Science* 2014;343(6176):1203-05.

110. Kandula S, Shaman J. Reappraising the utility of Google flu trends. *PLoS computational biology* 2019;15(8):e1007258.

111. Bernardo TM, Rajic A, Young I, et al. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *Journal of medical Internet research* 2013;15(7):e147.

112. You are what you tweet: Analyzing twitter for public health. Fifth international AAAI conference on weblogs and social media; 2011.

113. Chen L, Hossain KT, Butler P, et al. Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models. *Data mining and knowledge discovery* 2016;30(3):681-710.

114. Edo-Osagie O, Smith G, Lake I, et al. Twitter mining using semi-supervised classification for relevance filtering in syndromic surveillance. *PloS one* 2019;14(7):e0210689.

115. Velardi P, Stilo G, Tozzi AE, et al. Twitter mining for fine-grained syndromic surveillance. *Artificial intelligence in medicine* 2014;61(3):153-63.

116. Greenko J, Mostashari F, Fine A, et al. Clinical evaluation of the Emergency Medical Services (EMS) ambulance dispatch-based syndromic surveillance system, New York City. *Journal of Urban Health* 2003;80(1):i50-i56.

117. Harder K, Andersen P, Bæhr I, et al. Electronic real-time surveillance for influenza-like illness: experience from the 2009 influenza A (H1N1) pandemic in Denmark. *Eurosurveillance* 2011;16(3):19767.

118. Coory M, Kelly H, Tippett V. Assessment of ambulance dispatch data for surveillance of influenza-like illness in Melbourne, Australia. *Public Health* 2009;123(2):163-68.

119. Duijster JW, Doreleijers SD, Pilot E, et al. Utility of emergency call centre, dispatch and ambulance data for syndromic surveillance of infectious diseases: a scoping review. *European journal of public health* 2020;30(4):639-47.

120. Mostashari F, Fine A, Das D, et al. Use of ambulance dispatch data as an early warning system for communitywide influenzalike illness, New York City. *Journal of Urban Health* 2003;80(1):i43-i49.

121. Choi KR, Heilemann MV, Fauer A, et al. A second pandemic: mental health spillover from the novel coronavirus (COVID-19). *Journal of the American Psychiatric Nurses Association* 2020;26(4):340-43.

122. Leeb RT, Bitsko RH, Radhakrishnan L, et al. Mental health–related emergency department visits among children aged< 18 years during the COVID-19 pandemic—United States, January 1–October 17, 2020. *Morbidity and Mortality Weekly Report* 2020;69(45):1675.

123. Gruebner O, Sykora M, Lowe SR, et al. Mental health surveillance after the terrorist attacks in Paris. *The Lancet* 2016;387(10034):2195-96.

124. Larkin GL, Beautrais AL. Emergency departments are underutilized sites for suicide prevention: Hogrefe Publishing, 2010.

125. Royal College of Psychiatrists. Using real-time surveillance Information Page https://www.rcpsych.ac.uk/improving-care/nccmh/national-suicide-prevention-programme/using-real-time-surveillance2021 [accessed 29/07/21 2021].

126. Cheng Q, Li H, Silenzio V, et al. Suicide contagion: A systematic review of definitions and research utility. *PloS one* 2014;9(9):e108724.

127. Kuramoto-Crawford SJ, Spies EL, Davies-Cole J. Detecting suicide-related emergency department visits among adults using the District of Columbia syndromic surveillance system. *Public Health Reports* 2017;132(1_suppl):88S-94S.

128. Zwald ML, Holland KM, Annor F, et al. Monitoring suicide-related events using National Syndromic Surveillance Program data. *Online Journal of Public Health Informatics* 2019;11(1)

129. Whitlam G, Dinh M, Rodgers C, et al. Diagnosis-based emergency department alcohol harm surveillance: What can it tell us about acute alcohol harms at the population level? *Drug and alcohol review* 2016;35(6):693-701.

130. Vilain P, Larrieu S, Mougin-Damour K, et al. Emergency department syndromic surveillance to investigate the health impact and factors associated with alcohol intoxication in Reunion Island. *Emergency medicine journal* 2017;34(6):386-90.

131. Ising A, Proescholdbell S, Harmon KJ, et al. Use of syndromic surveillance data to monitor poisonings and drug overdoses in state and local public health agencies. *Injury prevention* 2016;22(Suppl 1):i43-i49.

132. Jones SA, Soto K, Grogan E, et al. Notes from the Field: Syndromic Surveillance used to monitor emergency department visits during a synthetic cannabinoid overdose outbreak—connecticut, August 2018. *Morbidity and Mortality Weekly Report* 2020;69(8):220.

133. Hope K, Durrheim DN, d'Espaignet ET, et al. Syndromic surveillance: is it a useful tool for local outbreak detection?: BMJ Publishing Group Ltd, 2006.

134. Hughes HE, Colón-González FJ, Fouillet A, et al. The influence of a major sporting event upon emergency department attendances; A retrospective cross-national European study. *PloS one* 2018;13(6):e0198665.

135. Vandentorren S, Paty A-C, Baffert E, et al. Syndromic surveillance during the Paris terrorist attacks. *The Lancet* 2016;387(10021):846-47.

136. Salazar MA, Law R, Winkler V. Health consequences of an armed conflict in Zamboanga, Philippines using a syndromic surveillance database. *International journal of environmental research and public health* 2018;15(12):2690.

137. Elliot AJ, Cooper DL, Loveridge P, et al. Real Time Syndromic Surveillance Response to UK Flooding Incident 2007.

138. Griffith MM, Yahata Y, Irie F, et al. Evaluation of an ad hoc paper-based syndromic surveillance system in Ibaraki evacuation centres following the 2011 Great East Japan Earthquake and Tsunami. *Western Pacific surveillance and response journal: WPSAR* 2018;9(4):21.

139. Elliot A, Singh N, Loveridge P, et al. Syndromic surveillance to assess the potential public health impact of the Icelandic volcanic ash plume across the United Kingdom, April 2010. *Eurosurveillance* 2010;15(23):19583.

140. Environmental causes of asthma. Seminars in respiratory and critical care medicine; 2018. Thieme Medical Publishers.

141. Bundle N, Verlander NQ, Morbey R, et al. Monitoring epidemiological trends in back to school asthma among preschool and school-aged children using real-time syndromic surveillance in England, 2012–2016. *J Epidemiol Community Health* 2019;73(9):825-31.

142. Davies J, Erbas B, Simunovic M, et al. Literature review on thunderstorm asthma and its implicaitons for public health advice. 2017

143. Hughes HE, Morbey R, Fouillet A, et al. Retrospective observational study of emergency department syndromic surveillance data during air pollution episodes across London and Paris in 2014. *BMJ open* 2018;8(4):e018732.

144. Green HK, Edeghere O, Elliot AJ, et al. Google search patterns monitoring the daily health impact of heatwaves in England: How do the findings compare to established syndromic surveillance systems from 2013 to 2017? *Environmental research* 2018;166:707-12.

145. Dirmyer VF. Using Real-Time Syndromic Surveillance to Analyze the Impact of a Cold Weather Event in New Mexico. *Journal of environmental and public health* 2018;2018

146. Hughes H, Morbey R, Hughes T, et al. Using an emergency department syndromic surveillance system to investigate the impact of extreme cold weather events. *Public Health* 2014;128(7):628-35.

147. Hale AC, Sánchez-Vizcaíno F, Rowlingson B, et al. A real-time spatio-temporal syndromic surveillance system with application to small companion animals. *Scientific reports* 2019;9(1):1-14.

148. Hiller KM, Stoneking L, Min A, et al. Syndromic surveillance for influenza in the emergency department–a systematic review. *PloS one* 2013;8(9):e73832.

149. Bourgeois FT, Olson KL, Brownstein JS, et al. Validation of syndromic surveillance for respiratory infections. *Annals of emergency medicine* 2006;47(3):265. e1.

150. World Health Organization. Global epidemiological surveillance standards for influenza. 2013

151. Papadomanolakis-Pakis N, Maier A, van Dijk A, et al. Development and assessment of a hospital admissions-based syndromic surveillance system for COVID-19 in Ontario, Canada: ACES Pandemic Tracker. *BMC Public Health* 2021;21(1):1-9.

152. Elliot AJ, Harcourt SE, Hughes HE, et al. The COVID-19 pandemic: a new challenge for syndromic surveillance. *Epidemiology & Infection* 2020;148

153. Bravata DM, Sundaram V, McDonald KM, et al. Evaluating detection and diagnostic decision support systems for bioterrorism response. *Emerging Infectious Diseases* 2004;10(1):100.

154. Group CW. Framework for evaluation Public Health Surveillance Systems for Early Detection of Outbreaks. *MMWR* 2004;53:1-11.

155. Tam CC, Rodrigues LC, Viviani L, 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.

156. Balter S, Weiss D, Hanson H, et al. Three years of emergency department gastrointestinal syndromic surveillance in New York City: what have we found. *Morbidity and Mortality Weekly Report* 2004;54(1):175-80.

157. Straetemans M, Altmann D, Eckmanns T, et al. Automatic outbreak detection algorithm versus electronic reporting system. *Emerging infectious diseases* 2008;14(10):1610.

158. Ziemann A, Rosenkötter N, Riesgo LG-C, et al. A concept for routine emergency-care data-based syndromic surveillance in Europe. *Epidemiology & Infection* 2014;142(11):2433-46.

159. Cooper DL, Verlander N, Smith G, et al. Can syndromic surveillance data detect local outbreaks of communicable disease? A model using a historical cryptosporidiosis outbreak. *Epidemiology & Infection* 2006;134(1):13-20.

160. Berger M, Shiau R, Weintraub JM. Review of syndromic surveillance: implications for waterborne disease detection. *Journal of Epidemiology & Community Health* 2006;60(6):543-50.

161. Hyllestad S, Amato E, Nygård K, et al. The effectiveness of syndromic surveillance for the early detection of waterborne outbreaks: a systematic review. *BMC Infectious Diseases* 2021;21(1):1-12.

162. Raubertas RF. An analysis of disease surveillance data that uses the geographic locations of the reporting units. *Statistics in Medicine* 1989;8(3):267-71.

163. Rogerson PA, Yamada I. Approaches to syndromic surveillance when data consist of small regional counts. *Morbidity and Mortality Weekly Report* 2004:79-85.

164. Noufaily A, Enki DG, Farrington P, et al. An improved algorithm for outbreak detection in multiple surveillance systems. *Statistics in medicine* 2013;32(7):1206-22.

165. Aghaali M, Kavousi A, Shahsavani A, et al. Performance of Bayesian outbreak detection algorithm in the syndromic surveillance of influenza-like illness in small region. *Transboundary and emerging diseases* 2020;67(5):2183-89.

166. Mathes RW, Lall R, Levin-Rector A, et al. Evaluating and implementing temporal, spatial, and spatio-temporal methods for outbreak detection in a local syndromic surveillance system. *PLoS One* 2017;12(9):e0184419.

167. Fricker Jr RD, Hegler BL, Dunfee DA. Comparing syndromic surveillance detection methods: EARS'versus a CUSUM-based methodology. *Statistics in Medicine* 2008;27(17):3407-29.

168. World Health Organization. Communicable disease alert and response for mass gatherings. Technical workshop; Geneva, Switzerland: April 2008 https://www.who.int/csr/resources/publications/WHO_HSE_EPR_2008_8c.pdf2008 [accessed 02/08/2021 2021].

169. Karami M, Doosti-Irani A, Ardalan A, et al. Public health threats in mass gatherings: A systematic review. *Disaster medicine and public health preparedness* 2019;13(5-6):1035-46.

170. Suzuki S, Ohyama T, Taniguchi K, et al. Web-based Japanese syndromic surveillance for FIFA World Cup 2002. *Target* 2003;100:10.

171. Mellou K, Potamiti-Komi M, Sideroglou T, et al. Detection and management of a norovirus gastroenteritis outbreak, Special Olympics World Summer Games, Greece, June 2011. *International Journal of Public Health and Epidemiology* 2012;1(2):1-6.

172. Kajita E, Luarca MZ, Wu H, et al. Harnessing syndromic surveillance emergency department data to monitor health impacts during the 2015 Special Olympics World Games. *Public Health Reports* 2017;132(1_suppl):99S-105S.

173. White P, Saketa S, Johnson E, et al. Mass gathering enhanced syndromic surveillance for the 8th Micronesian Games in 2014, Pohnpei State, Federated States of Micronesia. *Western Pacific surveillance and response journal: WPSAR* 2018;9(1):1.

174. Lami F, Asi W, Khistawi A, et al. Syndromic surveillance of communicable diseases in mobile clinics during the Arbaeenia Mass Gathering in Wassit Governorate, Iraq, in 2014: Cross-sectional study. *JMIR public health and surveillance* 2019;5(4):e10920.

175. Sokhna C, Goumballa N, Van Thuan Hoang BMM, et al. Senegal's grand Magal of Touba: syndromic surveillance during the 2016 mass gathering. *The American journal of tropical medicine and hygiene* 2020;102(2):476.

176. Carrico R, Goss L. Syndromic surveillance: hospital emergency department participation during the Kentucky Derby Festival. *Disaster Management & Response* 2005;3(3):73-79.

177. Hoy D, Saketa ST, Maraka RR, et al. Enhanced syndromic surveillance for mass gatherings in the Pacific: a case study of the 11th Festival of Pacific Arts in Solomon Islands, 2012. *Western Pacific surveillance and response journal: WPSAR* 2016;7(3):15.

178. Elliot A, Morbey R, Hughes H, et al. Syndromic surveillance-a public health legacy of the London 2012 Olympic and Paralympic Games. *Public Health* 2013;127(8):777-81.

179. Kraemer MU, Golding N, Bisanzio D, et al. Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings. *Scientific reports* 2019;9(1):1-11.

180. Kraemer MU, Yang C-H, Gutierrez B, et al. The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science* 2020;368(6490):493-97.

181. Wesolowski A, Eagle N, Tatem AJ, et al. Quantifying the impact of human mobility on malaria. *Science* 2012;338(6104):267-70.

182. Alsing J, Usher N, Crowley PJ. Containing Covid-19 outbreaks with spatially targeted short-term lockdowns and mass-testing. *medRxiv* 2020

183. Kanjo E, Younis EM, Ang CS. Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection. *Information Fusion* 2019;49:46-56.

184. Integrating Social Media with Ontologies for Real-Time Crowd Monitoring and Decision Support in Mass Gatherings. PACIS; 2013.

185. Ngo MQ, Haghighi PD, Burstein F. A crowd monitoring framework using emotion analysis of social media for emergency management in mass gatherings. *arXiv preprint arXiv:160600751* 2016

186. Yom-Tov E, Borsa D, Cox IJ, et al. Detecting disease outbreaks in mass gatherings using Internet data. *Journal of medical Internet research* 2014;16(6):e3156.

187. Resisting Surveillance: Responding to Wearable Device Privacy Policies. Proceedings of the 38th ACM International Conference on Design of Communication; 2020.

188. Arias O, Wurm J, Hoang K, et al. Privacy and security in internet of things and wearable devices. *IEEE Transactions on Multi-Scale Computing Systems* 2015;1(2):99-109.

189. Stoto MA, Dempsey JX, Baer A, et al. Expert meeting on privacy, confidentiality, and other legal and ethical issues in syndromic surveillance. *Adv Dis Surveill* 2009;7(2):1-10.

190. Gilbert GL, Degeling C, Johnson J. Communicable disease surveillance ethics in the age of big data and new technology. *Asian bioethics review* 2019;11(2):173-87.

191. Macdonald H. Can the NHS successfully deliver its GP data extraction scheme?: British Medical Journal Publishing Group, 2021.

12. LIST OF TABLES

13. LIST OF FIGURES

14. APPENDIX 1: FULL LIST OF PUBLISHED WORK BY THE CANDIDATE

- Flowers N, Hartley L, Todkill D, et al. Co-enzyme Q10 supplementation for the primary prevention of cardiovascular disease. *Cochrane Database of Systematic Reviews* 2014(12) 12:CD010405. doi: 10.1002/14651858.CD010405.pub2.
- Freeman K, Geppert J, Stinton C, et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. bmj 2021;374:n1872
- Goulding J, Todkill D, Carr R, et al. Pustules, plaques and pot-bellied pigs: difficulties in diagnosing tinea faciei. *Clinical and experimental dermatology* 2010;35(3):e10-e11.

- Hartley L, Lee MS, Kwong JS, et al. Qigong for the primary prevention of cardiovascular disease. Cochrane Database of Systematic Reviews 2015(6) Jun 11;6:CD010390. doi: 10.1002/14651858.CD010390.pub2.
- Hillman S, Shantikumar S, Ridha A, et al. Socioeconomic status and HRT prescribing: a study of practice-level data in England. British Journal of General Practice 2020;70(700):e772-e77.
- Kidy F, Shehata M, Stanbrook R, et al. Potential excess spend in primary care due to NHS drug tariff variability in vitamin D preparations. JRSM open 2020;11(3):2054270419894850.
- Mooney J, Yau R, Moiz H, et al. Associations between socioeconomic deprivation and pharmaceutical prescribing in primary care in England. Postgraduate Medical Journal 2020; 138944.
- Mulchandani R, Brehmer C, Butt S, et al. Outbreak of Shiga toxin-producing Escherichia coli O157 linked with consumption of a fast-food product containing imported cucumbers, United Kingdom, August 2020. *International Journal of Infectious Diseases* 2021
- Pereboom MTR, Todkill D, Knapper E, et.al. Shiga toxin–producing Escherichia coli (STEC) O157 outbreak associated with likely transmission in an inflatable home paddling pool in England, June 2017. Perspectives in Public Health, 2018, Vol XX, No XI
- Reichel H, Stanbrook R, Johnson H, et al. Guidance impact on primary care prescribing rates of simple analgesia: an interrupted time series analysis in England. British Journal of General Practice 2021;71(704):e201-e08.
- Shickle D, Todkill D, Chisholm C, et al. Addressing inequalities in eye health with subsidies and increased fees for General Ophthalmic Services in socio-economically deprived communities: a sensitivity analysis. Public Health 2015;129(2):131-37.
- Soyombo S, Stanbrook R, Aujla H, et al. Socioeconomic status and benzodiazepine and Z-drug prescribing: a cross-sectional study of practice-level data in England. Family practice 2020;37(2):194-99.
- Taylor-Phillips S, Mistry H, Leslie R, et al. Extending the diabetic retinopathy screening interval beyond 1 year: systematic review. British Journal of Ophthalmology 2016;100(1):105-14.

- Todkill D, Elliot A, Morbey R, et al. What is the utility of using syndromic surveillance systems during large subnational infectious gastrointestinal disease outbreaks? An observational study using case studies from the past 5 years in England. Epidemiology & Infection 2016;144(11):2241-50.
- Todkill D, Fowler T, Hawker J, Estimating the Incubation Period of Q fever, a Systematic Review. Epidemiology and Infection. 2018 Apr;146(6):665-672. Epub 2018 Mar *21.*
- Todkill D, Gonzalez FdJC, Morbey R, et al. Environmental factors associated with general practitioner consultations for allergic rhinitis in London, England: a retrospective time series analysis. BMJ open 2020;10(12):e036724.
- Todkill D, Goulding JMR, Bowers P, Gee B Frederick E Mohs (1910-2002) Dermatology's debt to the father of the micrographic surgical technique. British Journal of Dermatology. July Supplement.
- Todkill D, Hughes HE, Elliot AJ, et al. An observational study using English syndromic surveillance data collected during the 2012 London Olympics–what did syndromic surveillance show and what can we learn for future mass-gathering events? Prehospital and disaster medicine 2016;31(6):628-34.
- Todkill D, Loveridge P, Elliot AJ, et al. Utility of ambulance data for real-time syndromic surveillance: a pilot in the West Midlands region, United Kingdom. Prehospital and disaster medicine 2017;32(6):667-72.
- Todkill D, Powell J. Participant Experiences of an internet based intervention and randomised control trial: interview study. BMC Public Health, BMC Public Health 2013, 13:1017
- Todkill D, Pudney R, Terrell A, et al. An outbreak of Shigella boydii serotype 20 in January 2015 amongst United Kingdom healthcare workers involved in the Ebola response in Sierra Leone. Journal of medical microbiology 2018;67(11):1596-600.
- Todkill D, Taibjee S, Borg A, et al. Flagellate erythema due to bleomycin. British journal of haematology 2008;142(6):857-57.

15. APPENDIX 2: SIGNATURES FROM CONTRIBUTORS

Embedded here are the signed statements from collaborators. Collaborators who remained working for their organisations and / or where the author was not immediately involved in the

COVID-19 pandemic were contacted. The senior author for all papers was included. This approach was agreed to be acceptable with the Post-Graduate Office.

Paper_1_Supporting Signatures.docx
Paper_2_Supporting Signatures.docx
Paper_3_Supporting Signatures.docx
Paper_4_Supporting Signatures.docx
Paper_5_Supporting Signatures.docx