



Usefulness of linked data for infectious disease events: a systematic review

cambridge.org/hygEmma Field^{1,2} , Melanie Strathearn³, Christopher Boyd-Skinner⁴
and Amalie Dyda³ 

Review

Cite this article: Field E, Strathearn M, Boyd-Skinner C, Dyda A (2023). Usefulness of linked data for infectious disease events: a systematic review. *Epidemiology and Infection* **151**, e46, 1–10. <https://doi.org/10.1017/S0950268823000316>

Received: 13 October 2022
Revised: 20 January 2023
Accepted: 10 February 2023

Keywords:

Epidemiology; infectious disease
epidemiology; infectious disease; outbreaks;
surveillance

Author for correspondence:

Amalie Dyda, E-mail: a.dyda@uq.edu.au

¹National Centre for Epidemiology and Population Health, Australian National University, Canberra, Australia; ²Menzies School of Health Research, Charles Darwin University, Darwin, Australia; ³School of Population Health, University of Queensland, Brisbane, Australia and ⁴Australian Commission on Safety and Quality in Health Care, Sydney, Australia

Abstract

Surveillance is a key public health function to enable early detection of infectious disease events and inform public health action. Data linkage may improve the depth of data for response to infectious disease events. This study aimed to describe the uses of linked data for infectious disease events. A systematic review was conducted using Pubmed, CINAHL and Web of Science. Studies were included if they used data linkage for an acute infectious disease event (e.g. outbreak of disease). We summarised the event, study aims and designs; data sets; linkage methods; outcomes reported; and benefits and limitations. Fifty-four studies were included. Uses of linkage for infectious disease events included assessment of severity of disease and risk factors; improved case finding and contact tracing; and vaccine uptake, safety and effectiveness. The ability to conduct larger scale population level studies was identified as a benefit, in particular for rarer exposures, risk factors or outcomes. Limitations included timeliness, data quality and inability to collect additional variables. This review demonstrated multiple uses of data linkage for infectious disease events. As infectious disease events occur without warning, there is a need to establish pre-approved protocols and the infrastructure for data-linkage to enhance information available during an event.

Introduction

Infectious disease events cause significant impact around the globe [1, 2]. Surveillance is a key public health function to enable early detection of infectious disease events and inform public health action [3]. In many settings, for example Australia, surveillance systems are fragmented with data reported from numerous sources and shared responsibility across varying levels of government [4]. Rapid changes in technology have presented opportunities for improved timeliness, interoperability, analysis and interpretation of surveillance data. One example of this is data linkage.

Data linkage is the process of linking two or more datasets to provide more comprehensive information on individuals. For example, hospitalisation data can be linked to notifiable disease data to provide information on patient outcomes [5]. Data linkage can be performed using deterministic and probabilistic linkage methods or a combination of both [6]. Deterministic linkage is where a unique identifier is used for linkage, or a statistical linkage key is used from a combination of variables such as name, date of birth and sex [6]. Probabilistic linkage allows more flexibility to accommodate errors in data and calculate the likelihood of a match based on weightings from variables such as name, date of birth and address [6]. For both methods a linkage key is used to identify each record in place of identifiable data, ensuring that all identifiers are omitted from the final dataset to minimise risks to confidentiality [7].

There are numerous examples of the use of data linkage for infectious diseases. Data linkage has been used for infectious diseases for determining effectiveness and safety of routine immunisations [8], improving Indigenous status completeness of notification data [9] and improving case ascertainment for notifiable conditions [10, 11]. However, these examples are often for improving routine activities rather than for informing the response to an acute infectious disease event. Such events require a range of data to be collected and analysed rapidly to inform the response. These data may include, but are not limited to, notification, laboratory, hospitalisation, vaccination and mortality data. Typically, these data are collected through different systems, resulting in public health responders having to collect and analyse them separately.

Data linkage infrastructure has been established in many jurisdictions, and in some cases the addition of infectious disease data to these linked data sets [12, 13]. This provides a unique opportunity to use linked data for both surveillance of and response to infectious disease events. We hypothesise that linkage of routinely collected data may improve the depth of

© The Author(s), 2023. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

data for response to infectious disease events without additional primary data collection. We conducted a systematic review to describe the uses of linked data for infectious disease events.

Methods

Objectives

The objective of this review was to describe ways in which linked data has been used to assist in the response for acute infectious disease events (i.e., outbreaks/epidemics or pandemics). More specifically, this systematic review describes: the data sets used for data linkage; the study designs used; the methodologies used to link the data sets; the outcomes reported on; and methodological issues and limitations.

Criteria for considering studies for this review

Types of intervention

A study conducted to illicit information about an infectious disease event using linkage of routinely collected data OR linkage of data collected for the purposes of the outbreak investigation with routinely collected data. We considered studies where electronic records were linked using a common unique identifier(s) and/or probabilistic or deterministic linkage.

Types of outcome measures: phenomena of interest

Acute infectious disease events (epidemic or pandemic) where a rapid public health response was required. The study may be conducted during or after the infectious disease event.

Electronic searches

Pubmed, CINAHL and Web of Science were used to search for studies. The electronic database searches were conducted on 2 November 2021. The search was limited to studies published in 2000 or later and to studies published in English. The search terms were as follows: PubMed ('data linkage' OR 'record linkage' OR 'linked records' OR 'linked data' OR 'linked database') AND (outbreak OR epidemic OR pandemic OR communicable disease (MeSH Terms) OR 'infectious disease'); Web of Science – TOPIC: (('data linkage' OR 'record linkage' OR 'linked records' OR 'linked data' OR 'linked database') AND (outbreak OR epidemic OR pandemic OR 'communicable disease' OR 'infectious disease')) and CINAHL – ('data linkage' OR 'record linkage' OR 'linked records' OR 'linked data' OR 'linked database') AND (outbreak OR epidemic OR pandemic OR 'communicable disease' OR 'infectious disease').

Screening

The titles and abstracts from the search were screened by EF and AD to determine if they should be included in the full text review. The full text of those articles which met the inclusion criteria was then reviewed by EF, AD, MS and CBS to determine if they met the criteria for final inclusion. The reference lists of included articles were reviewed to identify further studies for inclusion.

Data extraction and synthesis

Data were extracted using a standard data extraction form by EF and MS. Data fields included on the data extraction form were:

author, year, event, study objective, study design, data sources, method for data linkage, data linkage category (study used: (1) pre-established linked dataset only, (2) pre-established linked dataset plus linkage to another dataset or (3) data linked for the purpose of study only) outcomes and limitations specifically in regards to data linkage.

Results

A total of 6006 studies were identified from Pubmed ($n = 5784$), Web of Science ($n = 150$) and CINAHL ($n = 72$) (Fig. 1). Additionally, 12 studies were identified through contacting state and territory health departments. A total of 376 duplicates were removed. The remaining 5642 articles were screened in title and abstract review, through which 5590 were excluded. There were 54 studies for which the full text was reviewed. Twenty of these studies were excluded for the following reasons: insufficient description of the data linkage process and datasets linked [14–20]; the infectious disease event was identified as a result of the linkage rather than being initiated by the event [21]; primary data collected specifically for the event were linked rather than routinely collected data [22, 23]; a perspective paper [24], an editorial [25]; outcomes not related to an infectious disease event [26, 27]; data not linked at an individual level [28, 29]; a description of a linked dataset [30, 31]; and study protocol only [32, 33]. The editorial referred to a study which was reviewed and included [34]. Fourteen additional studies were identified through reviewing the reference lists of included articles [35–48] plus an additional five from the OPENSafely website [49–53] (Table 1).

Infectious disease event

The majority of the studies were based on the COVID-19 pandemic ($n = 35$, 64.8%) [35–69] and to a lesser extent the influenza A(H1N1) 2009 pandemic ($n = 12$, 22.2%) [70–81]. Two studies (3.7%) involved cases of *Mycobacterium chimaera* associated with exposure to contaminated heater-cooler units used during open cardiac surgery in the United Kingdom and Queensland, Australia [82, 83]. One study each was identified investigating an Ebola virus disease outbreak in Guinea [84], an anthrax outbreak among injecting drug users in Scotland [85], a case of tuberculosis in a health care worker in the United States [86], a pertussis outbreak in Western Australia [87] and an outbreak of severe acute respiratory syndrome (SARS) in Toronto, Canada [88].

Study aims

The uses of linkage of routinely collected data for infectious disease events identified from these studies were in these broad categories: assessment of severity of disease and risk factors for specific populations (e.g. those with specific diseases (tuberculosis/HIV), rare diseases, pregnant women, infants, children); improve case finding/contact tracing investigations; determine uptake, safety and effectiveness of a vaccine during an outbreak/pandemic; and evaluate sensitivity and completeness of a surveillance system (e.g. for a notifiable disease or adverse events following vaccination).

The most common category of study aims was to assess the severity of outcomes and/or risk factors associated with infection and/or severe outcomes in the general population or specific population groups ($n = 33$, 61.1%), such as infants, pregnant women, children, people with rare autoimmune diseases or aged

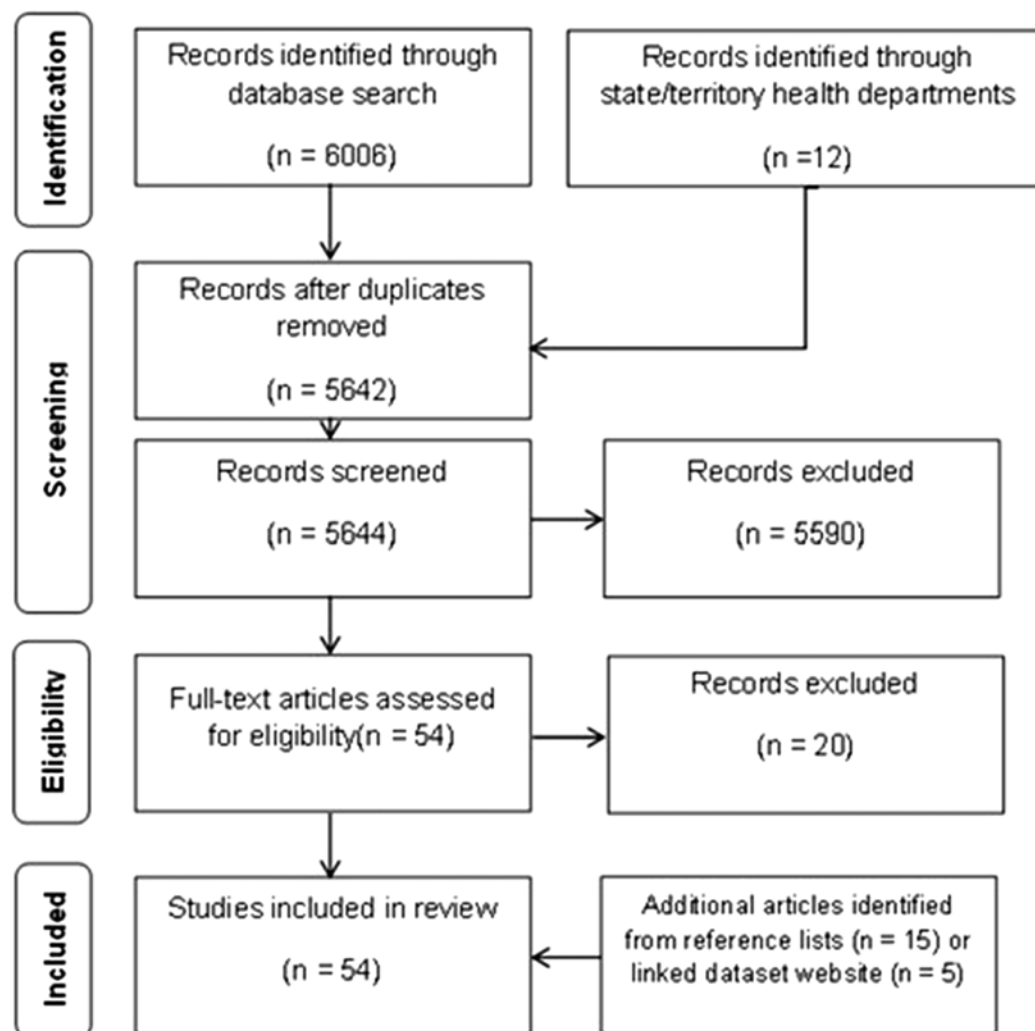


Fig. 1. PRISMA flow diagram [91]. This figure shows the number of studies included and excluded at each stage of the review process.

care residents [35, 55, 59, 60, 69–71, 74, 76, 85]. The second most common category of aims ($n = 14$, 25.9%) were associated with the safety, uptake and effectiveness of vaccines for either pandemic influenza A(H1N1) 2009 either in the general population, infants or in pregnant women [17, 72, 73, 75, 77, 78, 80, 81, 89] or a COVID-19 vaccination [39, 42, 46, 62, 65]. One of these studies specifically assessed the risk of a rare adverse event following vaccination, Guillian-Barre syndrome, in addition to other adverse events [80]. One study assessed the completeness of the adverse events reporting system in Taiwan [75]. Additionally, one study aimed to determine the effectiveness of preventing pertussis infection in infants through vaccinating new parents during a pertussis outbreak [8]. Two studies assessed the potential benefits of routinely prescribed pharmaceutical products on COVID-19 severity [50, 51] the first assessed the effect of hydroxychloroquine routinely prescribed for rheumatological disease on COVID-19 mortality; and the second assessed the association between routinely prescribed inhaled corticosteroids and COVID-19 related death in people with chronic obstructive pulmonary disease or asthma.

Two studies aimed to identify cases of *M. chimaera* associated with exposure to contaminated heater-cooler units used during open cardiac surgery in the United Kingdom and Queensland, Australia and one study aimed to identify contacts of a TB case

[82, 83, 86]. One study aimed to evaluate the sensitivity of two passive surveillance systems for Ebola [84]. One study assessed the performance of a medical decision algorithm to mitigate spread of SARS from inter-facility patient transfers in Toronto, Canada [88].

Study design

The cohort study design was most common ($n = 38$, 70.4%) [8, 35–53, 55, 57–59, 61–66, 68–71, 74, 77, 81, 82]. Five studies were descriptive analyses [54, 56, 60, 78, 80], three were case-control studies [72, 85, 89] and two studies used capture-recapture analysis [75, 76]. One study was a sensitivity calculation for a surveillance system [84]. Another study was a population-based self-controlled case series [73], one was a review of linked records [88], one was a retrospective case detection [83], one was a contact investigation [86] and one was a point prevalence study [67].

Data sources

Routinely collected data sources included births, deaths, drugs misuse, notifiable diseases, hospitalisations, primary care, laboratory,

Table 1. Summary of studies using data linkage for an acute infectious disease event

Author, year	Event, location	Objective category	Study design	Linkage category
MacDonald, 2006 [88]	Severe acute respiratory syndrome outbreak, Canada.	Evaluation	Review of linked records	Probabilistic
Huang, 2010 [80]	Influenza A(H1N1)pdm09 pandemic, Taiwan.	Vaccine effectiveness	Descriptive analysis	Unique identifier
Simpson, 2010 [81]	Influenza A(H1N1)pdm09 pandemic, Scotland.	Vaccine effectiveness	Cohort study	Unique identifier
Huang, 2011 [78]	Influenza A(H1N1)pdm09 pandemic, Taiwan.	Vaccine safety	Descriptive analysis	Unique identifier
Mahmud, 2011 [89]	Influenza A(H1N1)pdm09 pandemic, Canada.	Vaccine effectiveness	Case-control study	Unique identifier
Huang, 2012 [75]	Influenza A(H1N1)pdm09 pandemic, Taiwan.	Vaccine safety	Capture-recapture analysis	Unique identifier
Jules, 2012 [76]	Influenza A(H1N1)pdm09 pandemic, United States	Severity	Capture-recapture analysis	Linkage of multiple variables - not further described
Palmateer, 2012 [85]	Outbreak of anthrax infection among heroin users, Scotland.	Risk factors	Case control study	Probabilistic
Simpson, 2012 [77]	Influenza A(H1N1)pdm09 pandemic, Scotland.	Vaccine effectiveness and uptake	Cohort study with embedded case-control study	Unique identifier
Doyle, 2013 [74]	Influenza A(H1N1)pdm09 pandemic, United States.	Severity	Cohort study	Unique identifier plus other variables
Huang, 2013 [73]	Influenza A(H1N1)pdm09 pandemic, Taiwan.	Vaccine safety	Self-controlled case series	Unique identifier
Carcione, 2015 [8]	Pertussis epidemic, Western Australian, Australia.	Vaccine effectiveness	Cohort study	Probabilistic
Mahmud, 2015 [72]	Influenza A(H1N1)pdm09 pandemic, Canada.	Vaccine effectiveness	Nested case control study	Unique identifier
Sanderson, 2015 [86]	Contact investigation following diagnosis of a health care worker with infectious tuberculosis, United States.	Case finding	Contact investigation	Links of on multiple variables - not further described
Smith, 2015 [71]	Influenza A(H1N1)pdm09 pandemic, United Kingdom.	Severity	Cohort study	Links of on multiple variables - not further described
Lee, 2016 [84]	Ebola virus disease outbreak, Guinea.	Surveillance system evaluation	Sensitivity calculation	Probabilistic
Chand, 2017 [82]	<i>Mycobacterium chimera</i> associated with exposure to contaminated heater-cooler unit during cardiac surgery, United Kingdom.	Case finding	Cohort study	Unique identifier
Lee, 2018 [70]	Influenza A(H1N1)pdm09 pandemic, United Kingdom	Risk factors, severity	Cohort study	Deterministic
Robertson, 2018 [83]	<i>Mycobacterium chimaera</i> associated with exposure to contaminated heater-cooler unit during cardiac surgery, Queensland, Australia	Case finding	Case detection	Deterministic and probabilistic
Ayoubkhani, 2020 [36]	COVID-19 pandemic, England and Wales	Risk factors, severity	Cohort study	Unique identifier
Bhaskaran, 2020 [49]	COVID-19 pandemic, England	Risk factors	Cohort study	Unique identifier
Boulle 2020 [55]	COVID-19 pandemic, South Africa	Risk factors, severity	Cohort study	Unique identifier
Branden, 2020 [37]	COVID-19 pandemic, Stockholm, Sweden	Risk factors, severity	Cohort study	Unique identifier
Clift, 2020 [38]	COVID-19 pandemic, England	Risk factors, severity	Cohort study	Unique identifier
Drefahl, 2020 [47]	COVID-19 pandemic, Sweden	Risk factors	Cohort study	Unique identifier
Gobbato, 2020 [57]	COVID-19 pandemic, Northern Italy	Risk factors, severity	Cohort study	Unique identifier
Hollinghurst, 2020 [59]	COVID-19 pandemic, Wales	Severity	Cross-sectional and cohort study	Unique identifier

Liu, 2020 [60]	COVID-19 pandemic, New South Wales, Australia	Severity	Descriptive analysis	Probabilistic
Peach, 2020 [69]	COVID-19 pandemic, England	Risk factors, severity	Cohort study	Unique identifier
Reilev, 2020 [48]	COVID-19 pandemic, Denmark	Risk factors, severity	Cohort study	Unique identifier
Rentsch, 2020 [50]	COVID-19 pandemic, England	Prevention	Cohort study	Unique identifier
Schultze, 2020 [51]	COVID-19 pandemic, England	Protective factor	Cohort studies	Unique identifier
Shah, 2020 [45]	COVID-19 pandemic, Scotland	Risk factors, severity	Cohort study	Unique identifier
Williamson, 2020 [35]	COVID-19 pandemic, England	Risk factors, severity	Cohort study	Unique identifier
Wong, 2020 [52]	COVID-19 pandemic, England	Risk factors, severity	Cohort studies	Unique identifier
Bhattacharya, 2021 [54]	COVID-19 pandemic, England	Risk factors, severity	Descriptive analysis	Unique identifier
Burton, 2021 [56]	COVID-19 pandemic, Scotland	Risk factors	Descriptive analysis	Unique identifier
Curtis, 2021 [39]	COVID-19 pandemic, England	Vaccine uptake	Cohort study	Unique identifier
Forbes, 2021 [40]	COVID-19 pandemic, England	Risk factors	Cohort study	Unique identifier
Gaughan, 2021 [41]	COVID-19 pandemic, England and Wales	Risk factors	Cohort study	Unique identifier
Grint, 2021 [53]	COVID-19 pandemic, England	Severity	Cohort study	Unique identifier
Haas, 2021 [42]	COVID-19 pandemic, Israel	Vaccine effectiveness	Cohort study	Unique identifier
Hall, 2021 [58]	COVID-19 pandemic, England	Risk factors, severity	Cohort study	Unique identifier plus other variables
Liu, 2021 [61]	COVID-19 pandemic in New South Wales, Australia	Severity, risk factors	Cohort study.	Probabilistic
Mathur, 2021 [43]	COVID-19 in England	Risk factors, severity	Cohort study	Unique identifier
Nafilyan, 2021a [63]	COVID-19 in England	Risk factors	Cohort study	Deterministic and probabilistic and unique identifier
Nafilyan, 2021b [64]	COVID-19 in England	Risk factors	Cohort study	Unique identifier
Nafilyan, 2021c [62]	COVID-19 in England	Vaccine uptake	Cohort study	Deterministic and probabilistic and unique identifier
Nafilyan, 2021d [44]	COVID-19 in England	Risk factors, severity	Cohort study	Deterministic and probabilistic and unique identifier
Nunes, 2021 [65]	COVID-19 in Portugal	Vaccine effectiveness	Cohort study	Deterministic
Taji, 2021 [66]	COVID-19 pandemic, Canada	Risk factors, severity	Cohort study	Unique identifier plus other variable
Vasileiou, 2021 [46]	COVID-19 pandemic, Scotland	Vaccine effectiveness	Cohort study	Unique identifier
Walker, 2021 [67]	COVID-19 pandemic, United Kingdom	Risk factors, severity	Point prevalence study	Deterministic
Welsh, 2021 [67]	COVID-19 pandemic, Australia	Risk factors, severity	Cohort study	Probabilistic

pharmacy, national call centre, HIV and AIDS reporting, surveillance systems, disease registers, obstetrics, adverse drug reaction reporting, demographic databases, vaccination, patient transfer data (Table 1).

Methods of linkage

For the majority of studies, data linkage occurred for the purpose of the study ($n = 30$). However, in the more recent studies it was common that a pre-established linked database was used ($n = 24$), of which eight were from the OpenSAFELY linked dataset [35, 39, 40, 43, 50–53].

The studies described methods to link datasets in varying levels of detail. The majority of the studies referred to using a unique identifier ($n = 37$) for the linkage [35–43, 45–59, 62, 66, 69, 73–75, 77–82]. Of these studies, three used one or more variables in addition to the unique identifier for the linkage [58, 66, 74]. Seven studies referred to using probabilistic linkage only [8, 60, 61, 68, 84, 85, 88]. Four studies cited using both deterministic and probabilistic linkage methods [44, 62, 63, 83].

Outcomes reported

The most commonly reported outcomes focused on mortality and morbidity from influenza A(H1N1) 2009 or COVID-19. The predominant outcome reported was mortality rate ($n = 27$) from either COVID-19 ($n = 25$) [35–38, 40–45, 47–53, 55, 57, 59, 61, 63–65, 68] or H1N1 ($n = 2$) [70, 71]. Other common outcomes reported (for COVID-19 and influenza A(H1N1) 2009) included hospital admission ($n = 13$) [38, 40, 42, 43, 45, 46, 48, 57, 60, 61, 65, 68, 76], ICU admission (or severe/critical status) ($n = 8$) [40, 42, 43, 45, 48, 60, 61, 68]. Six papers reported on diagnosis of COVID-19 [40, 42, 43, 53, 58, 66], two of which separated cases into symptomatic and asymptomatic [42, 58]. Two papers reported rates of ventilation from COVID-19 [60, 68], one reported rates of emergency department presentation from COVID-19 [68], one reported on COVID-19 outbreaks in care-homes [56] and one reported on community onset *vs.* hospital onset of COVID-19 infection [54]. One paper reported complications (such as onset of pneumonia) from influenza A(H1N1) 2009 infection [70] and one reported on maternal characteristic and neonatal outcomes and maternal admission to ICU (influenza A(H1N1) 2009) [74].

Outcomes related to influenza A(H1N1) 2009 vaccine uptake ($n = 3$) [77, 80, 81], effectiveness ($n = 4$) [72, 77, 81, 89] and adverse events ($n = 4$) [73, 75, 78, 80] were also commonly reported. Two papers reported uptake of COVID-19 vaccines [39, 62] and three reported effectiveness of COVID-19 vaccines [42, 46, 65].

Additional outcomes included risk of infection in infants from pertussis between vaccinated and unvaccinated parents [8], risk of infection from *Mycobacterium chimera* [82] and sensitivity of calls to the national call centre and to local alerts regarding Ebola [84].

Benefits and limitations

A commonly identified benefit of these studies was the ability to study health in population-based cohorts [37, 43, 55, 61, 63, 69, 74]. The accuracy of data was also highlighted as a benefit. In one example, a study reported the use of hospital and health

records to provide accurate data which is less prone to selection and recall bias [72].

The ability to conduct more in-depth or large-scale analysis, due to increased information available through linkage from multiple sources was also identified as a strength. A paper linking hospital and primary care data allowed for more detailed analyses to investigate risk factors for complications from influenza in children. The linkage of the two data sets allowed for analysis of these risk factors managed in primary care as well as the risk of hospitalisation [70]. Large scale population analyses were common in the use of data linkage to investigate COVID-19 [43, 63]. In one example COVID-19 hospitalisation rates for all of New South Wales, Australia, were investigated using notifiable disease data and hospital record data [60].

A high proportion of the studies included in this analysis did not report limitations directly related to data linkage methods or processes. However, poor quality data – characterised by incomplete data sets, missing records or unique identifiers that were discovered during the linkage process – accounted for the most significant limitation. Mismatching of unique identifiers from probabilistic linkage methods in one study [84] saw decreased efficacy in results (sensitivity and specificity of record matching was 75%). The quality of datasets used varied greatly, with some studies reporting a substantial proportion of missing data [74, 82, 84]. Importantly these three studies were the least recent in the included studies.

Another commonly reported limitation reported was the reliance of data variables available [47, 49, 51, 55–57]. As data linkage relies on data already collected, collecting additional information is not possible. For example, a study investigating the mortality among influenza A patients admitted to hospital cited that the lack of information about comorbidities or co-existing infections was a limitation. However, the authors noted that this could be addressed with linkage to other data sources [71].

Timeliness was a clear limitation identified in the included studies. Several of the studies identified in this review were published well after the event [82, 85, 88]. For example, one of the earlier studies by MacDonald *et al.* investigating a decision support tool to assist in the mitigation of the spread of SARS was conducted using data from 2003 but published in 2006 [88].

Discussion

This systematic review demonstrates that the linkage of routinely collected administrative datasets can be used for a variety of purposes for acute infectious disease events. Most of the studies identified in this review had been conducted in relation to the COVID-19 pandemic. We identified several key benefits of linkage of routinely collected data for infectious disease events, importantly the ability to conduct larger scale population-level studies with more detailed data. However, there are limitations to these methods for the use in responding to infectious disease events. These include timeliness, data quality and relying on data already available which does not allow for the collection of new or additional information that may be required for specific studies.

In relation to infectious disease events, data linkage can provide additional data for assessment of severity of disease and risk factors. This is particularly useful for rare diseases or events affecting specific populations such as pregnant women, infants and children. A study within the United Kingdom investigated associations between ethnicity and COVID-19 mortality, made

possible by the use of linked data [36]. For outbreak response, data linkage was shown to improve case finding in a number of studies. These methods could compliment traditional case finding methods, demonstrated by Sanderson *et al.* (2015) who used hospital records and immunisation records to enhance contact tracing for infectious tuberculosis, showing improved efficiency by better targeting the response [86]. Additionally, the use of data linkage has been shown to be useful to determine uptake, safety and effectiveness of vaccines during an outbreak/pandemic [39, 46, 73]. However, these types of studies generally need to use pre-established linked data to provide findings in a timely manner [39, 46].

The primary benefit of data linkage is that population level datasets can be used allowing for population-based studies, whereby rare outcomes, exposures and risk factors can be studied. For example, the risk of Guillain-Barre syndrome after administration of the influenza A(H1N1) pandemic vaccine [80], and quantifying the risk of death from COVID-19 in people with autoimmune rheumatic disease [69]. This method also allows for more detailed and accurate analysis, as these data are not able to be collected in a study without linkage and the collection of primary data can be both time and resource intensive.

There are several limitations to data linkage studies which need to be navigated, including data availability. These types of studies can only use the data variables that are already collected, yet other variables may be required to answer certain public health research questions. Linked datasets can be complemented with primary data collection in such instances. For example, one study identified in this review investigated whether antibodies against SARS-COV-2 are associated with a decreased symptomatic and asymptomatic reinfection [58]. Questionnaires on symptoms and exposures were required to complete this study, as these data were not routinely collected.

Data linkage studies are also limited by data quality. Most commonly, studies within this review reported issues due to under-reporting [54], missing data in the original data source [82] or limitations with the linkage methods used [88]. Existing unique identifiers across multiple datasets makes linkage easier. An example of this is in Taiwan where each resident is assigned a personal identification number, which allows for ease of linkage across multiple datasets such as medical records (inpatient and outpatient), vaccination data, birth registry, household registration [73, 75, 78, 80]. Within this review, data quality was less cited as a limitation over time, particularly in relation to completeness suggesting that as data quality and linkage infrastructure improves, data linkage studies will be of higher quality.

One clear limitation of the use of data linkage for infectious disease events is timeliness. Several of the studies identified in this review were published well after the infectious disease event, resulting in the findings of the study not immediately available for the public health response [72, 74]. The data needs for infectious disease events vary based on pathogen, context and clinical and public health response needs; vary over the duration of the infectious disease event; and in some circumstances cannot be anticipated [3]. However, in line with all other preparedness activities for infectious disease events, frameworks for data linkage outlining which data sources could be linked and for what purposes, as identified in this review, would be help address this.

As noted, some of these issues may improve over time with the introduction of greater data linkage infrastructure and better interoperability of clinical information systems. In the studies included within this review, data linkage predominately occurred

for the purpose of the study such as the linkage of numerous data sources including general practice data, hospitalisation data and serology data to evaluate vaccination reporting for the A(H1N1) 2009 pandemic [90]. However, this appeared to change over time with recent studies, particularly those investigating COVID-19, using pre-established linked databases [30, 55, 59].

Existing linked datasets with ongoing linkage can help with timeliness as researchers can utilise the pre-existing dataset, rather than going through the process of linkage themselves. A key example of this is the OPENSafely COVID-19 dataset, open-source electronic health records data from England which can be accessed for research and analysis purposes. A number of studies within this review utilised these data in a timely manner, highlighting the utility of such resources [35, 39, 40, 43, 50, 51, 53]. The COVID-19 pandemic has demonstrated a proof of concept that data linkage can be completed in a timelier manner. COVID-19 publications were conducted rapidly in response to the pandemic. This strengthens the case for continuing to improve infrastructure and interoperability to assist with data linkage studies for possible future pandemics and ongoing infectious disease events.

Some studies that would have been eligible for inclusion in this review may not have been identified as they may have used linked data but not stated this explicitly or used terms for data linkage not included in our search terms. Further, health authorities may use data linkage for acute public health response but not published the results of such analyses. This may mean the uses of data linkage may be underreported.

This review demonstrated that data linkage has been used to answer important public health questions that can inform action during infectious disease events. A critical barrier to the use of data linkage for informing action during an infectious disease event is the time taken to gain approval for linked data, access the data and perform the linkage. This review has identified common data sets and variables used for infectious disease events, as well as proactively developed data linkage infrastructure established specifically for infectious diseases events. As infectious disease events occur without warning, it is possible to establish pre-approved protocols for data-linkage to enhance information available on case/contact finding, severity of disease; risk factors for disease; and vaccine uptake, safety and effectiveness for use during an event.

Acknowledgements. We would like to acknowledge Ross Andrews support in the early conception of this project.

Financial support. Emma Field received salary support through the Australian Partnership for Preparedness Research on Infectious Disease Emergencies is a Centre of Research Excellence funded by the Australian Government National Health and Medical Research Council (NHMRC) NT 1116530.

Conflict of interest. None.

Data availability statement. The data described in this article are available on request from the authors.

References

1. Fan VY, Jamison DT and Summers LH (2018) Pandemic risk: how large are the expected losses? *Bull World Health Organization* **96**, 129–134.
2. Kirk MD *et al.* (2015) World health organization estimates of the global and regional disease burden of 22 foodborne bacterial, protozoal, and viral diseases, 2010: a data synthesis. *PLoS Medicine* **12**, e1001921.

3. **World Health Organization** (2017) Asia Pacific Strategy for Emerging Diseases and Public Health Emergencies (APSED III): Advancing implementation of the International Health Regulations (2005).
4. **Commonwealth of Australia** (2014) *National Framework for Communicable Disease Control*. Canberra: Australian Department of Health.
5. **Eisen DP et al.** (2020) Linking administrative data sets of inpatient infectious diseases diagnoses in far North Queensland: a cohort profile. *BMJ Open* **10**, e034845.
6. **European Commission**. "Methodology: Data Linkage." Retrieved 13 October, 2022, from https://ec.europa.eu/eurostat/cros/system/files/s-dwh-m_4.2_methodology_data_linkage_v2.pdf.
7. **Queensland Health** (2017) *Queensland Data Linkage Framework*. State of Queensland (Queensland Health)
8. **Carcione D et al.** (2015) The impact of parental postpartum pertussis vaccination on infection in infants: a population-based study of cocooning in Western Australia. *Vaccine* **33**, 5654–5661.
9. **Rowe SL and Cowie BC** (2016) Using data linkage to improve the completeness of Aboriginal and Torres Strait Islander status in communicable disease notifications in Victoria. *Australian and New Zealand Journal of Public Health* **40**, 148–153.
10. **Oeser C et al.** (2017) Using data linkage to improve surveillance methods for acute hepatitis E infections in England and Wales 2010–2016. *Epidemiology and Infection* **145**, 2886–2889.
11. **Lim FJ et al.** (2017) Using record linkage to validate notification and laboratory data for a more accurate assessment of notifiable infectious diseases. *BMC Medical Informatics and Decision Making* **17**, 86.
12. **Rowe SL et al.** (2019) Use of data linkage to improve communicable disease surveillance and control in Australia: existing practices, barriers and enablers. *Australian and New Zealand Journal of Public Health* **43**, 33–40.
13. **Jutte DP, Roos LL and Brownell MD** (2011) Administrative record linkage as a tool for public health research. *Annual Review of Public Health* **32**, 91–108.
14. **Jian SW et al.** (2020) Contact tracing with digital assistance in Taiwan's COVID-19 outbreak response. *International Journal of Infectious Diseases: IJID: Official Publication of the International Society for Infectious Diseases* **101**, 348–352.
15. **Siddiqui MK et al.** (2020) Characteristics and outcomes of health and social care workers testing positive for SARS-CoV-2 in the Tayside region of Scotland. *The European Respiratory Journal* **56**, 2002568.
16. **Rossetto EV and Luna EJD** (2015) Clinical aspects of influenza A(H1N1) pdm09 cases reported during the pandemic in Brazil, 2009–2010. *Einstein-Sao Paulo* **13**, 177–182.
17. **Moro ML et al.** (2013) A population based cohort study to assess the safety of pandemic influenza vaccine Focetria (R) in Emilia-Romagna region, Italy-part two. *Vaccine* **31**, 1438–1446.
18. **Tan EH et al.** (2021) COVID-19 in patients with autoimmune diseases: characteristics and outcomes in a multinational network of cohorts across three countries. *Rheumatology (Oxford)* **60**, Si37–Si50.
19. **Diaz RA et al.** (2021) Extracorporeal membrane oxygenation for COVID-19-associated severe acute respiratory distress syndrome in Chile: a nationwide incidence and cohort study. *American Journal of Respiratory & Critical Care Medicine* **204**, 34–43.
20. **Sandrini M et al.** (2020) Assessment of the overall mortality during the COVID-19 outbreak in the Provinces of Milan and Lodi (Lombardy Region, Northern Italy). *Epidemiologia e prevenzione* **44**(suppl. 2), 244–251.
21. **Brum L and Kupek E** (2005) Record linkage and capture-recapture estimates for underreporting of human leptospirosis in a Brazilian health district. *The Brazilian Journal of Infectious Diseases: An Official Publication of the Brazilian Society of Infectious Diseases* **9**, 515–520.
22. **Cêtre JC et al.** (2005) Outbreaks of contaminated broncho-alveolar lavage related to intrinsically defective bronchoscopes. *Journal of Hospital Infection* **61**, 39–45.
23. **Weiser AA et al.** (2016) FoodChain-lab: a trace-back and trace-forward tool developed and applied during food-borne disease outbreak investigations in Germany and Europe. *Plos One* **11**, e0151977.
24. **Duchen R et al.** (2021) The role of a resilient information infrastructure in COVID-19 vaccine uptake in Ontario. *Healthcare Quarterly* **24**, 7–11.
25. **Ishigami J** (2021) Risk factors for severe COVID-19 in a large medical records linkage system in the United States. *Mayo Clinical Proceedings* **96**, 2508–2510.
26. **Greff DRL et al.** (2021) Epidemiology and seasonality of human parainfluenza serotypes 1–3 in Australian children. *Influenza and Other Respiratory Viruses* **15**, 661–669.
27. **Grimm F et al.** (2020) Hospital admissions from care homes in England during the COVID-19 pandemic: a retrospective, cross-sectional analysis using linked administrative data. *International Journal of Population Data Science* **5**, 1663.
28. **Cai S, Yan D and Intrator O** (2021) COVID-19 cases and death in nursing homes: the role of racial and ethnic composition of facilities and their communities. *Journal of the American Medical Directors Association* **22**, 1345–1351.
29. **Di Girolamo C et al.** (2020) Socioeconomic inequalities in overall and COVID-19 mortality during the first outbreak peak in Emilia-Romagna Region (Northern Italy). *Epidemiologia e Prevenzione* **44**(suppl. 2), 288–296.
30. **Pottgård A et al.** (2020) Existing data sources in clinical epidemiology: the Danish COVID-19 cohort. *Clinical Epidemiology* **12**, 875–881.
31. **Northstone K et al.** (2021) The avon longitudinal study of parents and children – A resource for COVID-19 research: home-based antibody testing results, October 2020. *Wellcome Open Research* **6**, 34.
32. **Stock SJ et al.** (2020) COVID-19 in pregnancy in Scotland (COPS): protocol for an observational study using linked Scottish national data. *BMJ Open* **10**, e042813.
33. **Grimaud O et al.** (2021) TRANSCOV cohort protocol: an epidemiological study assessing the impact of critically ill COVID-19 patients long distance transfers between intensive care units. *BMJ Open* **11**, e054774.
34. **St Sauver JL et al.** (2021) Factors associated with severe COVID-19 infection among persons of different ages living in a defined midwestern US population. *Mayo Clinical Proceedings* **96**, 2528–2539.
35. **Williamson EJ et al.** (2020) Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430–436.
36. **Ayoubkhani D et al.** (2021) Ethnic-minority groups in England and Wales-factors associated with the size and timing of elevated COVID-19 mortality: a retrospective cohort study linking census and death records. *International Journal of Epidemiology* **49**, 1951–1962.
37. **Brandén M et al.** (2020) Residential context and COVID-19 mortality among adults aged 70 years and older in Stockholm: a population-based, observational study using individual-level data. *The Lancet Healthy Longevity* **1**, e80–e88.
38. **Clift AK et al.** (2020) Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. *BMJ* **371**, m3731.
39. **Curtis HJ et al.** (2021) Trends and clinical characteristics of COVID-19 vaccine recipients: a federated analysis of 57.9 million patients' primary care records *in situ* using OpenSAFELY. medRxiv. 2021.01.25.21250356.
40. **Forbes H et al.** (2021) Association between living with children and outcomes from COVID-19: openSAFELY cohort study of 12 million adults in England. *BMJ (Clinical Research Edition)* **372**, n628–n.
41. **Gaughan CH et al.** (2021) Religious affiliation and COVID-19-related mortality: a retrospective cohort study of prelockdown and postlockdown risks in England and Wales. *Journal of Epidemiology and Community Health* **75**, 509–514.
42. **Haas EJ et al.** (2021) Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *The Lancet* **397**, 1819–1829.
43. **Mathur R et al.** (2021) Ethnic differences in SARS-CoV-2 infection and COVID-19-related hospitalisation, intensive care unit admission, and death in 17 million adults in England: an observational cohort study using the OpenSAFELY platform. *The Lancet* **397**, 1711–1724.
44. **Nafilyan V et al.** (2021) Ethnic differences in COVID-19 mortality during the first two waves of the Coronavirus Pandemic: a nationwide cohort study of 29 million adults in England. *European Journal of Epidemiology* **36**, 605–617.

45. **Shah ASV et al.** (2020) Risk of hospital admission with coronavirus disease 2019 in healthcare workers and their households: nationwide linkage cohort study. *BMJ* **371**, m3582.
46. **Vasileiou E et al.** (2021) Interim findings from first-dose mass COVID-19 vaccination roll-out and COVID-19 hospital admissions in Scotland: a national prospective cohort study. *The Lancet* **397**, 1646–1657.
47. **Drefahl S et al.** (2020) A population-based cohort study of socio-demographic risk factors for COVID-19 deaths in Sweden. *Nature Communications* **11**, 5097.
48. **Reilev M et al.** (2020) Characteristics and predictors of hospitalization and death in the first 11 122 cases with a positive RT-PCR test for SARS-CoV-2 in Denmark: a nationwide cohort. *International Journal of Epidemiology* **49**, 1468–1481.
49. **Bhaskaran K et al.** (2021) HIV infection and COVID-19 death: a population-based cohort analysis of UK primary care data and linked national death registrations within the OpenSAFELY platform. *The Lancet HIV* **8**, e24–e32.
50. **Rentsch CT et al.** (2021) Effect of pre-exposure use of hydroxychloroquine on COVID-19 mortality: a population-based cohort study in patients with rheumatoid arthritis or systemic lupus erythematosus using the OpenSAFELY platform. *The Lancet Rheumatology* **3**, e19–e27.
51. **Schultze A et al.** (2020) Risk of COVID-19-related death among patients with chronic obstructive pulmonary disease or asthma prescribed inhaled corticosteroids: an observational cohort study using the OpenSAFELY platform. *The Lancet Respiratory Medicine* **8**, 1106–1120.
52. **Wong AY et al.** (2021) Use of non-steroidal anti-inflammatory drugs and risk of death from COVID-19: an OpenSAFELY cohort analysis based on two cohorts. *Annals of the Rheumatic Diseases* **80**, 943–951.
53. **Grint DJ et al.** (2021) Case fatality risk of the SARS-CoV-2 variant of concern B.1.1.7 in England, 16 November to 5 February. *Eurosurveillance* **26**, 2100256.
54. **Bhattacharya A et al.** (2021) Healthcare-associated COVID-19 in England: a national data linkage study. *Journal of Infection* **83**, 565–572.
55. **Boule A et al.** (2020) Risk factors for COVID-19 death in a population cohort study from the Western Cape Province, South Africa. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **73**, e2005–e2015.
56. **Burton JK et al.** (2021) Care-home outbreaks of COVID-19 in Scotland March to May 2020: national linked data cohort analysis. *Age & Ageing* **50**, 1482–1492.
57. **Gobbato M et al.** (2020) Clinical, demographical characteristics and hospitalisation of 3,010 patients with COVID-19 in Friuli Venezia Giulia Region (Northern Italy). A multivariate, population-based, statistical analysis. *Epidemiologia e prevenzione* **44**(suppl. 2), 226–234.
58. **Hall VJ et al.** (2021) SARS-CoV-2 infection rates of antibody-positive compared with antibody-negative health-care workers in England: a large, multicentre, prospective cohort study (SIREN). *Lancet (London, England)* **397**, 1459–1469.
59. **Hollinghurst J et al.** (2021) The impact of COVID-19 on adjusted mortality risk in care homes for older adults in Wales, UK: a retrospective population-based cohort study for mortality in 2016–2020. *Age & Ageing* **50**, 25–31.
60. **Liu B et al.** (2020) Hospital outcomes after a COVID-19 diagnosis from January to May 2020 in New South Wales Australia. *Communicable Diseases Intelligence* **44**, 1–10.
61. **Liu B et al.** (2021) High risk groups for severe COVID-19 in a whole of population cohort in Australia. *BMC Infectious Diseases* **21**, 685.
62. **Nafilyan V et al.** (2021) Sociodemographic inequality in COVID-19 vaccination coverage among elderly adults in England: a national linked data study. *BMJ Open* **11**, e053402.
63. **Nafilyan V et al.** (2021) An external validation of the QCovid risk prediction algorithm for risk of mortality from COVID-19 in adults: a national validation cohort study in England. *Lancet Digit Health* **3**, e425–e433.
64. **Nafilyan V et al.** (2021) Ethnicity, household composition and COVID-19 mortality: a national linked data study. *Journal of the Royal Society of Medicine* **114**, 182–211.
65. **Nunes B et al.** (2021) mRNA vaccine effectiveness against COVID-19-related hospitalisations and deaths in older adults: a cohort study based on data linkage of national health registries in Portugal, February to August 2021. *Euro Surveillance* **26**, 2100833.
66. **Taji L et al.** (2021) COVID-19 in patients undergoing long-term dialysis in Ontario. *Canadian Medical Association Journal* **26**, 2100833.
67. **Walker JL et al.** (2021) UK prevalence of underlying conditions which increase the risk of severe COVID-19 disease: a point prevalence study using electronic health records. *BMC Public Health* **21**, 484.
68. **Welsh J et al.** (2021) The ATHENA COVID-19 study: cohort profile and first findings for people diagnosed with COVID-19 in Queensland, 1 January to 31 December 2020. *Communicable Diseases Intelligence (2018)* **45**, 1–19.
69. **Peach E et al.** (2021) Risk of death among people with rare autoimmune diseases compared with the general population in England during the 2020 COVID-19 pandemic. *Rheumatology (Oxford)* **60**, 1902–1909.
70. **Lee JJ et al.** (2018) Risk factors for influenza-related complications in children during the 2009/10 pandemic: a UK primary care cohort study using linked routinely collected data. *Epidemiology and Infection* **146**, 817–823.
71. **Smith C et al.** (2015) Use of linked electronic health records to assess mortality and length of stay associated with pandemic influenza A(H1N1)pdm09 at a UK teaching hospital. *Epidemiology and Infection* **143**, 1125–1128.
72. **Mahmud SM et al.** (2015) Did the H1N1 vaccine reduce the risk of admission with influenza and pneumonia during the pandemic? *Plos One* **10**, e0142754.
73. **Huang WT et al.** (2013) Safety of pandemic (H1N1) 2009 monovalent vaccines in Taiwan: a self-controlled case series study. *Plos One* **8**, e58827.
74. **Doyle TJ, Goodin K and Hamilton JJ** (2013) Maternal and neonatal outcomes among pregnant women with 2009 pandemic influenza A(H1N1) illness in Florida, 2009–2010: a population-based cohort study. *Plos One* **8**, e79040.
75. **Huang WT et al.** (2012) The reporting completeness of a passive safety surveillance system for pandemic (H1N1) 2009 vaccines: a capture-recapture analysis. *Vaccine* **30**, 2168–2172.
76. **Jules A et al.** (2012) Estimating age-specific influenza-related hospitalization rates during the pandemic (H1N1) 2009 in Davidson Co, TN. *Influenza Other Respiratory Viruses* **6**, e63–e71.
77. **Simpson CR et al.** (2012) Effectiveness of H1N1 vaccine for the prevention of pandemic influenza in Scotland, UK: a retrospective observational cohort study. *Lancet Infectious Diseases* **12**, 696–702.
78. **Huang WT et al.** (2011) Adverse events following pandemic A (H1N1) 2009 monovalent vaccines in pregnant women – Taiwan, November 2009–August 2010. *Plos One* **6**, e23049.
79. **Mahmud I et al.** (2011) Maternal and perinatal factors associated with hospitalised infectious mononucleosis in children, adolescents and young adults: record linkage study. *BMC Infectious Diseases* **11**, 51.
80. **Huang W-T et al.** (2010) Design of a robust infrastructure to monitor the safety of the pandemic A(H1N1) 2009 vaccination program in Taiwan. *Vaccine* **28**, 7161–7166.
81. **Simpson CR et al.** (2010) Vaccine effectiveness in pandemic influenza – primary care reporting (VIPER): an observational study to assess the effectiveness of the pandemic influenza A (H1N1)v vaccine. *Health Technology Assessment* **14**, 313–346.
82. **Chand M et al.** (2017) Insidious risk of severe *Mycobacterium chimaera* infection in cardiac surgery patients. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **64**, 335–342.
83. **Robertson J et al.** (2018) Responding to *Mycobacterium chimaera* heater-cooler unit contamination: international and national intersectoral collaboration coordinated in the state of Queensland, Australia. *Journal of Hospital Infection* **100**, E77–E84.
84. **Lee CT et al.** (2016) Evaluation of a national call center and a local alerts system for detection of New cases of ebola virus disease – Guinea, 2014–2015. *MMWR Morbidity and Mortality Weekly Report* **65**, 227–230.
85. **Palmateer NE et al.** (2012) Anthrax infection among heroin users in Scotland during 2009–2010: a case-control study by linkage to a national drug treatment database. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **55**, 706–710.

86. **Sanderson JM *et al.*** (2015) Increasing the efficiency and yield of a tuberculosis contact investigation through electronic data systems matching. *Journal of the American Medical Informatics Association: JAMIA* **22**, 1089–1093.
87. **Regan AK *et al.*** (2016) Effectiveness of seasonal trivalent influenza vaccination against hospital-attended acute respiratory infections in pregnant women: a retrospective cohort study. *Vaccine* **34**, 3649–3656.
88. **MacDonald RD, Henry B and Stuart R** (2006) Performance analysis of a medical decision algorithm to mitigate spread of SARS due to interfacility patient transfers. *Prehospital Emergency Care* **10**, 383–389.
89. **Mahmud S *et al.*** (2011) Effectiveness of the pandemic H1N1 influenza vaccines against laboratory-confirmed H1N1 infections: population-based case-control study. *Vaccine* **29**, 7975–7981.
90. **Simpson CR *et al.*** (2015) Early estimation of pandemic influenza Antiviral and Vaccine Effectiveness (EAVE): use of a unique community and laboratory national data-linked cohort study. *Health Technology Assessment* **19**, 1–32.
91. **Moher D *et al.*** (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine* **6**, e1000097.