


BMJ Open Linking administrative data sets of inpatient infectious diseases diagnoses in far North Queensland: a cohort profile

Damon P Eisen,^{1,2} Emma S McBryde,³ Luke Vasanthakumar,¹ Matthew Murray,⁴ Miriam Harings,¹ Oyelola Adegboye ³

To cite: Eisen DP, McBryde ES, Vasanthakumar L, *et al*. Linking administrative data sets of inpatient infectious diseases diagnoses in far North Queensland: a cohort profile. *BMJ Open* 2020;**10**:e034845. doi:10.1136/bmjopen-2019-034845

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (<http://dx.doi.org/10.1136/bmjopen-2019-034845>).

Received 09 October 2019
Revised 31 January 2020
Accepted 24 February 2020



© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Infectious Diseases, Townsville Hospital, Townsville, Queensland, Australia

²College of Medicine and Dentistry, Division of Tropical Health and Medicine, James Cook University, Townsville, Queensland, Australia

³Australian Institute of Tropical Health and Medicine, Division of Tropical Health and Medicine, James Cook University, Townsville, Queensland, Australia

⁴Commonline Pty Ltd, Townsville, Queensland, Australia

Correspondence to

Dr Oyelola Adegboye;
oyeadegboye@yahoo.com

ABSTRACT

Purpose To design a linked hospital database using administrative and clinical information to describe associations that predict infectious diseases outcomes, including long-term mortality.

Participants A retrospective cohort of Townsville Hospital inpatients discharged with an International Classification of Diseases and Related Health Problems 10th Revision Australian Modification code for an infectious disease between 1 January 2006 and 31 December 2016 was assembled. This used linked anonymised data from: hospital administrative sources, diagnostic pathology, pharmacy dispensing, public health and the National Death Registry. A Created Study ID was used as the central identifier to provide associations between the cohort patients and the subsets of granular data which were processed into a relational database. A web-based interface was constructed to allow data extraction and evaluation to be performed using editable Structured Query Language.

Findings to date The database has linked information on 41 367 patients with 378 487 admissions and 1 869 239 diagnostic/procedure codes. Scripts used to create the database contents generated over 24 000 000 database rows from the supplied data. Nearly 15% of the cohort was identified as Aboriginal or Torres Strait Islanders. Invasive staphylococcal, pneumococcal and Group A streptococcal infections and influenza were common in this cohort. The most common comorbidities were smoking (43.95%), diabetes (24.73%), chronic renal disease (17.93%), cancer (16.45%) and chronic pulmonary disease (12.42%). Mortality over the 11-year period was 20%.

Future plans This complex relational database reutilising hospital information describes a cohort from a single tropical Australian hospital of inpatients with infectious diseases. In future analyses, we plan to explore analyses of risks, clinical outcomes, healthcare costs and antimicrobial side effects in site and organism specific infections.

INTRODUCTION

Deriving a broad and detailed understanding of the epidemiology of infectious diseases is crucial as they are a common cause of admissions to hospitals and frequent cause of hospital complications. In 2016–2017, 7.2 per

Strengths and limitations of this study

- The linked database will serve as a basis for future studies unique to tropical Australia of incidence, risk factors and clinical outcomes of patients with hospital admissions involving infectious diseases.
- The incorporation of pathology results in the cohort will allow precise characterisation of many infectious diseases.
- The patient cohort was based on data sets from a single hospital, findings might not be generalisable to the Australian population.
- The validity of cohort studies rely on the accuracy of clinical coding; therefore, some important clinical information may be underrepresented.

1000 of Australia's population were hospitalised with a primary diagnosis of an infectious disease.¹ The rate in Australia's Indigenous population was double this. Of the principal causes of hospitalisation, pneumonia was fourth, cellulitis ninth and 'other sepsis' 16th. Regrettably, 103 000 patient episodes (1.2% of all hospital separations) involved a hospital-acquired infection. Urinary tract infection, pneumonia and blood stream infection are the third to fifth most common hospital-acquired complications. These infections contribute to the marked increase in the average length of stay (17 vs 4.4 days)¹ and may increase mortality.² Patterns of mortality for various illnesses, chronic and acute, are documented by the Australian Institute of Health and Welfare. Infectious and parasitic diseases (narrowly defined) are relatively infrequent single causes of mortality (<3%).³ However, more commonly, they are contributors to multiple causes of death in patients with chronic conditions. For instance, pneumonia and influenza are particularly common causes of death in patients with dementia.

Currently, there exists an opportunity to reutilise large amounts of data collected for administrative and routine clinical purposes to derive a more detailed picture of the incidence of diseases in Australian hospitals.⁴ Data-linkage processes are a powerful tool for analysis of various disease cohorts. These are a value-adding re-use of previously acquired patient information that represents a rich research resource. We have developed a database that will be used in the future to analyse the incidence, risk factors and clinical outcomes of patients with hospital admissions involving infectious disease.

COHORT DESCRIPTION

Setting

The Townsville Hospital is the tertiary referral centre for North Queensland, providing specialist care for 670 000 people. Townsville is located at 19.26° S and has a 'dry tropics' climate with a mean rainfall of 1100 mm.

Cohort selection

A cohort of Townsville Hospital inpatients was identified based on International Classification of Diseases and Related Health Problems 10th Revision Australian Modification (ICD-10-AM) discharge codes for an infectious disease. The cohort spanned for the 11-year period from 1 January 2006 to 31 December 2016. Information from the episode of care that led to cohort inclusion and all previous and subsequent inpatient admissions was provided.

The ICD-10-AM codes primarily used to select the patient cohort were infectious and parasitic diseases (A00–B99) (online supplementary table S1). However, for completeness, selected infection-related codes were also included from:

- ▶ Diseases of the nervous system G* describing intracranial infection.
- ▶ Diseases of the eye, ear and mastoid process H* describing intraocular and ear infection.
- ▶ Diseases of the circulatory system I* describing cardiac infections.
- ▶ Diseases of the respiratory system J* describing upper and lower respiratory tract infections.
- ▶ Diseases of the digestive system K* describing intra-abdominal infections.
- ▶ Diseases of the skin and subcutaneous tissues L* describing skin and soft tissue infections.
- ▶ Diseases of the musculoskeletal system and connective tissue M* describing infections of the bony skeleton and muscles.
- ▶ Diseases of the genitourinary system N* describing urinary tract infections.
- ▶ Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified R* describing fever of unknown origin and shock among others.

Databases

The following key data relating to the selected cohort were provided with the approval of Queensland Government Data Custodians:

- ▶ Queensland Health Admitted Patient Data Collection (QHAPDC): patient demographics, Indigenous status, principal and other diagnoses ICD-10-AM codes, procedure codes using Australian Classification of Health Interventions, length of stay and hospital separation.
- Admitted patient clinical coding is regulated by National Australian Coding Standards and QHAPDC data quality is managed via systematic internal audit, the State Government Queensland Audit Office and through periodic external audits.
- ▶ Date, primary and secondary causes of death over the 11-year study period.
- ▶ Emergency data collection: triage category, principal and other diagnoses.
- ▶ Pathology: results for; general microbiology, infective serology testing, infective PCR testing; haematology, full blood examination, coagulation; biochemistry results, urea and electrolytes, liver function tests, C-reactive protein.
- ▶ Antimicrobial dispensing: ipharmacy (central pharmacy dispensing) and Pyxis (ward dispensing); dose, date and price of selected anti-infective drug dispensing.
- ▶ Notifiable Conditions System: type and site of infection.

Data linkage

Extracted patient information was identifiable by the Medical Records Number. This was used by the Health Statistics Branch of Queensland Health to perform data-linkage processes described in the Queensland Data Linkage Framework. Anonymised data, identified by a unique Created Study ID, were provided to the research team.

Database construction

The data were supplied variously as comma or tab delimited text or as spreadsheet documents, and was processed into a relational database. The Created Study ID (PU_ID) was used as the central identifier to provide associations between the cohort patients and the subsets of granular data.

A web-based interface was constructed to allow data extraction and evaluation to be performed using either editable Structured Query Language or a selection of preset queries. The script and analysis interface were written in PHP/MySQL using a text editor.

Data analysis

Patient data extracts for analysis were imported into SAS V.9.4. Descriptive summaries are presented as frequencies and percentages for categorical variables, and means, quartiles and SDs for continuous variables. Charlson Comorbidity Index (CCI)^{5 6} was used to rank patient illness severity based on the number and importance of comorbid diseases (online supplementary table S2).

Patient and public involvement statement

Patients or members of the public were not involved in the development and design of the research. The anonymised data extraction does not require patient recruitment.

RESULTS

Cohort profile and database characteristics

The database consisted of linked information from 41 367 patients with 378 487 admissions and 1 869 239 diagnostic or procedure codes. The ICD-10-AM codes for infectious diseases that were used to select patients for inclusion in the cohort are listed in online supplementary table S1. A summary of the data and the datafields is included in online supplementary table S2. The individual datafields are listed in online supplementary table S3. The ICD-10-AM codes used to identify comorbidities are listed in online supplementary table S4. A database structure was designed to best accommodate the contents of the supplied data and the available identifiers within it. Its relational structure is shown in [figure 1](#). The resulting relational structure was designed to provide total freedom to retrieve grouped patient information from all the component sources as a single data set.

The database contents were created using a variety of purpose-built scripts to process, reshape and clean the data. These scripts generated over 24 000 000 database rows from the supplied data. The Created Study ID (PU_ID) was used as the central identifier to provide associations between the cohort patients and the data subsets.

Some assumptions were made during the processing of data. If pathology results were entered during the same date and time range as an admission, then this was included as part of the admission even though no admission identifier was available in the pathology data set.

Much of the collected data was entered as free text and preset values were inconsistently provided across different entry systems, resulting in variations in the expression of the same values. Scripts were written to standardise these results, extracting quantifiable values where possible. For example, the birth date of each person was not reliably supplied and the maximum detail was extracted from various data sources. Some sources using the same PU_ID recorded the age inconsistently at a certain admission date, others had birth month and day, and others incorporated full birth dates. The scripts analysed and prioritised each of these and consolidated all available information for each of 41 367 people. The year of birth was successfully generated for every person. Additionally, the ICD-10-AM codes were not consistently entered. For example, 'A064' was entered but the correct format is 'A06.4'. Each was analysed, broken down into its components and entered into the database. For 1130 of the 8274 deaths, principal and other causes of death were listed as free text not ICD-10-AM codes. Causes of these deaths were coded manually.

Summary statistics are presented to give a basic description of the cohort ([table 1](#)). The distribution of age at first

admission was skewed towards older subjects. Similarly, the total number of admissions was markedly skewed towards higher values. This is due to the significant number of haemodialysis patients who had a median of six admissions with IQR of 2–41 over the 11-year duration of the cohort study. A large proportion of the patients identified as Indigenous (14.88%). Of interest, 4.5% of patients in this cohort were admitted to the Townsville Hospital from correctional facilities and Indigenous peoples are overrepresented among these patients compared with the cohort as a whole. The overall 11-year all-cause mortality was 20%. A high proportion of patients smoked (44%). Other major modifiable risk factors included alcohol abuse, obesity and malnutrition ([table 1](#)).

This patient cohort had a moderately low burden of comorbidity with an average CCI score of 1.86 (IQR, 0–3). About 16% had a CCI of 5 and above. The major comorbidities are diabetes, cancer and renal disease. Other common comorbidities were chronic pulmonary disease, cerebrovascular disease and myocardial infarction. Multiple comorbidities were present in 67% of patients ([table 2](#)).

The geographic location of patient domicile as determined by postcode at the time of inpatient registration and numbers of patients per 100 000 resident in the Local Government Area are shown in [figure 2](#). The majority of cohort patients resided in the Townsville Local Government Areas.

[Table 3](#) lists common infectious diseases diagnoses along with others of note in the tropical setting of Townsville Hospital. These diagnoses represent aggregated codes that describe infection due to the same pathogen or the same site. Multiple codes often describe infection of the same organ. For common conditions such as *Staphylococcus aureus* (A41), urinary tract infection (N39.0) and influenza and pneumonia (J09–J18), many diagnoses are coded as 'other'. Precise study of these conditions, other microbial or organ specific infectious disease will require disaggregation of codes and incorporation of the available pathology results.

DISCUSSION

This longitudinal cohort study describes patients discharged from the largest tertiary referral hospital in the tropical region of Australia with an infectious disease diagnosis. The infectious diseases included in this cohort represent an exhaustive list of conditions prevalent in Northern Australia as well as in Australian communities in general.

When we consider the patterns of infectious diseases found in this cohort, *S. aureus* was the most common pathogen identified followed by influenza and Group A streptococcus. Skin and soft tissue was the most common site of infection followed by the respiratory tract. Future analysis of patient factors associated with mortality is underway. These data will allow comparison with other

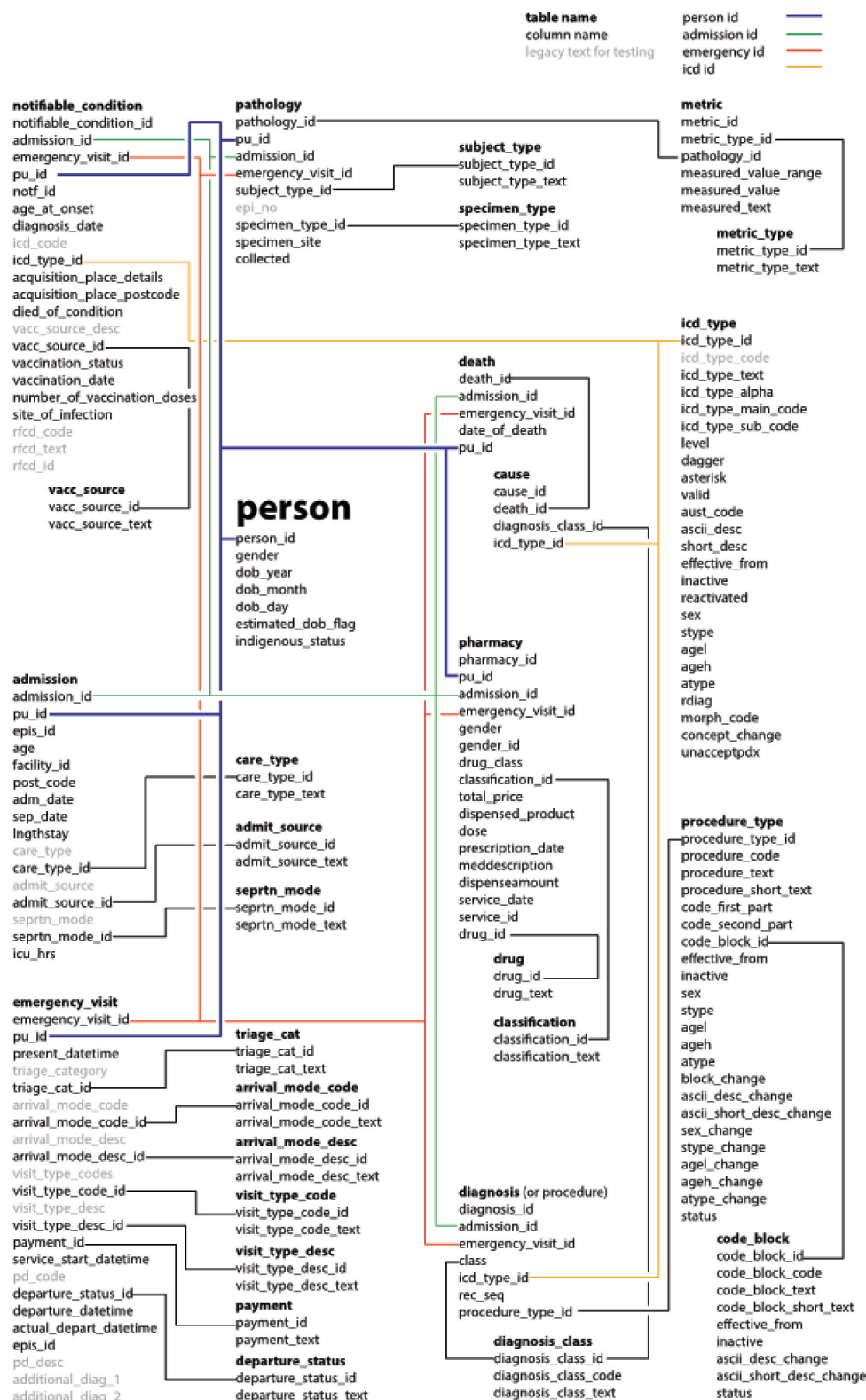


Figure 1 Representation of relational database constructed showing links between fields from incorporated administrative, clinical and death registry information. (see online supplementary table S1 for detailed description of fields).

mortality data from Australian studies of infectious diseases.

All-cause mortality rates from Australian cohorts of patients with selected, highly morbid, infections such as *S. aureus* bacteraemia (28%, 2–5 year follow-up),⁷

community-acquired pneumonia (60.4%, mean follow-up 6.1 years)⁸ and infective endocarditis (14.7%, 1–5 year follow-up) have been described.⁹ These studies all demonstrated increased all-cause mortality of the infectious diseases cohorts compared with controls.

Table 1 Cohort characteristics (n=41 367)

Characteristics	No	Mean	Median	SD	Q1	Q3
Age (years) at first admission	41 367	43.15	49	24.44	26	68
Total admissions	378 487	9.15	2	60.74	1	5
Demographics	Patients	Percentage				
Male	21 299	51.49				
Female	20 068	48.51				
11-year mortality						
Dead	8274	20				
Indigenous status						
Aboriginal but not TSI* origin	4763	11.51				
Aboriginal and TSI	550	1.33				
Neither Aboriginal nor TSI	35 187	85.06				
TSI but not Aboriginal	721	1.74				
Correctional facility	1853	4.47				
Aboriginal but not TSI origin	1450	78.25				
Aboriginal and TSI	8	0.43				
Neither Aboriginal nor TSI	402	21.69				
TSI but not Aboriginal	32	1.73				
Lifestyle						
Smoking	18 179	43.95				
Alcohol	4743	11.47				
Recreational drug use	600	1.45				
Malnutrition	4605	11.13				
Obesity	1633	3.95				

*Aboriginal and Torres Strait Islander.

This cohort will allow a wide range of future analyses on the epidemiology of severe infection in patients of the largest tertiary referral hospital in Northern Australia. Its size and complexity makes it a valuable resource. The variety of data that are incorporated allow for nuanced study of inpatients discharged with an infectious diagnosis. For example, linkage of microbiological, haematological and biochemical provides the opportunity to correlate numerous laboratory parameters with disease outcomes. Emergency department data will facilitate assessment of the numbers of hospital presentations made prior to a diagnosis such as cryptococcal meningitis. In a recent study based on a cohort of inpatients with pneumonia extracted from this data linkage, we found an immediate increase in risk of pneumonia associated with exposure to moderate low temperatures in late winter and early summer.¹⁰

There has been a sustained increase in the numbers of cohort studies using linked administrative hospital data sets, including in Australia.¹¹ However, infectious diseases studies are in the minority compared with cardiovascular, health services, cancer and maternal health research. Australian cohort studies that use data linkage to describe infectious diseases mostly rely on ICD-10-AM diagnostic

codes and death registry information. Some also incorporate notifiable diseases data¹² but, overall, studies incorporating pathology data are few.^{13 14}

Regrettably, in Australian jurisdictions, pathology data are only available for data linkage in Western Australia and Queensland due to their statewide diagnostic laboratories.⁴ Data-linkage studies incorporating pathology data have tested the precision of infectious diseases diagnosis in comparison with public health communicable diseases notifications systems¹⁵ and hospital discharge coding.¹³ These studies both demonstrated underascertainment of childhood respiratory tract diseases.

Australian infectious diseases cohort studies have involved: organ specific infections such as respiratory viral infections,¹³ infections such as Q fever¹² and *S. aureus* bacteraemia¹⁴ as well as specific patients such as asplenic¹⁶ and haematology-oncology.¹⁷ The value of Australian patient cohorts for infectious diseases research is further shown by the multiple studies deriving from the 45 and up study of ageing,¹⁸ Triple I Western Australian birth cohort¹⁵ and Victorian Post-Splenectomy Registry.¹⁹

There are inherent limitations of retrospective databases defined by ICD-10-AM codes. Some important clinical information is underrepresented. This is exemplified in

Table 2 Major comorbidities and Charlson Comorbidity Index

Major comorbidities	n	%
Myocardial infarction	3042	7.35
Peripheral vascular disease	1862	4.50
Cerebrovascular disease	3294	7.96
Heart failure	1754	4.24
Dementia	790	1.91
Chronic pulmonary disease	5140	12.42
Rheumatic disease	475	1.15
Peptic ulcer disease	568	1.37
Mild liver disease	1992	4.82
Moderate or severe liver disease	612	1.48
Diabetes without chronic complication	5131	12.40
Diabetes with chronic complication	5102	12.33
Hemiplegia or paraplegia	1907	4.61
Renal disease	7419	17.93
Any malignancy, including lymphoma and leukaemia, except malignant neoplasm of skin	5602	13.54
Metastatic solid tumour	1203	2.91
AIDS/HIV	114	0.26
Charlson Comorbidity Index (CCI)		
None: CCI score (0)	21 215	51.28
Mild: CCI score (1–2)	8270	19.99
Moderate: CCI score (3–4)	5492	15.28
Severe: CCI score (5+)	6390	15.45
Median (IQR)	0 (0–3)	
Mean (SD)	1.86 (2.72)	

this cohort study where only 3.95% of patients were coded as being obese. By contrast, among the general Australian population, as measured in 2017–2018, 31% of adults and 8.6% of children and adolescents were obese.²⁰ This inpatient underestimate may derive from ICD-10-AM coding for obesity only being allocated where active assessment is made by a dietitian for obesity. Inpatients at the Townsville Hospital were more frequently diagnosed (11.13%) with malnutrition reflecting documentation of clinical interventions. The administrative databases used to construct this linked database predated use of an electronic medical record at Townsville Hospital. Machine learning is being used in research settings to analyse free text in clinical notes and diagnostic imaging reports.²¹ However, owing to absence of free text data, we are unable to apply this methodology to our database. The absence of this clinical information may diminish the ability to determine precise case definitions and important comorbidities such as obesity.

Despite these potential limitations, ICD-10-AM codes for infectious diseases have been shown to be closely correlated with clinical diagnoses determined after medical chart review in Australian research, for example,

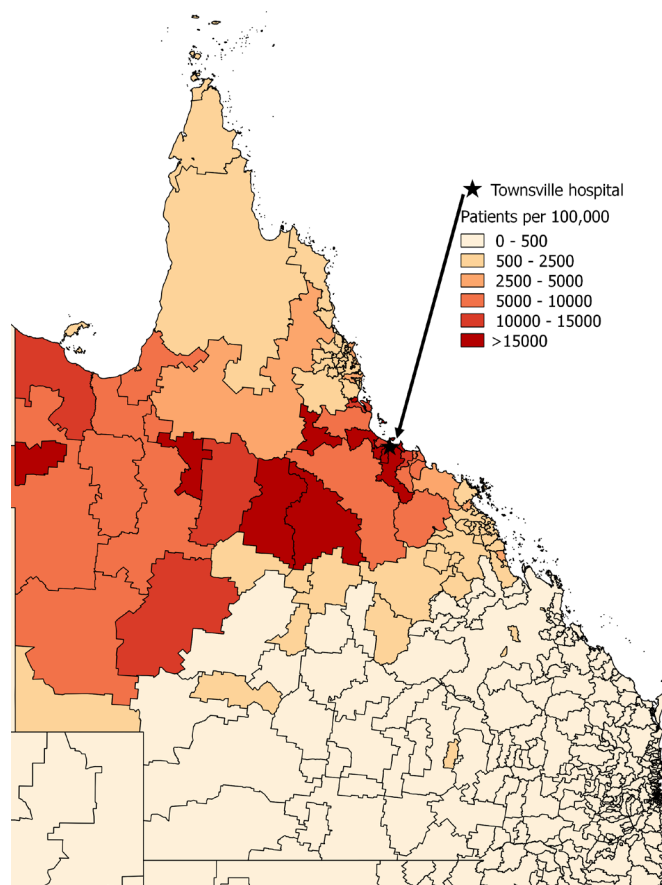


Figure 2 Heat map of cohort patients per 100 000 shown by postcode of domicile according to hospital registration at entry into cohort.

in two studies of community-acquired pneumonia.^{22 23} Linked administrative data was shown to reliably ascertain incident colorectal and lung cancer diagnoses when compared with the New South Wales Cancer Registry.²⁴ Other Australian researchers have studied the accuracy of ICD-10-AM codes for diagnoses of childhood influenza and pertussis.²⁵ While demonstrating high specificity and positive predictive value, the authors conclude that addition of laboratory data increases the precision of retrospective, population level diagnosis of paediatric respiratory infection. The incorporation of pathology results in the cohort described in this database will allow precise characterisation of the infectious diseases cohort we have assembled. For example, the large volume of microbiology data will allow for analysis of key areas such as antimicrobial resistant infections and their influence on clinical outcomes and provide greater precision for diagnosis (eg, site of infection in sepsis).

CONCLUSIONS

Numerous analysis of risks for, and outcomes of, disease and organism-specific infections, healthcare costs and antimicrobial side effects will all be undertaken in the future using these data. These studies will incorporate measures such as the Socio-Economic Index for Areas²⁶ to assess the

Table 3 Total cases of diseases due to selected microbial pathogens

Diseases	N
<i>Staphylococcus aureus</i> sepsis	6802
Skin and soft tissue infection	3182
Osteomyelitis	670
Arthritis	215
Phlebitis and thrombophlebitis	250
Infective endocarditis	172
<i>Streptococcus pyogenes</i> infection	1197
Skin and soft tissue infection	693
<i>Streptococcus pneumoniae</i> sepsis	515
Pneumonia	435
Urinary tract infection	
Pyelonephritis	1391
Cystitis	314
Urethritis	22
Prostatitis	118
Abscess	52
Other	9083
Pneumonia	
Viral	769
Bacterial	2853
Other	4151
Influenza	1738
Meningitis	
Viral	240
Bacterial	123
Tropical infection	
Melioidosis	84
Dengue	88
Ross River	48
Q fever	139

impact of socioeconomic disadvantage on outcomes of infectious diseases occurring in hospitalised patients. As hospitalisation data are available before the admission that led the patient to be included in the cohort, there will be an opportunity to assess presentations and investigation findings that predated diagnosis. Similarly, the extensive information from subsequent hospitalisations will allow detailed analysis of long-term health effects after severe infectious diseases. The use of linked pathology data may retrospectively improve definition of severe infectious diseases such as invasive group A streptococcal infection by a systematic search for positive cultures from sterile sites.

Strengths and limitations of this study

The main strength of this cohort is its large size and unique description of inpatients diagnosed with infectious diseases at an Australian tropical zone hospital. The intricate

relational database has provided a resource that can be easily searched. In future analyses, the linkage of numerous data sources to provide a granular description of patient disease and treatment will enable the use of a variety of statistical methods. Similarly, pathology and pharmacy antimicrobial dispensing data availability allows for precise case definition and analysis of treatment response.

The main study limitations are that it is based on data sets from a single hospital so future findings will not be applicable to the general Australian population and the validity of cohort studies rely on the accuracy of clinical coding. Despite these limitations, this database will be a rich source of information for future cohort studies of the epidemiology of infectious diseases in the catchment area of the only tertiary hospital in North Queensland.

Contributors DE conceived the study idea, defined the original study protocol and is responsible for the ethics applications and the ethical reporting of the study. DE, EM and LV are responsible for the study methodology. MM developed the relational database. MH and OA are responsible for ICD10-AM codes extraction, categorisation and quality assessment. OA carried out the data analysis. All authors have read and approved the final manuscript. DE and OA drafted the final version of this manuscript.

Funding This work was supported by a financial grant from the Townsville Hospital and Health Service Study Education Research Trust Account.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not required.

Ethics approval This project, HREC/16/QTHS/221, was approved by the Townsville Hospital and Health Service (THHS) Human Research Ethics Committee. A waiver of consent for access to anonymised data was approved under the Queensland Public Health Act (RD007802).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data were obtained from a third party and are not publicly available. Due to restrictions and confidentiality, the data sets generated during and/or analysed during this study are not publicly available.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iD

Oyelola Adegboye <http://orcid.org/0000-0002-9793-8024>

REFERENCES

- Burgess K, Gilbert M, McIntyre J, *et al*. *Admitted patient care 2016-17: Australian hospital statistics*. Canberra: Australian Institute of Health and Welfare, 2018.
- Barnett AG, Page K, Campbell M, *et al*. The increased risks of death and extra lengths of hospital and ICU stay from hospital-acquired bloodstream infections: a case-control study. *BMJ Open* 2013;3:e003587.
- Australian Institute of Health and Welfare. *Australian burden of disease study: impact and causes of illness and death in Australia 2015*. Canberra, 2019.
- Moore HC, Blyth CC. Optimising the use of linked administrative data for infectious diseases research in Australia. *Public Health Res Pract* 2018;28. doi:10.17061/phrp2821810. [Epub ahead of print: 14 Jun 2018].
- Charlson ME, Pompei P, Ales KL, *et al*. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;40:373-83.

- 6 Quan H, Sundararajan V, Halfon P, *et al.* Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med Care* 2005;43:1130–9.
- 7 Gotland N, Uhre ML, Mejer N, *et al.* Long-term mortality and causes of death associated with *Staphylococcus aureus* bacteremia. A matched cohort study. *J Infect* 2016;73:346–57.
- 8 Myint PK, Hawkins KR, Clark AB, *et al.* Long-Term mortality of hospitalized pneumonia in the EPIC-Norfolk cohort. *Epidemiol Infect* 2016;144:803–9.
- 9 Ternhag A, Cederström A, Törner A, *et al.* A nationwide cohort study of mortality risk and long-term prognosis in infective endocarditis in Sweden. *PLoS One* 2013;8:e67519.
- 10 Adegboye OA, McBryde ES, Eisen DP. Epidemiological analysis of association between lagged meteorological variables and pneumonia in wet-dry tropical North Australia, 2006–2016. *J Expo Sci Env Epid* 2019;1–11.
- 11 Tew M, Dalziel KM, Petrie DJ, *et al.* Growth of linked hospital data use in Australia: a systematic review. *Aust Health Rev* 2017;41:394–400.
- 12 Karki S, Gidding HF, Newall AT, *et al.* Risk factors and burden of acute Q fever in older adults in New South Wales: a prospective cohort study. *Med J Aust* 2015;203:438.
- 13 Lim FJ, Blyth CC, Fathima P, *et al.* Record linkage study of the pathogen-specific burden of respiratory viruses in children. *Influenza Other Respir Viruses* 2017;11:502–10.
- 14 Marquess J, Hu W, Nimmo GR, *et al.* Spatial analysis of community-onset *Staphylococcus aureus* bacteremia in Queensland, Australia. *Infect Control Hosp Epidemiol* 2013;34:291–8.
- 15 Lim FJ, Blyth CC, Levy A, *et al.* Using record linkage to validate notification and laboratory data for a more accurate assessment of notifiable infectious diseases. *BMC Med Inform Decis Mak* 2017;17:86.
- 16 Dendle C, Sundararajan V, Spelman T, *et al.* Splenectomy sequelae: an analysis of infectious outcomes among adults in Victoria. *Med J Aust* 2012;196:582–6.
- 17 Valentine JC, Morrissey CO, Tacey MA, *et al.* A population-based analysis of invasive fungal disease in haematology-oncology patients using data linkage of state-wide registries and administrative databases: 2005 - 2016. *BMC Infect Dis* 2019;19:274.
- 18 Sax Institute. 45 and up study: Sax Institute; 2019, 2019. Available: <https://www.saxinstitute.org.au/our-work/45-up-study/> [Accessed 9 July 2019].
- 19 Woolley I, Jones P, Spelman D, *et al.* Cost-Effectiveness of a post-splenectomy Registry for prevention of sepsis in the asplenic. *Aust N Z J Public Health* 2006;30:558–61.
- 20 Australian Institute of Health and Welfare. *Overweight and obesity: an interactive insight*. Canberra, Australia: Australian Government, 2019. <https://www.aihw.gov.au/reports/overweight-obesity/overweight-and-obesity-an-interactive-insight/contents/prevalence>
- 21 Ford E, Carroll JA, Smith HE, *et al.* Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016;23:1007–15.
- 22 Skull SA, Andrews RM, Byrnes GB, *et al.* Hospitalized community-acquired pneumonia in the elderly: an Australian case-cohort study. *Epidemiol Infect* 2009;137:194–202.
- 23 Skull SA, Andrews RM, Byrnes GB, *et al.* ICD-10 codes are a valid tool for identification of pneumonia in hospitalized patients aged > or = 65 years. *Epidemiol Infect* 2008;136:232–40.
- 24 Goldsbury D, Weber M, Yap S, *et al.* Identifying incident colorectal and lung cancer cases in health service utilisation databases in Australia: a validation study. *BMC Med Inform Decis Mak* 2017;17:23.
- 25 Moore HC, Lehmann D, de Klerk N, *et al.* How accurate are International classification of Diseases-10 diagnosis codes in detecting influenza and pertussis hospitalizations in children? *J Pediatric Infect Dis Soc* 2014;3:255–60.
- 26 Australian Bureau of Statistics. Census of population and housing: socio-economic indexes for areas (SEIFA), Australia, 2016 Canberra, Australia: Australian government; 2018. Available: <https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/2033.0.55.001~2016~Main%20Features~SEIFA%20Basics~5> [Accessed 06 Sep 2019].