



BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email [info.bmjopen@bmj.com](mailto:info.bmjopen@bmj.com)

# BMJ Open

## A linked hospital administrative data-set describes inpatient infectious diseases diagnoses in Far North Queensland

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2019-034845
Article Type:	Cohort profile
Date Submitted by the Author:	09-Oct-2019
Complete List of Authors:	eisen, damon; Townsville Hospital; James Cook University Division of Tropical Health and Medicine Mcbryde, Emma; James Cook University Division of Tropical Health and Medicine, Australian Institute of Tropical Health and Medicine Vasanthakumar, luke; Townsville Hospital Murray, Matthew; Commonline Pty Ltd Harings, Miriam; Townsville Hospital ADEGBOYE, Oyelola; James Cook University Division of Tropical Health and Medicine, Australian Institute of Tropical Health & Medicine
Keywords:	INFECTIOUS DISEASES, Epidemiology < INFECTIOUS DISEASES, Tropical medicine < INFECTIOUS DISEASES

SCHOLARONE™  
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

**A linked hospital administrative data-set describes inpatient infectious diseases diagnoses in Far  
North Queensland: A Cohort profile**

Damon P Eisen <sup>1,2</sup>

Emma S McBryde <sup>3</sup>

Luke Vasanthakumar <sup>1</sup>

Matthew Murray<sup>4</sup>

Miriam Harings <sup>1</sup>

Oyelola Adegboye <sup>3\*</sup>

<sup>1</sup> The Townsville Hospital, 100 Angus Smith Drive, Douglas, Queensland, Australia, 4814

<sup>2</sup> College of Medicine and Dentistry, James Cook University, 1 James Cook Drive, Douglas, Queensland,  
Australia, 4814

<sup>3</sup> Australian Institute of Tropical Health and Medicine, James Cook University, 1 James Cook Drive,  
Douglas, Queensland, Australia, 4814

<sup>4</sup> Commonline Pty Ltd, Townsville, Queensland, Australia, 4810

**\*Corresponding author:**

Oyelola Adegboye ([oyelola.adegboye@jcu.edu.au](mailto:oyelola.adegboye@jcu.edu.au))

Australian Institute of Tropical Health and Medicine, James Cook University, 1 James Cook Drive,  
Douglas, Queensland, Australia, 4814

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**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 study of Townsville Hospital inpatients discharged with an ICD-10-Australian Modification code for an infectious disease between 1/1/2006 and 31/12/2016 was undertaken. This utilised linked anonymised data from; hospital administrative sources, diagnostic pathology, pharmacy dispensing, public health and the National Death Registry. A Created Study ID (PU\_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 either editable Structured Query Language or a selection of preset queries.

**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 identify as Aboriginal or Torres Strait Islanders. Invasive staphylococcal, pneumococcal and Group A streptococcal infections and influenza were common in this cohort. The commonest comorbidities were; smoking (43.95%), chronic renal disease (17.93%), diabetes (16.71%), cancer (13.59%) chronic and pulmonary disease (12.42%). Mortality over the eleven-year period was 20%.

**Future plans** This complex relational-database reutilising hospital information describes a cohort from a single tropical Australian hospital of in-patients with infectious diseases. We plan to explore analyses of risks, clinical outcomes, health care costs and antimicrobial side effects in site and organism specific infections.

**Key words:** data-linkage, relational database, epidemiology, infectious diseases, hospital

**Funding statement:**

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

**Competing interests:**

None of the authors have any competing interests to declare.

**Word counts:**

Abstract: 257

Manuscript: 2546

For peer review only

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**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 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 4<sup>th</sup>, cellulitis 9<sup>th</sup> and ‘other sepsis’ 16<sup>th</sup>. 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 3<sup>rd</sup> to 5<sup>th</sup> 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.

There currently 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) The data-linkage process is a powerful tool for analysis of various disease cohorts. It is a value-adding re-use of previously acquired patient information that represents a rich research resource. We have developed a database that includes information from numerous sources to analyse the epidemiology of infection occurring in inpatients at the Townsville Hospital. This cohort database, from an eleven-year period, will be used to analyse the incidence, risk factors and clinical outcomes of patients with hospital admissions involving infectious disease.

## Cohort description

### Setting.

The Townsville Hospital and Health Service (THHS) serves a resident population of 239,000. 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 1100mm.

### Cohort selection.

A cohort of Townsville Hospital in-patients 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 the 11-year period January 1st 2006 to December 31st 2016. Information from the episode of care that led to cohort inclusion and all subsequent inpatient admissions was provided.

The ICD-10-AM codes primarily used to select the patient cohort were A00–B99 Infectious and parasitic diseases. 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.

Data bases.

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: 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.
- 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; all 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 was 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.

Some assumptions were made during the processing of data. If pathology data was collected during the same date range as an admission then this was inserted as happening as part of the admission even though we have no admission identifier recorded within the pathology.

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.

## Ethics approval.

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

## 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.

## Findings to date

The database consisted of linked information from 41,367 patients with 378,487 admissions and 1,869,239 diagnostic or procedure codes. A summary of the data and the datafields is included in Supplementary Tables S1, S2a, S2b and S2c. A database structure was designed to best accommodate

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

the contents of the supplied data and the available identifiers within it. The database relational structure is shown in Figure 1.

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 distribution of age at first admission was skewed towards older subjects. (Table 1) Similarly, the total number of admissions was markedly skewed towards higher numbers. This is due to the significant number of haemodialysis patients who had a median of six admissions with interquartile range of 2-41 over the eleven-year duration of the cohort study. A large proportion of the patients identify 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 over represented amongst these patients. The geographic location of patient domicile as determined by postcode at the time of inpatient registration is shown in Figure 2.

A high proportion of patients had comorbidities with 44% being smokers. Other common comorbidities were; renal impairment, cancer, alcohol abuse, diabetes mellitus and chronic respiratory disease. Multiple comorbidities were present in 67% of patients. Only four percent of patients had an ICD-10-AM code for obesity.

The overall eleven-year all-cause mortality was exactly 20%. Male sex, identifying as Aboriginal and Torres Strait Islander and all measured comorbidity groupings increased the risk of death after adjusting for other variables (Figure 3).

Table 2 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.

For peer review only

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Discussion

This longitudinal cohort study describes patients admitted to 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 and well as in Australian communities in general.

Cohort patients had numerous hospital admissions with a median of six during the study period. A high proportion of patients identified as Aboriginal and Torres Strait Islanders. Comorbidities were highly prevalent. The crude 11-year mortality rate was 20%. All-cause mortality rates from Australian cohorts of patients with selected, highly morbid, infections such as *Staphylococcus aureus* bacteraemia (28%, 2-5 year follow up) (5), community acquired pneumonia (60.4%, mean follow up 6.1 years) (6) and infective endocarditis (14.7%, 1 – 5 year follow up) have been described. (7) These studies all demonstrated increased all-cause mortality of the infectious diseases cohorts compared with controls.

Unsurprisingly, analysis of our cohort shows that comorbidities including diabetes, cancer, lung disease and renal failure are associated with higher rates of mortality. The failure to “close the gap” (8) in health outcomes in Indigenous patients is highlighted both by their high frequency in this cohort (nearly 15%) and by the increased all-cause mortality we have specifically demonstrated in this population, after taking age into account.

When we consider the patterns of infectious diseases found in this cohort *S. aureus* was the commonest pathogen identified followed by influenza and Group A streptococcus. Skin and soft tissue was the commonest site of infection followed by the lungs.

This data-linkage 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 patients hospitalized with an infectious disease. For example, linkage of pathology results (microbiological [routine, serological and PCR based diagnostic tests], haematology [all full blood examination parameters and coagulation profile] and biochemistry [urea and electrolytes, liver function tests and C-reactive protein]) 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.

There has been a sustained increase in the numbers of cohort studies using linked administrative hospital datasets including in Australia. (9) However, infectious diseases studies are in the minority compared with cardiovascular, health services, cancer and maternal health research. Australian cohort studies that utilise data linkage to describe infectious diseases mostly rely on ICD-10-AM diagnostic codes and death registry information. Some also incorporate notifiable diseases data (10) but, overall, studies incorporating pathology data are few (11, 12).

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 systems (13) and hospital discharge coding (11). These studies both demonstrated underascertainment of childhood respiratory tract diseases.

There are numerous Australian cohort studies of various infectious diseases. Patient outcomes, risks and other disease characteristics have been studied in; organ specific infections, respiratory viral infections (11); infection specific, Q fever (10) and *S. aureus* bacteraemia (12) as well as patient specific, such as asplenic (14) and haematology-oncology

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

patients. (15) 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 aging (16), Triple I Western Australian birth cohort (13) and Victorian Post-Splenectomy Registry (17).

There are inherent limitations of retrospective, databases defined by ICD-10-AM codes, such as this study. Some important clinical information is underrepresented. This is exemplified in this cohort study where only 3.9% 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. (18) This potential underestimate may derive from ICD-10-AM coding for obesity only being allocated where patients are being treated for actively assessed by a dietitian for obesity. More frequently, inpatients at the Townsville Hospital were 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. (19) 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 in Australian research, for example in two studies of community acquired pneumonia (20). Other Australian researchers have studied the accuracy of ICD-10-AM codes for diagnoses of childhood influenza and pertussis. (21) 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 microbiology, haematology and biochemistry results in the cohort described in this database allows 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 (e.g. site of infection in sepsis).

## Conclusions

Numerous analysis of risks for, and outcomes of, disease and organism specific infections, health care 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 Areas (22) to assess the impact of socioeconomic disadvantage on outcomes of infectious diseases occurring in hospitalized patients. As hospitalization 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 study is the large cohort of inpatients diagnosed with infectious diseases described and the multiple health outcomes therein. Analyses have not been limited to a prespecified range of conditions. The ability to link numerous pre-existing data sources to provide a granular description of patient disease and treatment will enable the use of a variety of statistical methods. Similarly, clinical pathology and pharmacy antimicrobial dispensing data availability allows for precise case definition and analysis of treatment response.

We acknowledge the following limitations: this study is based on data sets from a single hospital so our 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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

source of information for cohort studies of the epidemiology of infectious diseases in the catchment area of the only tertiary hospital in North Queensland.

Author’s contribution

DE conceived the study idea and defined the original study protocol. DE 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, categorization 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.

For peer review only

## References.

1. Burgess K, Gilbert M, McIntyre J, Mole T. Admitted patient care 2016-17: Australian hospital statistics. Canberra: Australian Institute of Health and Welfare. 2018.
2. Barnett AG, Page K, Campbell M, Martin E, Rashleigh-Rolls R, Halton K, 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(10):e003587.
3. Australian Institute of Health and Welfare. Australian Burden of Disease Study: impact and causes of illness and death in Australia 2015. Cat. No. BOD 22 ed. Canberra 2019.
4. Moore HC, Blyth CC. Optimising the use of linked administrative data for infectious diseases research in Australia. *Public Health Res Pract*. 2018;28(2).
5. Gotland N, Uhre ML, Mejer N, Skov R, Petersen A, Larsen AR, et al. Long-term mortality and causes of death associated with *Staphylococcus aureus* bacteremia. A matched cohort study. *J Infect*. 2016;73(4):346-57.
6. Myint PK, Hawkins KR, Clark AB, Luben RN, Wareham NJ, Khaw KT, et al. Long-term mortality of hospitalized pneumonia in the EPIC-Norfolk cohort. *Epidemiol Infect*. 2016;144(4):803-9.
7. Ternhag A, Cederstrom A, Torner A, Westling K. A nationwide cohort study of mortality risk and long-term prognosis in infective endocarditis in Sweden. *PLoS One*. 2013;8(7):e67519.
8. Hoy WE. "Closing the gap" by 2030: aspiration versus reality in Indigenous health. *Med J Aust*. 2009;190(10):542-4.
9. Tew M, Dalziel KM, Petrie DJ, Clarke PM. Growth of linked hospital data use in Australia: a systematic review. *Aust Health Rev*. 2017;41(4):394-400.
10. Karki S, Gidding HF, Newall AT, McIntyre PB, Liu BC. Risk factors and burden of acute Q fever in older adults in New South Wales: a prospective cohort study. *Med J Aust*. 2015;203(11):438.

11. Lim FJ, Blyth CC, Fathima P, de Klerk N, Moore HC. Record linkage study of the pathogen-specific burden of respiratory viruses in children. *Influenza Other Respir Viruses*. 2017;11(6):502-10.
12. Marquess J, Hu W, Nimmo GR, Clements AC. Spatial analysis of community-onset *Staphylococcus aureus* bacteremia in Queensland, Australia. *Infect Control Hosp Epidemiol*. 2013;34(3):291-8.
13. Lim FJ, Blyth CC, Levy A, Fathima P, de Klerk N, Giele C, 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(1):86.
14. Dendle C, Sundararajan V, Spelman T, Jolley D, Woolley I. Splenectomy sequelae: an analysis of infectious outcomes among adults in Victoria. *Med J Aust*. 2012;196(9):582-6.
15. Valentine JC, Morrissey CO, Tacey MA, Liew D, Patil S, Peleg AY, 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(1):274.
16. Sax Institute. 45 and Up Study. 2019 [Available from: <https://www.saxinstitute.org.au/our-work/45-up-study/>].
17. Woolley I, Jones P, Spelman D, Gold L. Cost-effectiveness of a post-splenectomy registry for prevention of sepsis in the asplenic. *Aust N Z J Public Health*. 2006;30(6):558-61.
18. Australian Institute of Health and Welfare. Overweight and obesity: an interactive insight Canberra, Australia: Australian Government; 2019 [updated 19/7/19. Available from: <https://www.aihw.gov.au/reports/overweight-obesity/overweight-and-obesity-an-interactive-insight/contents/prevalence>].
19. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc*. 2016;23(5):1007-15.

- 1  
2  
3 20. Skull SA, Andrews RM, Byrnes GB, Campbell DA, Kelly HA, Brown GV, et al. Hospitalized  
4 community-acquired pneumonia in the elderly: an Australian case-cohort study. *Epidemiol Infect.*  
5  
6 2009;137(2):194-202.  
7  
8  
9  
10  
11 21. Moore HC, Lehmann D, de Klerk N, Smith DW, Richmond PC, Keil AD, et al. How Accurate Are  
12 International Classification of Diseases-10 Diagnosis Codes in Detecting Influenza and Pertussis  
13 Hospitalizations in Children? *J Pediatric Infect Dis Soc.* 2014;3(3):255-60.  
14  
15  
16  
17  
18 22. Australian Bureau of Statistics. Census of Population and Housing: Socio-Economic Indexes for  
19 Areas (SEIFA), Australia, 2016 Canberra, Australia: Australian Government; 2018 [updated 27/3/2018].  
20  
21 Available from:  
22  
23 [https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/2033.0.55.001~2016~Main%20F](https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/2033.0.55.001~2016~Main%20Features~SEIFA%20Basics~5)  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Table 1: Cohort characteristics (n=41367)

Characteristics	Number	Mean	Median	SD	Q1	Q3
Age (years) at first admission	41367	43.15	49	24.44	26	68
Total admissions	37783	100.81	6	230.6	2	41
Demographics	Number of patients	Percentage				
Male	21299	51.49				
Female	20068	48.51				
11 year mortality						
Dead	8274	20				
Indigenous status						
Aboriginal but not TSI <sup>a</sup> origin	4763	11.51				
Aboriginal and TSI	550	1.33				
Neither Aboriginal nor TSI	35187	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				
Smoking	18179	43.95				
Alcohol	4743	11.47				
Recreational drug use	600	1.45				
Immunocompromised						

Cancer <sup>b</sup>	5621	13.59
Kidney failure	7419	17.93
HIV	114	0.28
Splenectomy	77	0.19
Coronary vascular disease	4107	9.9
Diabetes mellitus	6912	15.31
Pulmonary disease	5140	12.42
Malnutrition	4605	11.13
Obesity	1633	3.95
Cerebrovascular disease (Stroke)	3294	7.96

<sup>a</sup> Aboriginal and Torres Strait Islander

<sup>b</sup> Includes; melanoma, breast, gastro-intestinal, lung and genitourinary, hematological, head and neck malignancies.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Table 2. 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>S. 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
---------	-----

---

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Figure 1. Representation of relational database constructed showing links between fields from incorporated administrative, clinical and death registry information. (See Appendix Table 1 for detailed description of fields.)**

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

**Figure 3. Multivariable model for predictors of mortality**

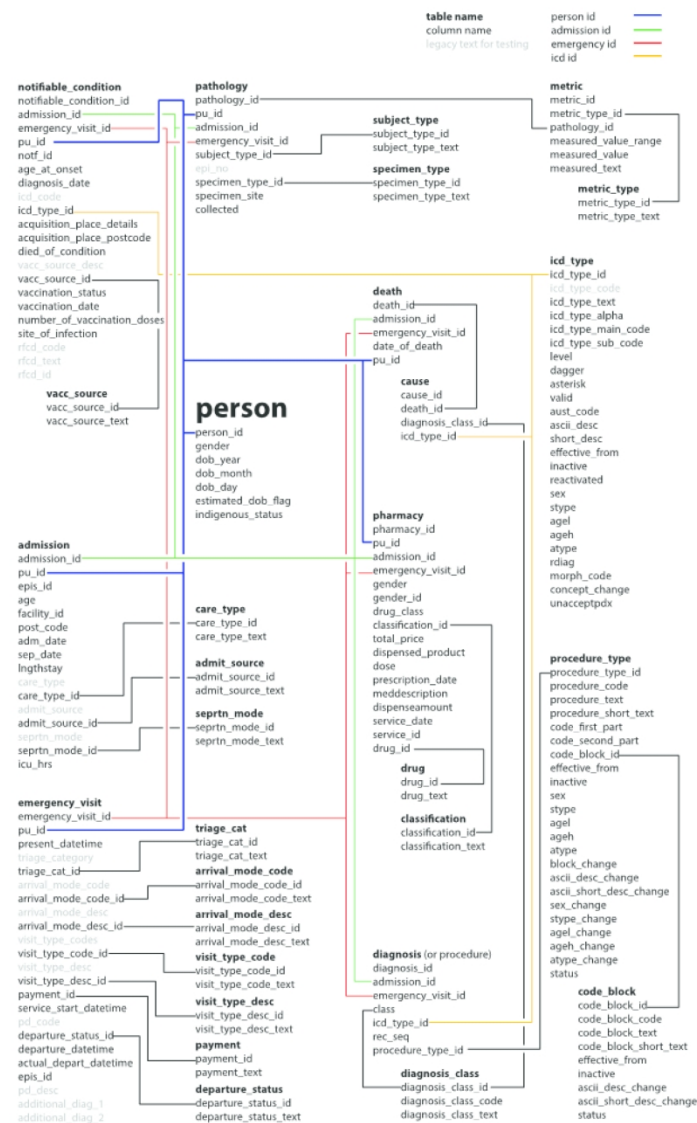


Figure 1. Representation of relational database constructed showing links between fields from incorporated administrative, clinical and death registry information. (See Supplementary materials for detailed description of fields.)

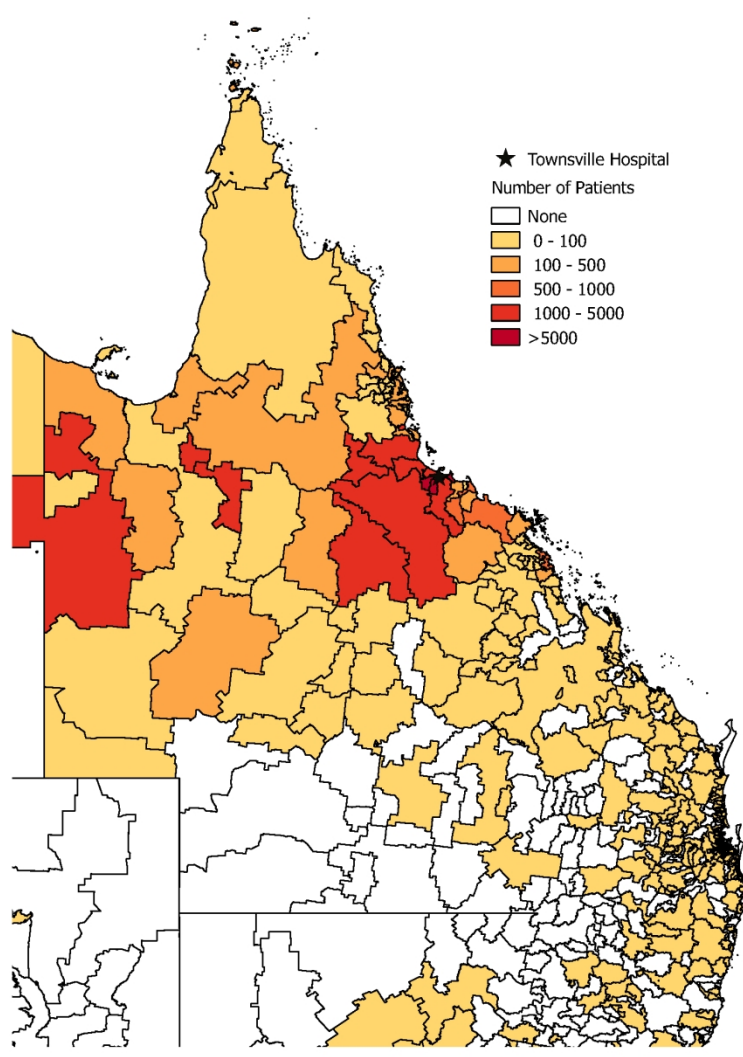


Figure 2. Heat map of cohort patients shown by postcode of domicile according to hospital registration at entry into cohort.

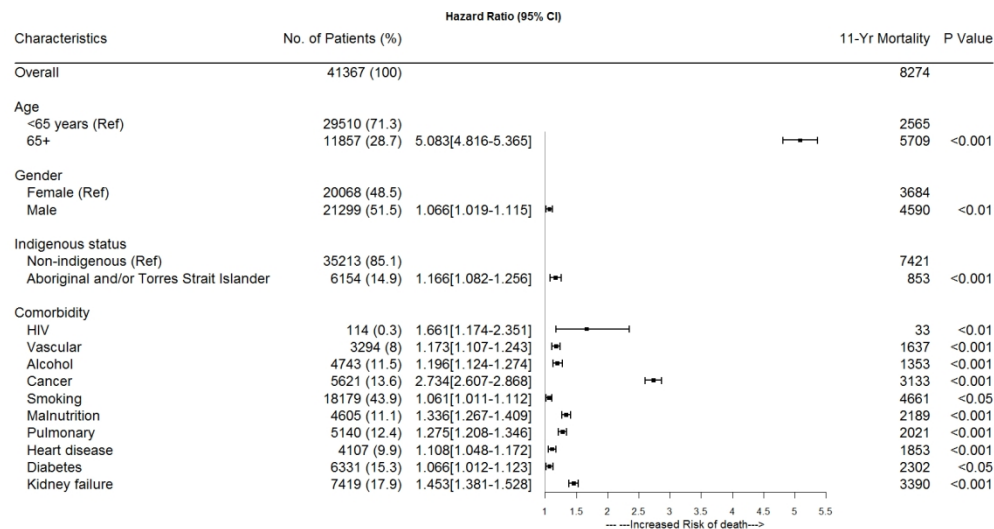


Figure 3. Multivariable model for predictors of mortality

Table S1. Sources and numbers of patient records from the cohort of 41,367 unique people, identified and extracted from Queensland Health Admitted Patient Data Collection (QHAPDC) and the Queensland Health Statistical Services Branch Master Linkage File, based on a selection of infectious disease codes for the period 1st January 2006 to 31st December 2016.

Table Name (Output Name)	Counts	Description
<b>QHAPDC_Single_Out</b> (QHAPDC_Single_Out_oct_18.csv)	records= 378,487 -- unique people= 41,367	QHAPDC Admissions data
<b>QHAPDC_Morb_Out</b> (QHAPDC_Morb_Out.csv)	records= 1,869,239 -- unique people= 41,367	QHAPDC Morbidity data
<b>EDIS</b> (EDIS_Out.csv)	records= 204,112 -- unique people= 35,316	EDIS data
<b>Death_Out</b> (Death_Out_oct_18.csv)	records= 8,274 -- unique people= 8,274	Death data
<b>VivNocs_out</b> (VivNocs_out.csv)	records= 10,501 -- unique people= 7,060	NOCS and VIVAS data
<b>Pathology data (AUSLAB)</b>		
<b>Microbiology_out</b> (Microbiology_out.csv)	records= 296,949 -- unique people= 35,736	Pathology data (AUSLAB)
<b>Serology_out</b> (Serology_out.csv)	records= 12,840 -- unique people= 6,852	Pathology data (AUSLAB)
<b>PCR_out</b> (PCR_out.csv)	records= 10,376 -- unique people= 6,309	Pathology data (AUSLAB)
<b>Haematology_out</b> (Haematology_out.csv)	records= 621,030 -- unique people= 38,406	Pathology data (AUSLAB)
<b>Biochemistry_out</b> (Biochemistry_out.csv)	records= 603,337 -- unique people= 38,110	Pathology data (AUSLAB)
<b>Remainder_out</b> (Remainder.csv)	records= 21,537 -- unique people= 10,229	Pathology data (AUSLAB)
<b>Drug Dispensing Data</b>		
<b>Pyxis_out</b> (Pyxis_out.csv)	records= 147,405 -- unique people= 10,404	Drug Dispensing Data
<b>iPharmacy</b> (iPharm_out.csv)	records= 115,505 -- unique people= 21,335	Drug Dispensing Data

Table S2a. Queensland Hospital Admitted Patient Data Collection, Emergency Department Information System, Death Registry and Notifiable Conditions fields included in the dataset.

Queensland Hospital Admitted Patient Data Collection (QHAPDC)		Emergency Department Information System		Death Registry	
Variable Name	Variable Description	Variable Name	Variable description	Variable Name	Variable Description
PU_ID	Person level identifier	PU_ID	Person level identifier	PU_ID	Person level identifier
EPIS_ID	Episode level identifier	PAT_SEX	Sex	Date_Of_Death	Person date of death (dd-mm-yyyy)
AGE	Age of the person at time of admission	PAT_DATE_OF_BIRTH	Date of birth	AUSE	Cause of death (free text)
POST_CODE	Post code at the time of admission	PAT_DOB_EST_FLAG	Estimated DOB Flag	COD	Underlying cause of death (ICD-10)
ADM_DATE	Full date of the persons admission	PRESENTATION_DATETIME	Present datetime (dd.mmm.yyyy.hh.mm)	EC_AXIS_1 - 14	Other cause of death (ICD-10) 1 - 14
SEP_DATE	Full date of the persons discharge	TRIAGE_CATEGORY	Triage category		
SEX	Sex of the person relating to the admission	TRIAGE_DATETIME	Triage date time		
INDIG_STATUS	Indigenous status of the person	ARRIVAL_TRANSPORT_MODE	Arrival mode description		
LNGTHSTAY	Length the person was in hospital excluding periods of leave	VISIT_TYPE_CODE	Visit type codes		
CARE_TYPE	The nature of the treatment/care provided to a patient during an episode of care	VISIT_TYPE	Visit type description		
ADMIT_SOURCE	The source of referral/transfer (admission source) of a patient immediately before they are admitted	PAYMENT_CLASS	Payment ttatus		
SEPRTN_MODE	The mode of separation: place to which a patient is referred immediately following separation	SERVICE_START_DATETIME	Service start datetime		
ICU_HRS	Total number of hours and minutes a patient has spent in ICU	PRINCIPAL_DIAGNOSIS	PD Code		
ICD_TYPE	Principal Diagnoses (PD), Other/Additional Diagnoses (OD), Procedure Code (PR)	PRINCIPAL_DIAGNOSIS	PD Description		
ICD_CODE	ICD-10-AM and ACHI classifications of diagnoses and procedures	ADDITIONAL_DIAGNOSIS_1	Additional Diagnosis 1		
		ADDITIONAL_DIAGNOSIS_2	Additional Diagnosis 2		
		EPISODE_END_STATUS_CODE	Departure status		
		EPISODE_END_DATETIME	Departure datetime		
		PHYSICAL_DEPART_DATETIME	Actual departure datetime		

Table S2b. AUSLAB microbiological results included in dataset

Patient identifier	Patient identifier
Specimen	Specimen
Specimen Site	Specimen Site
Collection time (dd-mmm-yyyy hh-mm)	Collection time (dd-mmm-yyyy hh-mm)
Blood Culture	Parainfluenza 1 (NAA)
Incubation Time Until Positive	
Positive Bottles	Parainfluenza 2 (NAA)
Culture Comment	Parainfluenza 3 (NAA)
Organism	Influenza A RNA (NAA)
Fungal Culture	Resp Syn Virus (NAA)
Mycobacterial Culture Result	Adenovirus DNA (NAA)
Organism	Influenza B RNA (RNA)
Sensitivities	HSV 1 DNA (NAA)
	HSV 2 DNA (NAA)
Q Fever Ph2 IgG (EIA)	Human Metapneumovirus RNA (NAA)
Q Fever Ph2 IgM (EIA)	Human herpes 6 DNA (NAA)
Q Fever Ph1 IgG (IF)	Human herpes 7 DNA (NAA)
Q Fever Ph2 IgG (IF)	Human herpes 8 DNA (NAA)
Q Fever Ph2 IgM (IF)	Varicella zoster DNA (NAA)
Q Fever DNA	Enterovirus RNA (NAA)
Leptospira sp. IgM (EIA)	Dengue group RNA (NAA)
Leptospirosis NAA	Denque Universal (TAQ)
Atypical Serology	Barmah Forest RNA (TAQ)
M. pneumoniae Total Ab	Ross River Virus RNA (TAQ)
L. pneumophila 1 antigen	J. encephalitis RNA (TAQ)
S. pneumoniae antigen	Kunjin Virus (TAQ)
CMV DNA	Murray Valley RNA (TAQ)
CMV IgM (EIA)	Chikungunya RNA (TAQ)
CMV IgM (EIA)	West Nile Virus (TAQ)
EBV IgG (EIA)	Rift Valley Fever Virus (TAQ)
EBV IgM (EIA)	P. jiroveci DNA (NAA)
Epstein-Barr Virus IgA	N. meningitidis DNA (NAA)
EBV DNA	S. pneumoniae DNA (NAA)
Dengue NS1 antigen	Non tuberculous mycobacteria PCR
Cryptococcal antigen	M. ulcerans PCR
Aspergillus galactomannan antigen	TB PCR
	M. leprae PCR

Table S2c. AUSLAB pathology results and Pharmacy drug dispensing information included in the dataset

Biochemistry	Haematology	Other	Pharmacy
Patient identifier	Patient identifier	Patient identifier	iPharm
Specimen	Specimen	Specimen	
Specimen Site	Specimen Site	Specimen Site	
Collection time (dd-mmm-yyyy hh-mm)	Collection time (dd-mmm-yyyy hh-mm)	Collection time (dd-mmm-yyyy hh-mm)	Patient identifier
Sodium	Haemoglobin	HbA1c	Gender
Potassium	White Cell Count	25-Hydroxy-Vitamin D	Drug class
Chloride	Platelets	1,25-Dihydroxy-Vitamin D	Classification
Bicarbonate	Haematocrit	IGG	Total price
Urea/Creat. Ratio	Mean corpuscular haemoglobin	IGG1	Dispensed product
Glucose	Red Cell Count	IGG2	Dose
Calcium	Mean corpuscular volume	IGG3	Prescription date
Magnesium	Neutrophils	IGG4	
Phosphate	Lymphocytes	IGA	Pyxis
Urate	Monocytes	IGM	
Protein	Eosinophils	Antinuclear antibody	
Albumin	Basophils	Rheumatoid factor	Patient identifier
Bilirubin	Prothrombin Time	Rheumatoid factor (fluid)	Medication description
Bilirubin (conjugated)	APTT	Total protein	Dispensed amount
Alkaline phosphatase		Albumin	Service date
Gamma GT		Total globulin	
Aspartate transaminase		Monoclonal protein	
Alanine transaminase		Serum EPP comment	
Lactate dehydrogenase		Angiotensin converting enzyme	
Anion gap		ACE (CSF)	
Osmolality (calculated)			
Urea/Creat. Ratio			
Globulin			
Corrected calcium			
Corrected potassium			

estimated glomerular filtration rate  
C-reactive protein

---

For peer review only

# BMJ Open

## Linking administrative datasets of inpatient infectious diseases diagnoses in Far North Queensland: A cohort profile

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2019-034845.R1
Article Type:	Cohort profile
Date Submitted by the Author:	15-Dec-2019
Complete List of Authors:	eisen, damon; Townsville Hospital; James Cook University Division of Tropical Health and Medicine Mcbryde, Emma; James Cook University Division of Tropical Health and Medicine, Australian Institute of Tropical Health and Medicine Vasanthakumar, luke; Townsville Hospital Murray, Matthew; Commonline Pty Ltd Harings, Miriam; Townsville Hospital ADEGBOYE, Oyelola; James Cook University Division of Tropical Health and Medicine, Australian Institute of Tropical Health & Medicine
<b>Primary Subject Heading</b>:	Infectious diseases
Secondary Subject Heading:	Epidemiology, Public health, Research methods
Keywords:	INFECTIOUS DISEASES, Epidemiology < INFECTIOUS DISEASES, Tropical medicine < INFECTIOUS DISEASES

SCHOLARONE™  
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

# Linking administrative datasets of inpatient infectious diseases diagnoses in Far North Queensland: A cohort profile

Damon P Eisen <sup>1,2</sup>, Emma S McBryde <sup>3</sup>, Luke Vasanthakumar <sup>1</sup>, Matthew Murray<sup>4</sup>, Miriam Harings <sup>1</sup>, Oyelola Adegboye <sup>3\*</sup>

<sup>1</sup> The Townsville Hospital, 100 Angus Smith Drive, Douglas, Queensland, Australia, 4814

<sup>2</sup> College of Medicine and Dentistry, James Cook University, 1 James Cook Drive, Douglas, Queensland, Australia, 4814

<sup>3</sup> Australian Institute of Tropical Health and Medicine, James Cook University, 1 James Cook Drive, Douglas, Queensland, Australia, 4814

<sup>4</sup> Commonline Pty Ltd, Townsville, Queensland, Australia, 4810

## \*Corresponding author:

Oyelola Adegboye ([oyelola.adegboye@jcu.edu.au](mailto:oyelola.adegboye@jcu.edu.au))

Australian Institute of Tropical Health and Medicine, James Cook University, 1 James Cook Drive, Douglas, Queensland, Australia, 4814

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**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 ICD-10-Australian Modification code for an infectious disease between 1/1/2006 and 31/12/2016 was assembled. This utilised 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 identify as Aboriginal or Torres Strait Islanders. Invasive staphylococcal, pneumococcal and Group A streptococcal infections and influenza were common in this cohort. The commonest 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 eleven-year period was 20%.

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

**Key words:** data-linkage, relational database, epidemiology, infectious diseases, hospital

### 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 several infectious diseases cohort.
- Patients cohort was based on data sets from a single hospital, findings might not be generalizable to the Australian population.
- The validity of cohort studies rely on the accuracy of clinical coding, therefore some important clinical information may be underrepresented.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Funding statement:**

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

**Competing interests:**

None of the authors have any competing interests to declare.

**Word counts:**

Abstract: 250

Manuscript: 2905

For peer review only

## 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 1000 of Australia's population were hospitalised with a primary diagnosis of an infectious disease. <sup>1</sup> The rate in Australia's Indigenous population was double this. Of the principal causes of hospitalisation, pneumonia was 4<sup>th</sup>, cellulitis 9<sup>th</sup> and 'other sepsis' 16<sup>th</sup>. 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 3<sup>rd</sup> to 5<sup>th</sup> most common hospital acquired complications. These infections contribute to the marked increase in the average length of stay (17 vs 4.4 days) <sup>1</sup> and may increase mortality. <sup>2</sup> 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%). <sup>3</sup> 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. <sup>4</sup> 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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**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 1100mm.

Cohort selection.

A cohort of Townsville Hospital in-patients 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 the 11-year period January 1st 2006 to December 31st 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 A00–B99 Infectious and parasitic diseases (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.

#### Data bases.

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: 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.
- 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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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 was 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 9.4 software (SAS Institute Inc., Cary, NC, USA). Descriptive summaries are presented as frequencies and percentages for

1  
2  
3 categorical variables, and means, quartiles and standard deviations for continuous variables.  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Charlson Comorbidity Index (CCI) <sup>5 6</sup> was used to rank patient illness severity based on the number and importance of comorbid diseases (Supplementary Table S2).

#### 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).

#### 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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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 Supplementary Table S1. A summary of the data and the datafields is included in Supplementary Table S2. The individual datafields are listed in Supplementary Table S3. ICD-10-AM codes used to identify comorbidities are listed in 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 dataset.

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 efforts were made to extract the maximum detail from various data sources. Some sources using the

1  
2  
3 same PU\_ID recorded the age inconsistently at a certain admission date, others had birth  
4 month and day, and others incorporated full birth dates. The scripts analysed and prioritised  
5 each of these and consolidated all available information for each of 41,367 people. The year  
6 of birth was successfully generated for every person. Additionally, the ICD Codes were not  
7 consistently entered. For example "A064" was entered but the correct format is "A06.4".  
8 Each was analysed, broken down into its components and entered into the database.  
9

10  
11 Summary statistics are presented to give a basic description of the cohort. (Table 1) The  
12 distribution of age at first admission was skewed towards older subjects. Similarly, the total  
13 number of admissions was markedly skewed towards higher values. This is due to the  
14 significant number of haemodialysis patients who had a median of six admissions with  
15 interquartile range of 2-41 over the eleven-year duration of the cohort study. A large  
16 proportion of the patients identified as Indigenous (14.88%). Of interest, 4.5% of patients in  
17 this cohort were admitted to the Townsville Hospital from correctional facilities and  
18 Indigenous peoples are over represented amongst these patients compared with the cohort  
19 as a whole. The overall eleven-year all-cause mortality was 20%. A high proportion of patients  
20 smoked (44%). Other major modifiable risk factors included alcohol abuse, obesity and  
21 malnutrition (Table 1).  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45

46 This patient cohort had a moderately low burden of comorbidity with an average Charlson  
47 comorbidity index (CCI) score of 1.86 (interquartile range, 0-3). About 16% had a CCI of 5 and  
48 above. The major comorbidities are diabetes, cancer and renal disease. Other common  
49 comorbidities were; chronic pulmonary disease, cerebrovascular disease and myocardial  
50 infarction. Multiple comorbidities were present in 67% of patients (Table 2). .  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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 2 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 and well as in Australian communities in general.

When we consider the patterns of infectious diseases found in this cohort *S. aureus* was the commonest pathogen identified followed by influenza and Group A streptococcus. Skin and soft tissue was the commonest 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 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) <sup>7</sup>, community acquired pneumonia (60.4%, mean follow up 6.1 years) <sup>8</sup> and infective endocarditis (14.7%, 1 – 5 year follow up) have been described. <sup>9</sup> These studies all demonstrated increased all-cause mortality of the infectious diseases cohorts compared with controls.

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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.<sup>10</sup> There has been a sustained increase in the numbers of cohort studies using linked administrative hospital datasets including in Australia.<sup>11</sup> However, infectious diseases studies are in the minority compared with cardiovascular, health services, cancer and maternal health research. Australian cohort studies that utilise data linkage to describe infectious diseases mostly rely on ICD-10-AM diagnostic codes and death registry information. Some also incorporate notifiable diseases data<sup>12</sup> but, overall, studies incorporating pathology data are few<sup>13 14</sup>.

Regrettably, in Australian jurisdictions, pathology data are only available for data-linkage in Western Australia and Queensland due to their statewide diagnostic laboratories<sup>4</sup>. Data-linkage studies incorporating pathology data have tested the precision of infectious diseases diagnosis in comparison with public health communicable diseases notifications systems<sup>15</sup> and hospital discharge coding<sup>13</sup>. 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<sup>13</sup>, infections such as Q fever<sup>12</sup> and *S. aureus* bacteraemia<sup>14</sup> as well as specific patients such as asplenic<sup>16</sup> and haematology-oncology.<sup>17</sup> 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 aging<sup>18</sup>, Triple I Western Australian birth cohort<sup>15</sup> and Victorian Post-Splenectomy Registry<sup>19</sup>.

There are inherent limitations of retrospective, databases defined by ICD-10-AM codes. Some important clinical information is underrepresented. This is exemplified in this cohort study where only 3.95% of patients were coded as being obese. By contrast, among the general

1  
2  
3 Australian population, as measured in 2017-2018, 31% of adults and 8.6% of children and  
4  
5 adolescents were obese.<sup>20</sup> This in-patient underestimate may derive from ICD-10-AM coding  
6  
7 for obesity only being allocated where active assessment is made by a dietitian for obesity.  
8  
9  
10 Inpatients at the Townsville Hospital were more frequently diagnosed (11.13%) with  
11  
12 malnutrition reflecting documentation of clinical interventions. The administrative databases  
13  
14 used to construct this linked-database predated use of an electronic medical record at  
15  
16 Townsville Hospital. Machine learning is being used in research settings to analyse free text  
17  
18 in clinical notes and diagnostic imaging reports.<sup>21</sup> However, owing to absence of free text  
19  
20 data, we are unable to apply this methodology to our database. The absence of this clinical  
21  
22 information may diminish the ability to determine precise case definitions and important  
23  
24 comorbidities such as obesity.  
25  
26  
27  
28

29  
30 Despite these potential limitations, ICD-10-AM codes for infectious diseases have been shown  
31  
32 to be closely correlated with clinical diagnoses determined after medical chart review in  
33  
34 Australian research, for example in two studies of community acquired pneumonia<sup>22 23</sup>.  
35  
36 Linked administrative data was shown to reliably ascertain incident colorectal and lung cancer  
37  
38 diagnoses when compared with the New South Wales Cancer Registry<sup>24</sup>. Other Australian  
39  
40 researchers have studied the accuracy of ICD-10-AM codes for diagnoses of childhood  
41  
42 influenza and pertussis<sup>25</sup>. While demonstrating high specificity and positive predictive value,  
43  
44 the authors conclude that addition of laboratory data increases the precision of retrospective,  
45  
46 population level diagnosis of paediatric respiratory infection. The incorporation of pathology  
47  
48 results in the cohort described in this database will allow precise characterisation of the  
49  
50 infectious diseases cohort we have assembled. For example, the large volume of microbiology  
51  
52 data will allow for analysis of key areas such as antimicrobial resistant infections and their  
53  
54  
55  
56  
57  
58  
59  
60

influence on clinical outcomes and provide greater precision for diagnosis (e.g. site of infection in sepsis).

Conclusions

Numerous analysis of risks for, and outcomes of, disease and organism specific infections, health care 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 Areas <sup>26</sup> to assess the impact of socioeconomic disadvantage on outcomes of infectious diseases occurring in hospitalized patients. As hospitalization 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 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.

For peer review only

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Author’s contribution

DE conceived the study idea and defined the original study protocol. DE 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, categorization 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.

## References.

1. Burgess K, Gilbert M, McIntyre J, et al. Admitted patient care 2016-17: Australian hospital statistics. *Canberra: Australian Institute of Health and Welfare* 2018
2. 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(10):e003587. doi: 10.1136/bmjopen-2013-003587 [published Online First: 2013/11/02]
3. Welfare AloHa. Australian Burden of Disease Study: impact and causes of illness and death in Australia 2015. Cat. No. BOD 22 ed. Canberra, 2019.
4. Moore HC, Blyth CC. Optimising the use of linked administrative data for infectious diseases research in Australia. *Public Health Res Pract* 2018;28(2) doi: 10.17061/phrp2821810 [published Online First: 2018/06/21]
5. Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases* 1987;40(5):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. *Medical Care* 2005;1130-39.
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(4):346-57. doi: 10.1016/j.jinf.2016.07.005 [published Online First: 2016/07/16]
8. Myint PK, Hawkins KR, Clark AB, et al. Long-term mortality of hospitalized pneumonia in the EPIC-Norfolk cohort. *Epidemiol Infect* 2016;144(4):803-9. doi: 10.1017/S0950268815001971 [published Online First: 2015/08/25]
9. Ternhag A, Cederstrom A, Torner A, et al. A nationwide cohort study of mortality risk and long-term prognosis in infective endocarditis in Sweden. *PLoS One* 2013;8(7):e67519. doi: 10.1371/journal.pone.0067519 [published Online First: 2013/07/19]
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. *Journal of Exposure Science & Wnvironmental Epidemiology* 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(4):394-400. doi: 10.1071/AH16034 [published Online First: 2016/07/23]
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(11):438. doi: 10.5694/mja15.00391 [published Online First: 2015/12/15]

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(6):502-10. doi: 10.1111/irv.12508 [published Online First: 2017/10/11]
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(3):291-8. doi: 10.1086/669522 [published Online First: 2013/02/08]
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(1):86. doi: 10.1186/s12911-017-0484-7 [published Online First: 2017/06/19]
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(9):582-6. [published Online First: 2012/05/25]
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(1):274. doi: 10.1186/s12879-019-3901-y [published Online First: 2019/03/23]
18. Institute S. 45 and Up Study: Sax Institute; 2019 [Available from: <https://www.saxinstitute.org.au/our-work/45-up-study/> accessed 9 July 2019 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 NZ J Public Health* 2006;30(6):558-61. [published Online First: 2007/01/11]
20. Welfare AloHa. Overweight and obesity: an interactive insight Canberra, Australia: Australian Government; 2019 [updated 19/7/19. Available from: <https://www.aihw.gov.au/reports/overweight-obesity/overweight-and-obesity-an-interactive-insight/contents/prevalence> accessed 6/9/19 2019.
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(5):1007-15. doi: 10.1093/jamia/ocv180 [published Online First: 2016/02/26]
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(2):194-202. doi: 10.1017/S0950268808000812 [published Online First: 2008/06/19]
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(2):232-40. doi: 10.1017/S0950268807008564 [published Online First: 2007/04/21]

- 1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60
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(1):23. doi: 10.1186/s12911-017-0417-5 [published Online First: 2017/03/01]
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(3):255-60. doi: 10.1093/jpids/pit036 [published Online First: 2014/09/01]
26. Statistics ABo. Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2016 Canberra, Australia: Australian Government; 2018 [updated 27/3/2018. Available from: <https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/2033.0.55.001~2016~Main%20Features~SEIFA%20Basics~5> accessed 6/9/2019 2019.

Table 1: Cohort characteristics (n=41367)

Characteristics	Number	Mean	Median	SD	Q1	Q3
Age (years) at first admission	41367	43.15	49	24.44	26	68
Total admissions	37783	100.81	6	230.6	2	41
Demographics	Number of patients	Percentage				
Male	21299	51.49				
Female	20068	48.51				
11 year mortality						
Dead	8274	20				
Indigenous status						
Aboriginal but not TSI <sup>a</sup> origin	4763	11.51				
Aboriginal and TSI	550	1.33				
Neither Aboriginal nor TSI	35187	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	18179	43.95				
Alcohol	4743	11.47				
Recreational drug use	600	1.45				
Malnutrition	4605	11.13				
Obesity	1633	3.95				

<sup>a</sup> Aboriginal and Torres Strait Islander

Table 2. 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	
Meliodosis	84
Dengue	88
Ross River	48
Q fever	139

Table 3. 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)	21215	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)	

Figure 1. Representation of relational database constructed showing links between fields from incorporated administrative, clinical and death registry information. (See Supplementary Table S1 for detailed description of fields.)

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

For peer review only

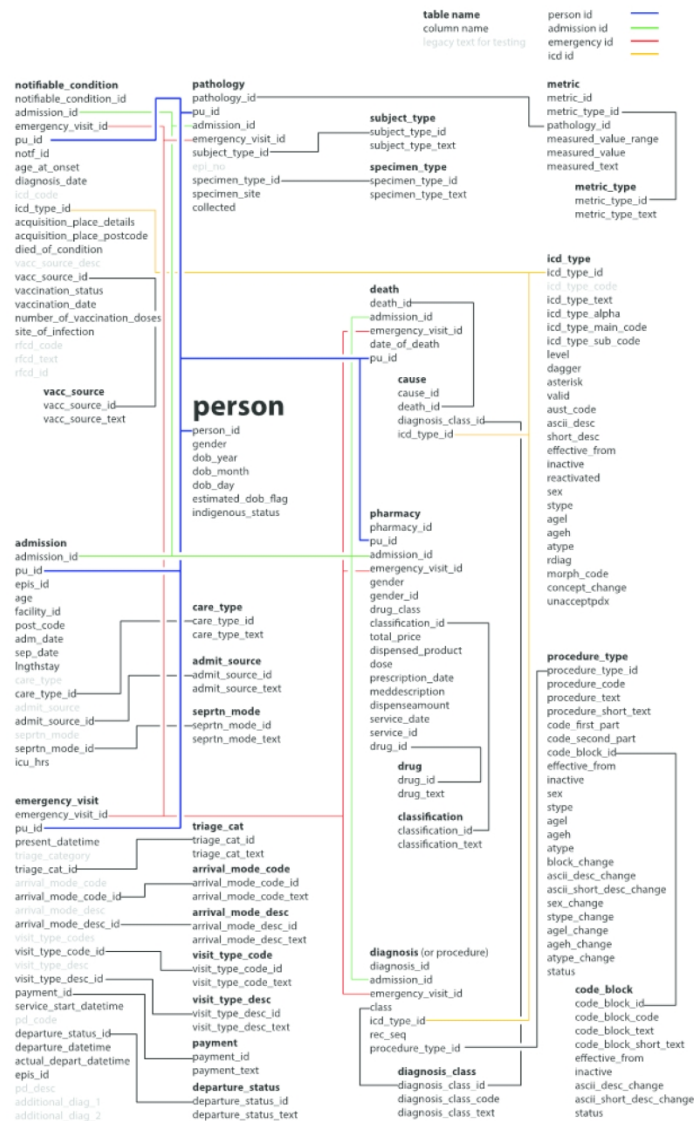


Figure 1. Representation of relational database constructed showing links between fields from incorporated administrative, clinical and death registry information. (See Supplementary materials for detailed description of fields.)

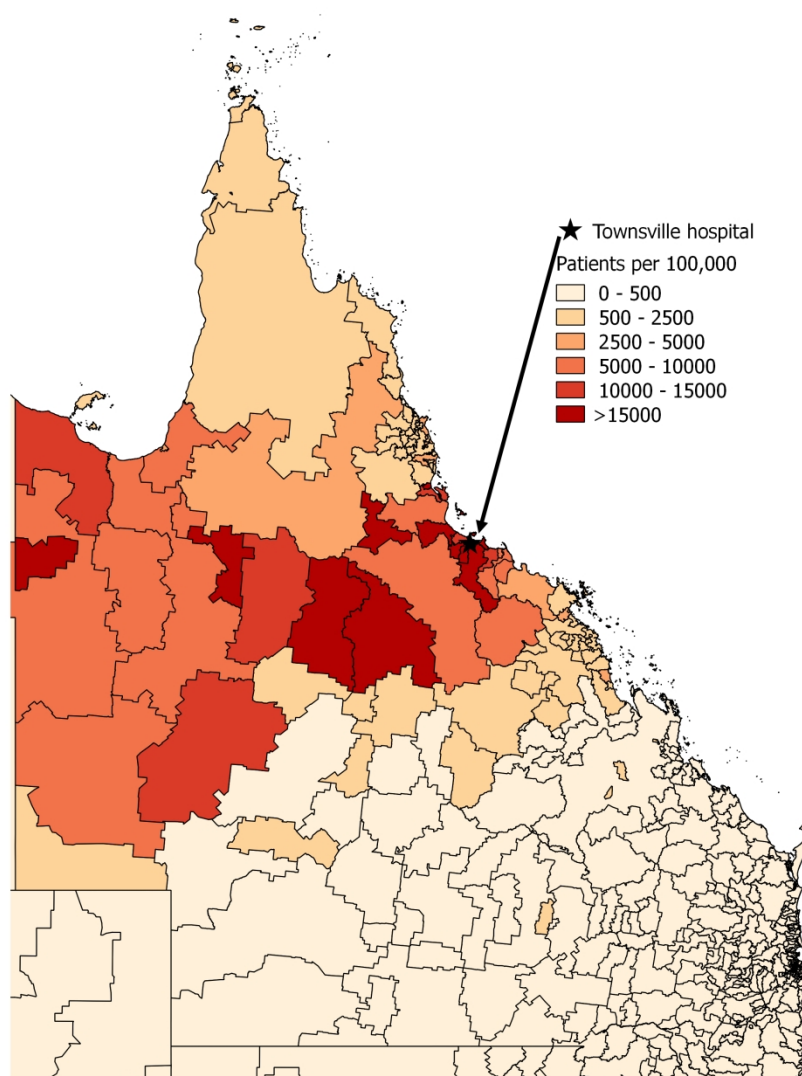


Figure 2. Heat map of cohort patients per 100,000 shown by postcode of domicile according to hospital registration at entry into cohort.

209x296mm (300 x 300 DPI)

Table S1. ICD-10-AM infectious disease codes used to select the patient cohort

	ICD10 codes
Infectious and parasitic diseases	A00–B99
Bacterial meningitis, not elsewhere classified	G00
Encephalitis, myelitis and encephalomyelitis	G04
Intracranial and intraspinal abscess and granuloma	G06
Intracranial and intraspinal phlebitis and thrombophlebitis	G08
Focal chorioretinal inflammation	H30.0
Purulent endophthalmitis	H44.0
Other endophthalmitis	H44.1
Disorders of vitreous body and globe	H45
Otitis Externa	H60
Mastoiditis	H70
Infective pericarditis	I30.1
Chronic constrictive pericarditis	I31.1
Pericarditis in diseases classified elsewhere	I30.0
Acute and subacute infective endocarditis	I33.0
Acute endocarditis unspecified	I33.9
Endocarditis and heart valve disorders in diseases classified elsewhere	I39
Infective myocarditis	I40.0
Myocarditis in diseases classified elsewhere	I41
Acute sinusitis	J01
Streptococcal pharyngitis	J02.0
Acute pharyngitis due to other specified organisms	J02.8
influenza and Pneumonia	J09-J18
Influenza due to certain identified influenza virus	J09
Influenza due to other identified influenza virus	J10
Influenza, virus not identified	J11
Viral pneumonia, not elsewhere classified	J12
Pneumonia due to Streptococcus pneumoniae	J13
Pneumonia due to Haemophilus influenzae	J14
Bacterial pneumonia, not elsewhere classified	J15
Pneumonia due to other infectious organisms, not elsewhere classified	J16
Pneumonia in diseases classified elsewhere	J17
Pneumonia, organism unspecified	J18
Peritonsillar abscess	J36
Retropharyngeal and parapharyngeal abscess	J39.0
Abscess of lung and mediastinum	J85
Pyothorax	J86.9
Acute peritonitis	K65.0
Disorders of peritoneum in infectious diseases classified elsewhere	K67
Abscess of liver	K75.0
Granulomatous hepatitis not elsewhere classified	K75.0
Liver disorders in infectious and parasitic diseases classified elsewhere	K77.0
Acute Cholecystitis	K81.0
Cellulitis	L03
Pyoderma	L08.0
Pyogenic arthritis	M00
Direct infections of joint in infectious and parasitic diseases classified elsewhere	M01
Post infective and reactive arthropathies in diseases classified elsewhere	M03
Necrotising Fasciitis	M72.6
Osteomyelitis	M86
UTI	N39.0
Inflammatory disease of the prostate	N41
Orchitis and epididymitis	N45.9
Fever of other and unknown origin	R50
Shock, not elsewhere classified	R57
Abnormal findings in cerebrospinal fluids	R83

Table S2. Sources and numbers of patient records from the cohort of 41,367 unique people, identified and extracted from Queensland Health Admitted Patient Data Collection (QHAPDC) and the Queensland Health Statistical Services Branch Master Linkage File, based on a selection of infectious disease codes for the period 1st January 2006 to 31st December 2016.

Table Name (Output Name)	Counts	Description
<b>QHAPDC_Single_Out</b> (QHAPDC_Single_Out_oct_18.csv)	records= 378,487 -- unique people= 41,367	QHAPDC Admissions data
<b>QHAPDC_Morb_Out</b> (QHAPDC_Morb_Out.csv)	records= 1,869,239 -- unique people= 41,367	QHAPDC Morbidity data
<b>EDIS</b> (EDIS_Out.csv)	records= 204,112 -- unique people= 35,316	EDIS data
<b>Death_Out</b> (Death_Out_oct_18.csv)	records= 8,274 -- unique people= 8,274	Death data
<b>VivNocs_out</b> (VivNocs_out.csv)	records= 10,501 -- unique people= 7,060	NOCs and VIVAS data
<b>Pathology data (AUSLAB)</b>		
<b>Microbiology_out</b> (Microbiology_out.csv)	records= 296,949 -- unique people= 35,736	Pathology data (AUSLAB)
<b>Serology_out</b> (Serology_out.csv)	records= 12,840 -- unique people= 6,852	Pathology data (AUSLAB)
<b>PCR_out</b> (PCR_out.csv)	records= 10,376 -- unique people= 6,309	Pathology data (AUSLAB)
<b>Haematology_out</b> (Haematology_out.csv)	records= 621,030 -- unique people= 38,406	Pathology data (AUSLAB)
<b>Biochemistry_out</b> (Biochemistry_out.csv)	records= 603,337 -- unique people= 38,110	Pathology data (AUSLAB)
<b>Remainder_out</b> (Remainder.csv)	records= 21,537 -- unique people= 10,229	Pathology data (AUSLAB)
<b>Drug Dispensing Data</b>		
<b>Pyxis_out</b> (Pyxis_out.csv)	records= 147,405 -- unique people= 10,404	Drug Dispensing Data
<b>iPharmacy</b> (iPharm_out.csv)	records= 115,505 -- unique people= 21,335	Drug Dispensing Data

Table S3. Queensland Hospital Admitted Patient Data Collection, Emergency Department Information System, Death Registry and Notifiable Conditions fields included in the dataset.

Queensland Hospital Admitted Patient Data Collection (QHAPDC)		Emergency Department Information System		Death Registry	
Variable Name	Variable Description	Variable Name	Variable description	Variable Name	Variable Description
PU_ID	Person level identifier	PU_ID	Person level identifier	PU_ID	Person level identifier
EPIS_ID	Episode level identifier	PAT_SEX	Sex	Date_Of_Death	Person date of death (dd-mm-yyyy)
AGE	Age of the person at time of admission	PAT_DATE_OF_BIRTH	Date of birth	CAUSE	Cause of death (free text)
POST_CODE	Post code at the time of admission	PAT_DOB_EST_FLAG	Estimated DOB Flag	COD	Underlying cause of death (ICD-10)
ADM_DATE	Full date of the persons admission	PRESENTATION_DATETIME	Present datetime (dd.mmm.yyyy.hh.mm)	EC_AXIS_1 - 14	Other cause of death (ICD-10) 1 - 14
SEP_DATE	Full date of the persons discharge	TRIAGE_CATEGORY	Triage category		
SEX	Sex of the person relating to the admission	TRIAGE_DATETIME	Triage date time		
INDIG_STATUS	Indigenous status of the person	ARRIVAL_TRANSPORT_MODE	Arrival mode description		
LNGTHSTAY	Length the person was in hospital excluding periods of leave	VISIT_TYPE_CODE	Visit type codes		
CARE_TYPE	The nature of the treatment/care provided to a patient during an episode of care	VISIT_TYPE	Visit type description		
ADMIT_SOURCE	The source of referral/transfer (admission source) of a patient immediately before they are admitted	PAYMENT_CLASS	Payment ttatus		
SEPRTN_MODE	The mode of separation: place to which a patient is referred immediately following separation	SERVICE_START_DATETIME	Service start datetime		
ICU_HRS	Total number of hours and minutes a patient has spent in ICU	PRINCIPAL_DIAGNOSIS	PD Code		
ICD_TYPE	Principal Diagnoses (PD), Other/Additional Diagnoses (OD), Procedure Code (PR)	PRINCIPAL_DIAGNOSIS	PD Description		
ICD_CODE	ICD-10-AM and ACHI classifications of diagnoses and procedures	ADDITIONAL_DIAGNOSIS_1	Additional Diagnosis 1		
		ADDITIONAL_DIAGNOSIS_2	Additional Diagnosis 2		
		EPISODE_END_STATUS_CODE	Departure status		
		EPISODE_END_DATETIME	Departure datetime		
		PHYSICAL_DEPART_DATETIME	Actual departure datetime		

Table S3. AUSLAB microbiological results included in dataset

Patient identifier	Patient identifier
Specimen	Specimen
Specimen Site	Specimen Site
Collection time (dd-mmm-yyyy hh-mm)	Collection time (dd-mmm-yyyy hh-mm)
Blood Culture	Parainfluenza 1 (NAA)
Incubation Time Until Positive	
Positive Bottles	Parainfluenza 2 (NAA)
Culture Comment	Parainfluenza 3 (NAA)
Organism	Influenza A RNA (NAA)
Fungal Culture	Resp Syn Virus (NAA)
Mycobacterial Culture Result	Adenovirus DNA (NAA)
Organism	Influenza B RNA (RNA)
Sensitivities	HSV 1 DNA (NAA)
	HSV 2 DNA (NAA)
Q Fever Ph2 IgG (EIA)	Human Metapneumovirus RNA (NAA)
Q Fever Ph2 IgM (EIA)	Human herpes 6 DNA (NAA)
Q Fever Ph1 IgG (IF)	Human herpes 7 DNA (NAA)
Q Fever Ph2 IgG (IF)	Human herpes 8 DNA (NAA)
Q Fever Ph2 IgM (IF)	Varicella zoster DNA (NAA)
Q Fever DNA	Enterovirus RNA (NAA)
Leptospira sp. IgM (EIA)	Dengue group RNA (NAA)
Leptospirosis NAA	Denque Universal (TAQ)
Atypical Serology	Barmah Forest RNA (TAQ)
M. pneumoniae Total Ab	Ross River Virus RNA (TAQ)
L. pneumophila 1 antigen	J. encephalitis RNA (TAQ)
S. pneumoniae antigen	Kunjin Virus (TAQ)
CMV DNA	Murray Valley RNA (TAQ)
CMV IgM (EIA)	Chikungunya RNA (TAQ)
CMV IgM (EIA)	West Nile Virus (TAQ)
EBV IgG (EIA)	Rift Valley Fever Virus (TAQ)
EBV IgM (EIA)	P. jiroveci DNA (NAA)
Epstein-Barr Virus IgA	N. meningitidis DNA (NAA)
EBV DNA	S. pneumoniae DNA (NAA)
Dengue NS1 antigen	Non tuberculous mycobacteria PCR
Cryptococcal antigen	M. ulcerans PCR
Aspergillus galactomannan antigen	TB PCR
	M. leprae PCR

Table S3. AUSLAB pathology results and Pharmacy drug dispensing information included in the dataset

Biochemistry	Haematology	Other	Pharmacy
Patient identifier	Patient identifier	Patient identifier	iPharm
Specimen	Specimen	Specimen	
Specimen Site	Specimen Site	Specimen Site	
Collection time (dd-mmm-yyyy hh-mm)	Collection time (dd-mmm-yyyy hh-mm)	Collection time (dd-mmm-yyyy hh-mm)	Patient identifier
Sodium	Haemoglobin	HbA1c	Gender
Potassium	White Cell Count	25-Hydroxy-Vitamin D	Drug class
Chloride	Platelets	1,25-Dihydroxy-Vitamin D	Classification
Bicarbonate	Haematocrit	IGG	Total price
Urea/Creat. Ratio	Mean corpuscular haemaglobin	IGG1	Dispensed product
Glucose	Red Cell Count	IGG2	Dose
Calcium	Mean corpuscular volume	IGG3	Prescription date
Magnesium	Neutrophils	IGG4	
Phosphate	Lymphocytes	IGA	Pyxis
Urate	Monocytes	IGM	
Protein	Eosinophils	Antinuclear antibody	
Albumin	Basophils	Rheumatoid factor	Patient identifier
Bilirubin	Prothrombin Time	Rheumatoid factor (fluid)	Medication description
Bilirubin (conjugated)	APTT	Total protein	Dispensed amount
Alkaline phosphatase		Albumin	Service date
Gamma GT		Total globulin	
Aspartate transaminase		Monoclonal protein	
Alanine transaminase		Serum EPP comment	
Lactate dehydrogenase		Angiotensin converting enzyme	
Anion gap		ACE (CSF)	
Osmolality (calculated)			
Urea/Creat. Ratio			
Globulin			
Corrected calcium			
Corrected potassium			
estimated glomerular filtration rate			
C-reactive protein			

Table S4. ICD-10 codes used by compute Charlson comorbidity score.

Description	ICD-10 codes	Charlson score
Myocardial infarction	I21.x, I22.x, I25.2	1
Peripheral vascular disease	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9	1
Congestive heart failure	I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5 - I42.9, I43.x, I50.x, P29.0	1
Cerebrovascular disease	G45.x, G46.x, H34.0, I60.x - I69.x	1
Dementia	F00.x - F03.x, F05.1, G30.x, G31.1	1
Chronic pulmonary disease	I27.8, I27.9, J40.x - J47.x, J60.x - J67.x, J68.4, J70.1, J70.3	1
Rheumatic disease	M05.x, M06.x, M31.5, M32.x - M34.x, M35.1, M35.3, M36.0	1
Peptic ulcer disease	K25.x - K28.x	1
Mild liver disease	B18.x, K70.0 - K70.3, K70.9, K71.3 - K71.5, K71.7, K73.x, K74.x, K76.0, K76.2 - K76.4, K76.8, K76.9, Z94.4	1
Diabetes without chronic complication	E10.0, E10.1, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, E12.6, E12.8, E12.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9	1
Diabetes with chronic complication	E10.2 - E10.5, E10.7, E11.2 - E11.5, E11.7, E12.2 - E12.5, E12.7, E13.2 - E13.5, E13.7, E14.2 - E14.5, E14.7	2
Hemiplegia or paraplegia	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0 - G83.4, G83.9	2
Renal disease	I12.0, I13.1, N03.2 - N03.7, N05.2 - N05.7, N18.x, N19.x, N25.0, Z49.0 - Z49.2, Z94.0, Z99.2	2
Any malignancy, including lymphoma and leukaemia, except malignant neoplasm of skin	C00.x - C26.x, C30.x - C34.x, C37.x - C41.x, C43.x, C45.x - C58.x, C60.x - C76.x, C81.x - C85.x, C88.x, C90.x - C97.x	2
Moderate or severe liver disease	I85.0, I85.9, I86.4, I98.2, K70.4, K71.1, K72.1, K72.9, K76.5, K76.6, K76.7	3
Metastatic solid tumour	C77.x - C80.x	6
AIDS/HIV	B20.x - B22.x, B24.x	6

# BMJ Open

## Linking administrative datasets of inpatient infectious diseases diagnoses in Far North Queensland: A cohort profile

Journal:	<i>BMJ Open</i>
Manuscript ID	bmjopen-2019-034845.R2
Article Type:	Cohort profile
Date Submitted by the Author:	31-Jan-2020
Complete List of Authors:	eisen, damon; Townsville Hospital; James Cook University Division of Tropical Health and Medicine Mcbryde, Emma; James Cook University Division of Tropical Health and Medicine, Australian Institute of Tropical Health and Medicine Vasanthakumar, luke; Townsville Hospital Murray, Matthew; Commonline Pty Ltd Harings, Miriam; Townsville Hospital ADEGBOYE, Oyelola; James Cook University Division of Tropical Health and Medicine, Australian Institute of Tropical Health & Medicine
<b>Primary Subject Heading</b>:	Infectious diseases
Secondary Subject Heading:	Epidemiology, Public health, Research methods
Keywords:	INFECTIOUS DISEASES, Epidemiology < INFECTIOUS DISEASES, Tropical medicine < INFECTIOUS DISEASES

SCHOLARONE™  
Manuscripts



I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our [licence](#).

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which [Creative Commons](#) licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

# Linking administrative datasets of inpatient infectious diseases diagnoses in Far North Queensland: A cohort profile

Damon P Eisen <sup>1,2</sup>, Emma S McBryde <sup>3</sup>, Luke Vasanthakumar <sup>1</sup>, Matthew Murray<sup>4</sup>, Miriam Harings <sup>1</sup>,  
Oyelola Adegboye <sup>3\*</sup>

<sup>1</sup> The Townsville Hospital, 100 Angus Smith Drive, Douglas, Queensland, Australia, 4814

<sup>2</sup> College of Medicine and Dentistry, James Cook University, 1 James Cook Drive, Douglas, Queensland,  
Australia, 4814

<sup>3</sup> Australian Institute of Tropical Health and Medicine, James Cook University, 1 James Cook Drive,  
Douglas, Queensland, Australia, 4814

<sup>4</sup> Commonline Pty Ltd, Townsville, Queensland, Australia, 4810

## \*Corresponding author:

Oyelola Adegboye ([oyelola.adegboye@jcu.edu.au](mailto:oyelola.adegboye@jcu.edu.au))

Australian Institute of Tropical Health and Medicine, James Cook University, 1 James Cook Drive,  
Douglas, Queensland, Australia, 4814

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**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 ICD-10-Australian Modification code for an infectious disease between 1/1/2006 and 31/12/2016 was assembled. This utilised 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 identify as Aboriginal or Torres Strait Islanders. Invasive staphylococcal, pneumococcal and Group A streptococcal infections and influenza were common in this cohort. The commonest 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 eleven-year period was 20%.

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

Key words: data-linkage, relational database, epidemiology, infectious diseases, hospital

**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 several infectious diseases cohort.
- Patients cohort was based on data sets from a single hospital, findings might not be generalizable to the Australian population.
- The validity of cohort studies rely on the accuracy of clinical coding, therefore some important clinical information may be underrepresented.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Funding statement:**

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

**Competing interests:**

None of the authors have any competing interests to declare.

**Word counts:**

Abstract: 250

Manuscript: 2905

For peer review only

## 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 1000 of Australia's population were hospitalised with a primary diagnosis of an infectious disease. <sup>1</sup> The rate in Australia's Indigenous population was double this. Of the principal causes of hospitalisation, pneumonia was 4<sup>th</sup>, cellulitis 9<sup>th</sup> and 'other sepsis' 16<sup>th</sup>. 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 3<sup>rd</sup> to 5<sup>th</sup> most common hospital acquired complications. These infections contribute to the marked increase in the average length of stay (17 vs 4.4 days) <sup>1</sup> and may increase mortality. <sup>2</sup> 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%). <sup>3</sup> 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. <sup>4</sup> 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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**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 1100mm.

Cohort selection.

A cohort of Townsville Hospital in-patients 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 the 11-year period January 1st 2006 to December 31st 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 A00–B99 Infectious and parasitic diseases (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.

## Data bases.

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Queensland Data Linkage Framework. Anonymised data, identified by a unique Created Study ID, were provided to the research team.

Database construction.

The data was 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 9.4 software (SAS Institute Inc., Cary, NC, USA). Descriptive summaries are presented as frequencies and percentages for categorical variables, and means, quartiles and standard deviations for continuous variables. Charlson Comorbidity Index (CCI) <sup>5 6</sup> was used to rank patient illness severity based on the number and importance of comorbid diseases (Supplementary Table S2).

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).

1  
2  
3 Patient and public involvement statement  
4  
5

6 Patients or members of the public were not involved in the development and design of the research.  
7

8 The anonymised data extraction does not require patient recruitment.  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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 Supplementary Table S1. A summary of the data and the datafields is included in Supplementary Table S2. The individual datafields are listed in Supplementary Table S3. ICD-10-AM codes used to identify comorbidities are listed in 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 dataset.

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 interquartile range of 2-41 over the eleven-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 over represented amongst these patients compared with the cohort as a whole. The overall eleven-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 Charlson comorbidity index (CCI) score of 1.86 (interquartile range, 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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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.

For peer review only

## 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 and well as in Australian communities in general.

When we consider the patterns of infectious diseases found in this cohort *S. aureus* was the commonest pathogen identified followed by influenza and Group A streptococcus. Skin and soft tissue was the commonest 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 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) <sup>7</sup>, community acquired pneumonia (60.4%, mean follow up 6.1 years) <sup>8</sup> and infective endocarditis (14.7%, 1 – 5 year follow up) have been described. <sup>9</sup> These studies all demonstrated increased all-cause mortality of the infectious diseases cohorts compared with controls.

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

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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.<sup>10</sup>

There has been a sustained increase in the numbers of cohort studies using linked administrative hospital datasets including in Australia.<sup>11</sup> However, infectious diseases studies are in the minority compared with cardiovascular, health services, cancer and maternal health research. Australian cohort studies that utilise data linkage to describe infectious diseases mostly rely on ICD-10-AM diagnostic codes and death registry information. Some also incorporate notifiable diseases data<sup>12</sup> but, overall, studies incorporating pathology data are few<sup>13 14</sup>.

Regrettably, in Australian jurisdictions, pathology data are only available for data-linkage in Western Australia and Queensland due to their statewide diagnostic laboratories<sup>4</sup>. Data-linkage studies incorporating pathology data have tested the precision of infectious diseases diagnosis in comparison with public health communicable diseases notifications systems<sup>15</sup> and hospital discharge coding<sup>13</sup>. 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<sup>13</sup>, infections such as Q fever<sup>12</sup> and *S. aureus* bacteraemia<sup>14</sup> as well as specific patients such as asplenic<sup>16</sup> and haematology-oncology.<sup>17</sup> 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 aging<sup>18</sup>, Triple I Western Australian birth cohort<sup>15</sup> and Victorian Post-Splenectomy Registry<sup>19</sup>.

There are inherent limitations of retrospective, databases defined by ICD-10-AM codes. Some important clinical information is underrepresented. This is exemplified in this cohort study where only 3.95% of patients were coded as being obese. By contrast, among the general

1  
2  
3 Australian population, as measured in 2017-2018, 31% of adults and 8.6% of children and  
4  
5 adolescents were obese.<sup>20</sup> This in-patient underestimate may derive from ICD-10-AM coding  
6  
7 for obesity only being allocated where active assessment is made by a dietitian for obesity.  
8  
9  
10 Inpatients at the Townsville Hospital were more frequently diagnosed (11.13%) with  
11  
12 malnutrition reflecting documentation of clinical interventions. The administrative databases  
13  
14 used to construct this linked-database predated use of an electronic medical record at  
15  
16 Townsville Hospital. Machine learning is being used in research settings to analyse free text  
17  
18 in clinical notes and diagnostic imaging reports.<sup>21</sup> However, owing to absence of free text  
19  
20 data, we are unable to apply this methodology to our database. The absence of this clinical  
21  
22 information may diminish the ability to determine precise case definitions and important  
23  
24 comorbidities such as obesity.  
25  
26  
27  
28  
29

30 Despite these potential limitations, ICD-10-AM codes for infectious diseases have been shown  
31  
32 to be closely correlated with clinical diagnoses determined after medical chart review in  
33  
34 Australian research, for example in two studies of community acquired pneumonia<sup>22 23</sup>.  
35  
36 Linked administrative data was shown to reliably ascertain incident colorectal and lung cancer  
37  
38 diagnoses when compared with the New South Wales Cancer Registry<sup>24</sup>. Other Australian  
39  
40 researchers have studied the accuracy of ICD-10-AM codes for diagnoses of childhood  
41  
42 influenza and pertussis<sup>25</sup>. While demonstrating high specificity and positive predictive value,  
43  
44 the authors conclude that addition of laboratory data increases the precision of retrospective,  
45  
46 population level diagnosis of paediatric respiratory infection. The incorporation of pathology  
47  
48 results in the cohort described in this database will allow precise characterisation of the  
49  
50 infectious diseases cohort we have assembled. For example, the large volume of microbiology  
51  
52 data will allow for analysis of key areas such as antimicrobial resistant infections and their  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

influence on clinical outcomes and provide greater precision for diagnosis (e.g. site of infection in sepsis).

Conclusions

Numerous analysis of risks for, and outcomes of, disease and organism specific infections, health care 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 Areas <sup>26</sup> to assess the impact of socioeconomic disadvantage on outcomes of infectious diseases occurring in hospitalized patients. As hospitalization 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 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.

For peer review only

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Author’s contribution

DE conceived the study idea and defined the original study protocol. DE 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, categorization 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.

Data availability statement

Due to restrictions and confidentiality, the datasets generated during and/or analysed during this study are not publicly available and may be obtained from a third party.

## References.

1. Burgess K, Gilbert M, McIntyre J, et al. Admitted patient care 2016-17: Australian hospital statistics. *Canberra: Australian Institute of Health and Welfare* 2018
2. 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(10):e003587. doi: 10.1136/bmjopen-2013-003587 [published Online First: 2013/11/02]
3. Welfare AloHa. Australian Burden of Disease Study: impact and causes of illness and death in Australia 2015. Cat. No. BOD 22 ed. Canberra, 2019.
4. Moore HC, Blyth CC. Optimising the use of linked administrative data for infectious diseases research in Australia. *Public Health Res Pract* 2018;28(2) doi: 10.17061/phrp2821810 [published Online First: 2018/06/21]
5. Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases* 1987;40(5):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. *Medical Care* 2005;1130-39.
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(4):346-57. doi: 10.1016/j.jinf.2016.07.005 [published Online First: 2016/07/16]
8. Myint PK, Hawkins KR, Clark AB, et al. Long-term mortality of hospitalized pneumonia in the EPIC-Norfolk cohort. *Epidemiol Infect* 2016;144(4):803-9. doi: 10.1017/S0950268815001971 [published Online First: 2015/08/25]
9. Ternhag A, Cederstrom A, Torner A, et al. A nationwide cohort study of mortality risk and long-term prognosis in infective endocarditis in Sweden. *PLoS One* 2013;8(7):e67519. doi: 10.1371/journal.pone.0067519 [published Online First: 2013/07/19]
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. *Journal of Exposure Science & Wnvironmental Epidemiology* 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(4):394-400. doi: 10.1071/AH16034 [published Online First: 2016/07/23]
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(11):438. doi: 10.5694/mja15.00391 [published Online First: 2015/12/15]
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(6):502-10. doi: 10.1111/irv.12508 [published Online First: 2017/10/11]

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(3):291-8. doi: 10.1086/669522 [published Online First: 2013/02/08]
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(1):86. doi: 10.1186/s12911-017-0484-7 [published Online First: 2017/06/19]
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(9):582-6. [published Online First: 2012/05/25]
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(1):274. doi: 10.1186/s12879-019-3901-y [published Online First: 2019/03/23]
18. Institute S. 45 and Up Study: Sax Institute; 2019 [Available from: <https://www.saxinstitute.org.au/our-work/45-up-study/> accessed 9 July 2019 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 NZ J Public Health* 2006;30(6):558-61. [published Online First: 2007/01/11]
20. Welfare AloHa. Overweight and obesity: an interactive insight Canberra, Australia: Australian Government; 2019 [updated 19/7/19. Available from: <https://www.aihw.gov.au/reports/overweight-obesity/overweight-and-obesity-an-interactive-insight/contents/prevalence> accessed 6/9/19 2019.
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(5):1007-15. doi: 10.1093/jamia/ocv180 [published Online First: 2016/02/26]
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(2):194-202. doi: 10.1017/S0950268808000812 [published Online First: 2008/06/19]
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(2):232-40. doi: 10.1017/S0950268807008564 [published Online First: 2007/04/21]
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(1):23. doi: 10.1186/s12911-017-0417-5 [published Online First: 2017/03/01]
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(3):255-60. doi: 10.1093/jpids/pit036 [published Online First: 2014/09/01]
26. Statistics ABo. Census of Population and Housing: Socio-Economic Indexes for Areas (SEIFA), Australia, 2016 Canberra, Australia: Australian Government; 2018 [updated 27/3/2018. Available from:

<https://www.abs.gov.au/ausstats/abs@.nsf/Lookup/by%20Subject/2033.0.55.001~2016~Main%20Features~SEIFA%20Basics~5> accessed 6/9/2019 2019.

For peer review only

Table 1: Cohort characteristics (n=41367)

Characteristics	Number	Mean	Median	SD	Q1	Q3
Age (years) at first admission	41367	43.15	49	24.44	26	68
Total admissions	378487	9.15	2	60.74	1	5
Demographics	Number of patients	Percentage				
Male	21299	51.49				
Female	20068	48.51				
11 year mortality						
Dead	8274	20				
Indigenous status						
Aboriginal but not TSI <sup>a</sup> origin	4763	11.51				
Aboriginal and TSI	550	1.33				
Neither Aboriginal nor TSI	35187	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	18179	43.95				
Alcohol	4743	11.47				
Recreational drug use	600	1.45				
Malnutrition	4605	11.13				
Obesity	1633	3.95				

<sup>a</sup> Aboriginal and Torres Strait Islander

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)	21215	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)	

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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	
Meliodosis	84
Dengue	88
Ross River	48
Q fever	139

Figure 1. Representation of relational database constructed showing links between fields from incorporated administrative, clinical and death registry information. (See Supplementary Table S1 for detailed description of fields.)

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

For peer review only

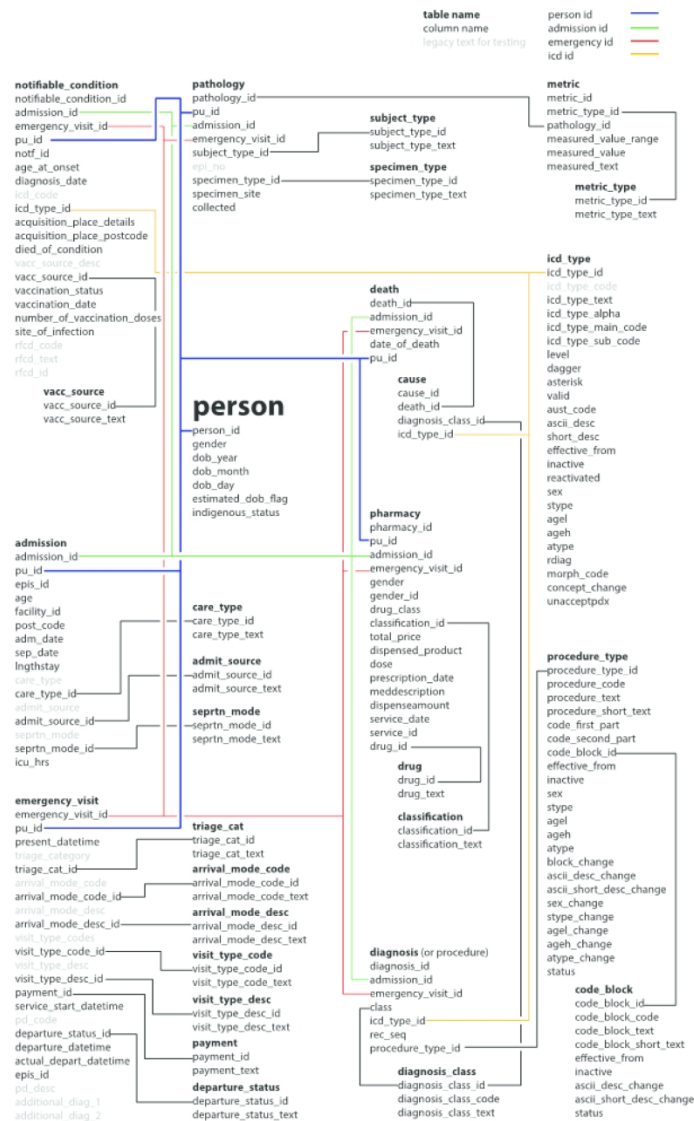


Figure 1. Representation of relational database constructed showing links between fields from incorporated administrative, clinical and death registry information. (See Supplementary materials for detailed description of fields.)

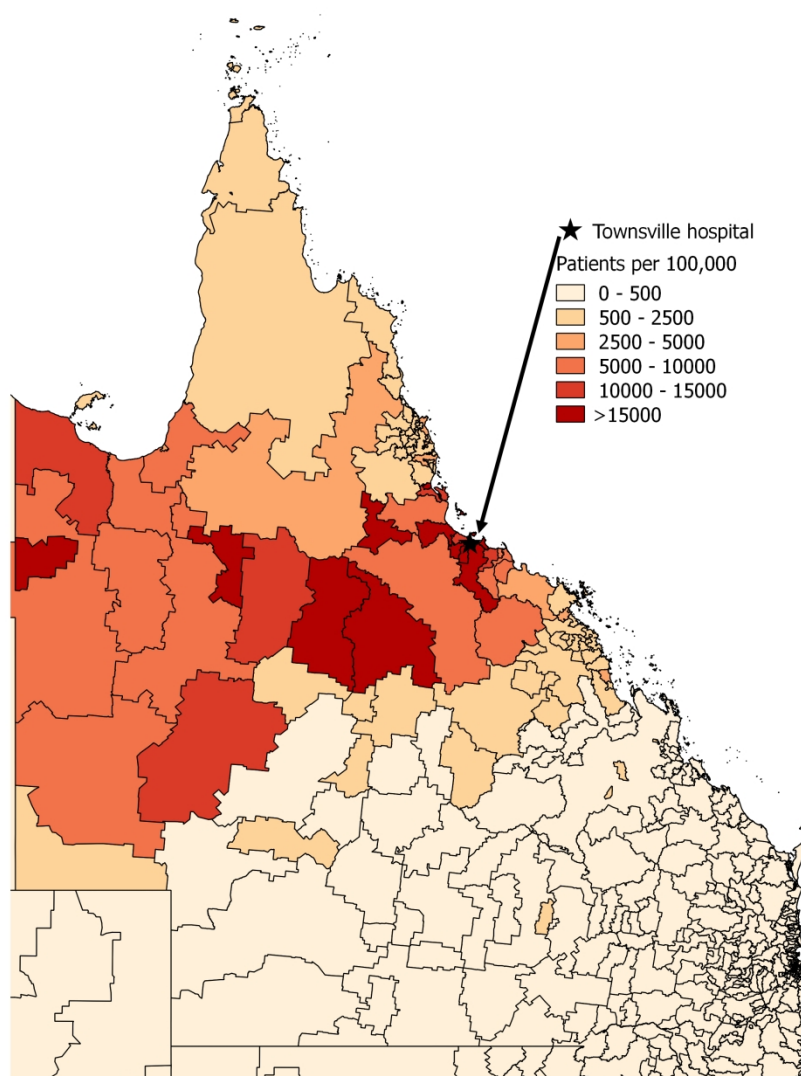


Figure 2. Heat map of cohort patients per 100,000 shown by postcode of domicile according to hospital registration at entry into cohort.

209x296mm (300 x 300 DPI)

Table S1. ICD-10-AM infectious disease codes used to select the patient cohort

	ICD10 codes
Infectious and parasitic diseases	A00–B99
Bacterial meningitis, not elsewhere classified	G00
Encephalitis, myelitis and encephalomyelitis	G04
Intracranial and intraspinal abscess and granuloma	G06
Intracranial and intraspinal phlebitis and thrombophlebitis	G08
Focal chorioretinal inflammation	H30.0
Purulent endophthalmitis	H44.0
Other endophthalmitis	H44.1
Disorders of vitreous body and globe	H45
Otitis Externa	H60
Mastoiditis	H70
Infective pericarditis	I30.1
Chronic constrictive pericarditis	I31.1
Pericarditis in diseases classified elsewhere	I30.0
Acute and subacute infective endocarditis	I33.0
Acute endocarditis unspecified	I33.9
Endocarditis and heart valve disorders in diseases classified elsewhere	I39
Infective myocarditis	I40.0
Myocarditis in diseases classified elsewhere	I41
Acute sinusitis	J01
Streptococcal pharyngitis	J02.0
Acute pharyngitis due to other specified organisms	J02.8
influenza and Pneumonia	J09-J18
Influenza due to certain identified influenza virus	J09
Influenza due to other identified influenza virus	J10
Influenza, virus not identified	J11
Viral pneumonia, not elsewhere classified	J12
Pneumonia due to Streptococcus pneumoniae	J13
Pneumonia due to Haemophilus influenzae	J14
Bacterial pneumonia, not elsewhere classified	J15
Pneumonia due to other infectious organisms, not elsewhere classified	J16
Pneumonia in diseases classified elsewhere	J17
Pneumonia, organism unspecified	J18
Peritonsillar abscess	J36
Retropharyngeal and parapharyngeal abscess	J39.0
Abscess of lung and mediastinum	J85
Pyothorax	J86.9
Acute peritonitis	K65.0
Disorders of peritoneum in infectious diseases classified elsewhere	K67
Abscess of liver	K75.0
Granulomatous hepatitis not elsewhere classified	K75.0
Liver disorders in infectious and parasitic diseases classified elsewhere	K77.0
Acute Cholecystitis	K81.0
Cellulitis	L03
Pyoderma	L08.0
Pyogenic arthritis	M00
Direct infections of joint in infectious and parasitic diseases classified elsewhere	M01
Post infective and reactive arthropathies in diseases classified elsewhere	M03
Necrotising Fasciitis	M72.6
Osteomyelitis	M86
UTI	N39.0
Inflammatory disease of the prostate	N41
Orchitis and epididymitis	N45.9
Fever of other and unknown origin	R50
Shock, not elsewhere classified	R57
Abnormal findings in cerebrospinal fluids	R83

Table S2. Sources and numbers of patient records from the cohort of 41,367 unique people, identified and extracted from Queensland Health Admitted Patient Data Collection (QHAPDC) and the Queensland Health Statistical Services Branch Master Linkage File, based on a selection of infectious disease codes for the period 1st January 2006 to 31st December 2016.

Table Name (Output Name)	Counts	Description
<b>QHAPDC_Single_Out</b> (QHAPDC_Single_Out_oct_18.csv)	records= 378,487 -- unique people= 41,367	QHAPDC Admissions data
<b>QHAPDC_Morb_Out</b> (QHAPDC_Morb_Out.csv)	records= 1,869,239 -- unique people= 41,367	QHAPDC Morbidity data
<b>EDIS</b> (EDIS_Out.csv)	records= 204,112 -- unique people= 35,316	EDIS data
<b>Death_Out</b> (Death_Out_oct_18.csv)	records= 8,274 -- unique people= 8,274	Death data
<b>VivNocs_out</b> (VivNocs_out.csv)	records= 10,501 -- unique people= 7,060	NOIS and VIVAS data
<b>Pathology data (AUSLAB)</b>		
<b>Microbiology_out</b> (Microbiology_out.csv)	records= 296,949 -- unique people= 35,736	Pathology data (AUSLAB)
<b>Serology_out</b> (Serology_out.csv)	records= 12,840 -- unique people= 6,852	Pathology data (AUSLAB)
<b>PCR_out</b> (PCR_out.csv)	records= 10,376 -- unique people= 6,309	Pathology data (AUSLAB)
<b>Haematology_out</b> (Haematology_out.csv)	records= 621,030 -- unique people= 38,406	Pathology data (AUSLAB)
<b>Biochemistry_out</b> (Biochemistry_out.csv)	records= 603,337 -- unique people= 38,110	Pathology data (AUSLAB)
<b>Remainder_out</b> (Remainder.csv)	records= 21,537 -- unique people= 10,229	Pathology data (AUSLAB)
<b>Drug Dispensing Data</b>		
<b>Pyxis_out</b> (Pyxis_out.csv)	records= 147,405 -- unique people= 10,404	Drug Dispensing Data
<b>iPharmacy</b> (iPharm_out.csv)	records= 115,505 -- unique people= 21,335	Drug Dispensing Data

Table S3. Queensland Hospital Admitted Patient Data Collection, Emergency Department Information System, Death Registry and Notifiable Conditions fields included in the dataset.

Queensland Hospital Admitted Patient Data Collection (QHAPDC)		Emergency Department Information System		Death Registry	
Variable Name	Variable Description	Variable Name	Variable description	Variable Name	Variable Description
PU_ID	Person level identifier	PU_ID	Person level identifier	PU_ID	Person level identifier
EPIS_ID	Episode level identifier	PAT_SEX	Sex	Date_Of_Death	Person date of death (dd-mm-yyyy)
AGE	Age of the person at time of admission	PAT_DATE_OF_BIRTH	Date of birth	CAUSE	Cause of death (free text)
POST_CODE	Post code at the time of admission	PAT_DOB_EST_FLAG	Estimated DOB Flag	COD	Underlying cause of death (ICD-10)
ADM_DATE	Full date of the persons admission	PRESENTATION_DATETIME	Present datetime (dd.mmm.yyyy.hh.mm)	EC_AXIS_1 - 14	Other cause of death (ICD-10) 1 - 14
SEP_DATE	Full date of the persons discharge	TRIAGE_CATEGORY	Triage category		
SEX	Sex of the person relating to the admission	TRIAGE_DATETIME	Triage date time		
INDIG_STATUS	Indigenous status of the person	ARRIVAL_TRANSPORT_MODE	Arrival mode description		
LNGTHSTAY	Length the person was in hospital excluding periods of leave	VISIT_TYPE_CODE	Visit type codes		
CARE_TYPE	The nature of the treatment/care provided to a patient during an episode of care	VISIT_TYPE	Visit type description		
ADMIT_SOURCE	The source of referral/transfer (admission source) of a patient immediately before they are admitted	PAYMENT_CLASS	Payment ttatus		
SEPRTN_MODE	The mode of separation: place to which a patient is referred immediately following separation	SERVICE_START_DATETIME	Service start datetime		
ICU_HRS	Total number of hours and minutes a patient has spent in ICU	PRINCIPAL_DIAGNOSIS	PD Code		
ICD_TYPE	Principal Diagnoses (PD), Other/Additional Diagnoses (OD), Procedure Code (PR)	PRINCIPAL_DIAGNOSIS	PD Description		
ICD_CODE	ICD-10-AM and ACHI classifications of diagnoses and procedures	ADDITIONAL_DIAGNOSIS_1	Additional Diagnosis 1		
		ADDITIONAL_DIAGNOSIS_2	Additional Diagnosis 2		
		EPISODE_END_STATUS_CODE	Departure status		
		EPISODE_END_DATETIME	Departure datetime		
		PHYSICAL_DEPART_DATETIME	Actual departure datetime		

Table S3. AUSLAB microbiological results included in dataset

Patient identifier	Patient identifier
Specimen	Specimen
Specimen Site	Specimen Site
Collection time (dd-mmm-yyyy hh-mm)	Collection time (dd-mmm-yyyy hh-mm)
Blood Culture	Parainfluenza 1 (NAA)
Incubation Time Until Positive	
Positive Bottles	Parainfluenza 2 (NAA)
Culture Comment	Parainfluenza 3 (NAA)
Organism	Influenza A RNA (NAA)
Fungal Culture	Resp Syn Virus (NAA)
Mycobacterial Culture Result	Adenovirus DNA (NAA)
Organism	Influenza B RNA (RNA)
Sensitivities	HSV 1 DNA (NAA)
	HSV 2 DNA (NAA)
Q Fever Ph2 IgG (EIA)	Human Metapneumovirus RNA (NAA)
Q Fever Ph2 IgM (EIA)	Human herpes 6 DNA (NAA)
Q Fever Ph1 IgG (IF)	Human herpes 7 DNA (NAA)
Q Fever Ph2 IgG (IF)	Human herpes 8 DNA (NAA)
Q Fever Ph2 IgM (IF)	Varicella zoster DNA (NAA)
Q Fever DNA	Enterovirus RNA (NAA)
Leptospira sp. IgM (EIA)	Dengue group RNA (NAA)
Leptospirosis NAA	Denque Universal (TAQ)
Atypical Serology	Barmah Forest RNA (TAQ)
M. pneumoniae Total Ab	Ross River Virus RNA (TAQ)
L. pneumophila 1 antigen	J. encephalitis RNA (TAQ)
S. pneumoniae antigen	Kunjin Virus (TAQ)
CMV DNA	Murray Valley RNA (TAQ)
CMV IgM (EIA)	Chikungunya RNA (TAQ)
CMV IgM (EIA)	West Nile Virus (TAQ)
EBV IgG (EIA)	Rift Valley Fever Virus (TAQ)
EBV IgM (EIA)	P. jiroveci DNA (NAA)
Epstein-Barr Virus IgA	N. meningitidis DNA (NAA)
EBV DNA	S. pneumoniae DNA (NAA)
Dengue NS1 antigen	Non tuberculous mycobacteria PCR
Cryptococcal antigen	M. ulcerans PCR
Aspergillus galactomannan antigen	TB PCR
	M. leprae PCR

6/bmjopen-2019-034845 on 18 March 2020. Downloaded from <http://bmjopen.bmj.com/> on April 8, 2024 by guest. Protected by copyright.

Table S3. AUSLAB pathology results and Pharmacy drug dispensing information included in the dataset

Biochemistry	Haematology	Other	Pharmacy
Patient identifier	Patient identifier	Patient identifier	iPharm
Specimen	Specimen	Specimen	
Specimen Site	Specimen Site	Specimen Site	
Collection time (dd-mmm-yyyy hh-mm)	Collection time (dd-mmm-yyyy hh-mm)	Collection time (dd-mmm-yyyy hh-mm)	Patient identifier
Sodium	Haemoglobin	HbA1c	Gender
Potassium	White Cell Count	25-Hydroxy-Vitamin D	Drug class
Chloride	Platelets	1,25-Dihydroxy-Vitamin D	Classification
Bicarbonate	Haematocrit	IGG	Total price
Urea/Creat. Ratio	Mean corpuscular haemaglobin	IGG1	Dispensed product
Glucose	Red Cell Count	IGG2	Dose
Calcium	Mean corpuscular volume	IGG3	Prescription date
Magnesium	Neutrophils	IGG4	
Phosphate	Lymphocytes	IGA	Pyxis
Urate	Monocytes	IGM	
Protein	Eosinophils	Antinuclear antibody	
Albumin	Basophils	Rheumatoid factor	Patient identifier
Bilirubin	Prothrombin Time	Rheumatoid factor (fluid)	Medication description
Bilirubin (conjugated)	APTT	Total protein	Dispensed amount
Alkaline phosphatase		Albumin	Service date
Gamma GT		Total globulin	
Aspartate transaminase		Monoclonal protein	
Alanine transaminase		Serum EPP comment	
Lactate dehydrogenase		Angiotensin converting enzyme	
Anion gap		ACE (CSF)	
Osmolality (calculated)			
Urea/Creat. Ratio			
Globulin			
Corrected calcium			
Corrected potassium			
estimated glomerular filtration rate			
C-reactive protein			

Table S4. ICD-10 codes used by compute Charlson comorbidity score.

Description	ICD-10 codes	Charlson score
Myocardial infarction	I21.x, I22.x, I25.2	1
Peripheral vascular disease	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9	1
Congestive heart failure	I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5 - I42.9, I43.x, I50.x, P29.0	1
Cerebrovascular disease	G45.x, G46.x, H34.0, I60.x - I69.x	1
Dementia	F00.x - F03.x, F05.1, G30.x, G31.1	1
Chronic pulmonary disease	I27.8, I27.9, J40.x - J47.x, J60.x - J67.x, J68.4, J70.1, J70.3	1
Rheumatic disease	M05.x, M06.x, M31.5, M32.x - M34.x, M35.1, M35.3, M36.0	1
Peptic ulcer disease	K25.x - K28.x	1
Mild liver disease	B18.x, K70.0 - K70.3, K70.9, K71.3 - K71.5, K71.7, K73.x, K74.x, K76.0, K76.2 - K76.4, K76.8, K76.9, Z94.4	1
Diabetes without chronic complication	E10.0, E10.1, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, E12.6, E12.8, E12.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9	1
Diabetes with chronic complication	E10.2 - E10.5, E10.7, E11.2 - E11.5, E11.7, E12.2 - E12.5, E12.7, E13.2 - E13.5, E13.7, E14.2 - E14.5, E14.7	2
Hemiplegia or paraplegia	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0 - G83.4, G83.9	2
Renal disease	I12.0, I13.1, N03.2 - N03.7, N05.2 - N05.7, N18.x, N19.x, N25.0, Z49.0 - Z49.2, Z94.0, Z99.2	2
Any malignancy, including lymphoma and leukaemia, except malignant neoplasm of skin	C00.x - C26.x, C30.x - C34.x, C37.x - C41.x, C43.x, C45.x - C58.x, C60.x - C76.x, C81.x - C85.x, C88.x, C90.x - C97.x	2
Moderate or severe liver disease	I85.0, I85.9, I86.4, I98.2, K70.4, K71.1, K72.1, K72.9, K76.5, K76.6, K76.7	3
Metastatic solid tumour	C77.x - C80.x	6
AIDS/HIV	B20.x - B22.x, B24.x	6