Visualising a State-wide Patient Data Collection: A Case Study to Expand the Audience for Healthcare Data

Wei Luo[1] **Marcus Gallagher**[1] **Di O'Kane**[2] **Jason Connor**[3] **Mark Dooris**[4]
Col Roberts[2] **Lachlan Mortimer**[2] **Janet Wiles**[1]

[1] School of Information Technology & Electrical Engineering
The University of Queensland
Brisbane, Australia
Email: {luo, marcusg, wiles}@itee.uq.edu.au

[2] Clinical Practice Improvement Centre
Centre for Healthcare Improvement
Queensland Health
Brisbane, Australia

[3] Department of Psychiatry
The University of Queensland
Brisbane, Australia

[4] Department of Cardiology
Royal Brisbane and Women's Hospital
Brisbane, Australia

Abstract

This paper describes the application of existing and novel adaptations of visualisation techniques to routinely collected health data. The aim of this case study is to examine the capacity for visualisation approaches to quickly and effectively inform clinical, policy, and fiscal decision making to improve healthcare provision. We demonstrate the use of interactive graphics, fluctuation plots, mosaic plots, time plots, heatmaps, and disease maps to visualise patient admission, transfer, in-hospital mortality, morbidity coding, execution of diagnosis and treatment guidelines, and the temporal and spatial variations of diseases. The relative effectiveness of these techniques and associated challenges are discussed.

Keywords: Visualisation, Exploratory Data Analysis, Routine Data Collection

1 Introduction

The state of Queensland has the third largest population in Australia [20]. In the financial year 2006-2007, public hospitals in Queensland treated more than 780,000 inpatients [11, page 11-12]. All such inpatient encounters are routinely collected in the Queensland Hospital Admitted Patient Data Collection (QHAPDC) [4]. This centralized database setup represents an invaluable resource for knowledge discovery and evidence based medicine. Since 2005, the health department of the state government, Queensland Health, has implemented a series of initiatives to improve performance monitoring and governance. As an example, the VLAD (Variable Life Adjusted Display) system is in place to detect extraordinary trends and occurrences [1], using data from QHAPDC. Significant challenges exist in developing efficient and effective ways of maximizing the utility of this data resource. A visualisation toolkit tailored for health data such as QHAPDC is likely to have significant benefits to both Queensland Health and the broader medical community. This article describes a step towards developing such a toolkit.

In the following sections, we describe various techniques used to visualise QHAPDC data. Our visualisation is exploratory in nature, with the overall aim of expanding the audience for healthcare data, and the following specific aims guiding the selection of techniques:

1. To assess data quality, and hence to identify potential improvements to the data collection process. (See Section 3.1 for an example where coding issues were identified through a simple histogram.)
2. To detect anomalies (both positive and negative) in clinical practices, and hence to promote clinical practice improvement. (See Section 3.2 for such an attempt.)
3. To identify temporal trends and spatial variation in the data for better allocation of health care resources. (See Section 3.5 and 3.6.)
4. To identify the potential research value of the routinely collected data; to generate medical hypotheses that lead to further research projects. (See Section 3.4.)

Visualisation of public health data has been demonstrated to enhance knowledge and support decision making (see for example [12, 22, 23, 24]). A unique challenge of the state-wide QHAPDC database is that it is essentially a repository for a number of largely independently generated data collection sites, typically different hospital campuses. While this provides comprehensive and rich data for visualisation techniques, there are key challenges of "noisy" and missing data associated with possible non-uniformity of data coding practices across hospital sites.

This article is organised as follows: Section 2 describes the dataset to be used and briefly explains

Copyright 2010, Australian Computer Society, Inc. This paper appeared at the Australasian Workshop on Health Informatics and Knowledge Management (HIKM 2010), Brisbane, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 108. Anthony Maeder and David Hansen, Eds. Reproduction for academic, not-for-profit purposes permitted provided this text is included.

terms such as International Classification of Diseases, 10th Revision (ICD-10) and Diagnosis-Related Group (DRG). Section 3 reports on a collection of visualisation plots within specific areas of health data. Section 4 summarises the findings, highlights the unique contributions of these techniques and discusses opportunities for further research in this field.

2 Data Description

QHAPDC provides comprehensive data on patient demographic and clinical/treatment occurrences. In this preliminary study, we focus on the clinical data as coded by the International Classification of Diseases.

2.1 International Classification of Diseases

The morbidity information in QHAPDC is encoded following the *International Classification of Diseases and Related Health Problems, 10th Revision, Australian Modification* (ICD-10-AM, [6]). ICD is published by the World Health Organization (WHO) as a standard way of coding morbidity and mortality statistics [21]. The 10th Revision (ICD-10) is the current standard. ICD-10-AM includes the Australian extensions to ICD-10 and contains more than 20,000 codes in total.

ICD-10 codes are organized in 22 chapters, each denoted by a capital letter. A chapter is further divided into blocks, each denoted by a number from 0 to 99. A block again can be further refined by appending a fraction number[1]. Take the code `I21.4` as an example. The leading letter `I` indicates it belongs to the chapter of *diseases of the circulatory systems*; `I21` is the block for *Acute Myocardial Infarction* (AMI); The fraction `.4` indicates the infarction is *subendocardial* [21].

Based on diagnoses and procedures, hospital cases are classified into over 600 *Diagnosis-Related groups* (DRGs)[2]. DRGs can be further grouped into 25 *Major Diagnostic Categories* (MDCs) [17]. For example, MDC 05 *Diseases and disorders of the circulatory system* contains DRGs ranging from `F01A`—*Implantation or Replacement of AICD, Total System with Catastrophic complication* to `F76B`—*Arrhythmia, Cardiac Arrest and Conduction Disorders without Catastrophic or Severe complication*. In later discussions, we focus on cases in MDC 05 that have a principal ICD code in the `I21` (AMI) block.

2.2 AMI data

Ischaemic heart disease is the number one cause of death for Australians [14, page 44]. Acute myocardial infarction (AMI), also known as a *heart attack*, is a major clinical form of ischaemic heart disease that affects many Australians. The Australian Institute of Health and Welfare, citing the 2004-2005 National Health Survey, suggests that about 1.8% of Australians reported a history of AMI [16, page 183].

In this article, all visualisations are demonstrated using an AMI dataset from QHAPDC[3]. Specifically, the dataset was created by extracting QHAPDC records that met the following criteria:

1. The patient was treated in one of 6 *principal referral and specialised* public hospitals.
2. The principal diagnosis was "acute myocardial infarction" (i.e., the ICD is in the I21 group)[4].
3. Records were extracted from January 2005 to December 2008.

3 Visualisation

3.1 Histograms for data cleaning

As a preliminary step prior to involved statistical analysis, it is desirable to "feel" the structure of data and assess the data quality to avoid "garbage in, garbage out".

We start out by showing how simple graphics such as a histogram can expose anomalies in data, and how interactions such as colour brushing [26] can help discover unexpected relations.

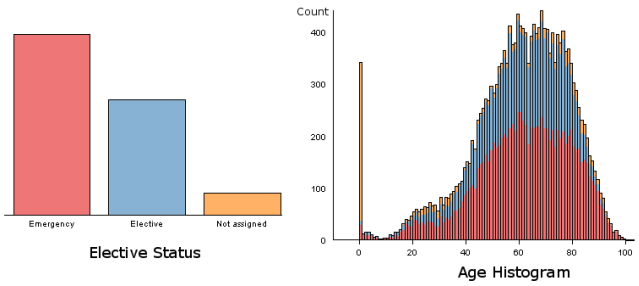


Figure 1: Histogram of the age of the AMI patients described in Section 2.2.

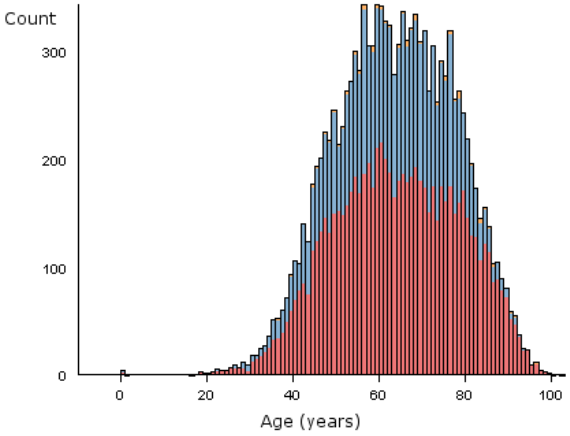


Figure 2: Histogram of the age of patients in the circulatory disease MDC. The shape is less skewed than in Figure 1. The colour scheme for elective status is the same as in Figure 1.

As the AMI dataset has about 20,000 cases, we expect certain continuous variables to approximately follow a unimodal distribution[5]; otherwise we may suspect the data come from multiple distinct sources.

As an example, Figure 1 displays the histogram of the age of all patients. One distinctive feature of the histogram is the spike at age 0. Colour-brushing various discrete variables in the data revealed a connection between elective status and patient's age.

[1] In ICD-10-AM, morphology codes (for identifying the morphology of neoplasms) follow a different convention: a morphology code consists of 4 digits following the letter `M`.

[2] DRGs are used in the casemix funding model in Australia.

[3] The visualisation techniques we present are to gain understanding of the data. Further confirmatory studies are necessary before definitive conclusions can be made.

[4] Late in Section 3.1 we shall see that extraction based on the principal diagnosis may not be a good choice, as some coding practice is either unreliable or unintuitive.

[5] It is useful here to distinguish measurement variables from nominal variables. For example, a patient's age is a measurement variable. But his or her length of hospital stay is not a measurement variable, as it is an aggregated function of whether the patient is discharged on a particular day, which is nominal.

Namely, most newborns in the dataset did not have elective status assigned. Moreover, the histogram also shows a group of patients of age 16 to 37 who did not have elective status. Most of these patients were female and in obstetric DRGs. On subsequent investigation we found in the QHAPDC manual ([4, Section 7.30]) that admissions for normal delivery and admissions which begin with the birth of the patient do not have elective status assigned. Why did these patients in obstetric DRGs have AMI as their principal diagnosis? Consultation with an obstetrician and coding staff may help answer the question.

Figure 1 has highlighted the heterogeneous nature of the dataset. Upon further examination, we restricted the data to cases in the MDC 05—Diseases and disorders of the circulatory system [15], which account for 70% of the original cohort. The age distribution of the selected subgroup is shown in Figure 2. The new histogram has a shape more closely resembling a normal distribution, indicating a more homogeneous group of patients.

3.2 Fluctuation plots: visualising patient transfer among hospitals

With a state-wide database, we can visualise patient flow among hospitals within the state, which can potentially help to facilitate the logistics involved in patient transfer. A fluctuation plot serves this purpose well. A *fluctuation plot* is used to visualise the contingency table of two discrete variables, where the count of each combination is represented by a square of proportional area.

Figure 3 shows AMI patient transfers since 2005. The size of a square indicates the relative number of transfers from one hospital to another. One can see that most patients were transferred to hospital **A**, as the vertical column **A** has most large squares). One can also infer that hospital **B** and hospital **D** are referral destinations for hospital **C** and hospital **F**, respectively, as indicated by the two large squares at the coordinates (**B**,**C**) and (**D**,**F**).

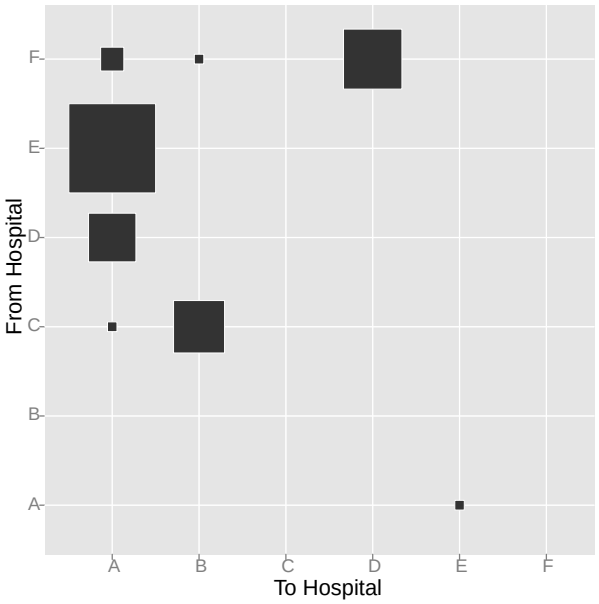


Figure 3: Fluctuation plot showing number of patients transferred among 6 Queensland hospitals. The size of a square indicates the number of transfers. It shows that hospital A is a major transfer destination for AMI patients.

For patients being transferred, it would be expected that they would receive the same principal diagnosis in the second hospital. We used a fluctuation plot to verify this assumption. In Figure 4, we see that principal diagnoses were largely identical, as the squares along the diagonal contain much of the total area of all the squares. However, small squares off the diagonal indicate potential inconsistencies in diagnosis or coding, or a logical clinical phenomenon not reflected in the current data set. As the principal diagnosis affects how a patient is treated, these data anomalies warrant further investigation.

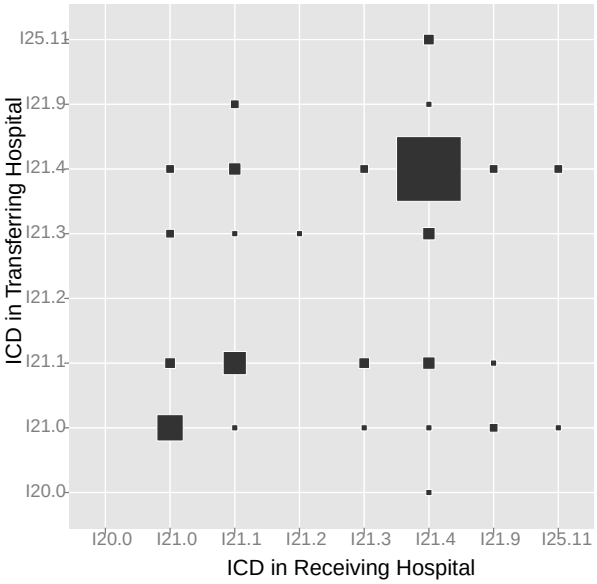


Figure 4: Fluctuation plot showing principal diagnoses before and after transfer. Squares along the diagonal correspond to cases with an unchanged principal diagnosis after transfer. Squares off the diagonal correspond to cases with different principal diagnoses before and after transfer. Size of a square reflects the relative number of particular combination.

3.3 Mosaic plots: visualising clinical pathways

The *Clinical Practice Improvement Centre* (CPIC) at Queensland Health has developed a set of state-wide Cardiac pathways [6].)

As a simplified example, suppose a patient presented at an emergency department with chest pain. Based on her ECG result, a medical officer decided whether she had a ST-segment elevation myocardial infarction (STEMI) case or a Non-ST-Segment Elevation case. If it was a STEMI case, then percutaneous coronary intervention (PCI, insertion of a catheter into coronary vessels to remove blockages or improve blood flow) was needed; otherwise PCI might not be necessary. With a mosaic plot, we can gain insight into how these clinical pathways have been followed in hospitals[7].

The mosaic plot is closely related to the fluctuation plot. In a mosaic plot of two discrete variables, each

[6] A clinical pathway is "a document outlining a standardised, evidence-based multidisciplinary management plan, which identifies the appropriate sequence of clinical interventions, timeframes, milestones and expected outcomes for a homogenous patient group" [10]

[7] A more data-driven approach to visualise the potential causal relation between comorbidities and procedures is to learn a Bayesian network using the database. One such learning algorithm can be found at [25]).

major cell shows the relative frequency of data observations corresponding to values of the two variables. With an additional discrete variable, each cell is further divided according to the conditional frequency of the variable in the cell. Hence at both the local (cell) and global level, the visualisation shows the degree of non-uniformity across variable values.

Here we use mosaic plots to visualize the conditional relationship between PCI and ST-segment elevation. Unfortunately ST-segment elevation is not explicitly encoded in ICD-10-AM or anywhere else in QHAPDC[8]. After consulting coding staff from a Queensland hospital, we used the following rule to estimate ST-segment-elevation status: For patients with diagnoses I20.0-I20.3, we assumed ST-segment elevation (STEMI in Figure 5); for patients with diagnosis I20.4, we assumed no ST-segment elevation (NSTEMI in Figure 5); for patients with diagnosis I20.9, we assume the ST-segment elevation status was unknown (UNKNOWN in Figure 5). In addition, we assume that PCI was performed if and only if the case is in one of the DRGs: `F10Z`, `F15Z`, or `F16Z`.

Figure 5 shows that practices are relatively consistent across 3 hospitals of similar sizes. It also shows that not every patient in the STEMI group received PCI, whereas some patients in Non-STEMI group did receive PCI.

A possible explanation for the above inconsistency between pathways and data is that we have not correctly estimated STEMI status for patients. If that is the case, then Figure 5 shows the importance of coding the STEMI status in ICD-10-AM.

Mosaic plots have been used in Queensland Health for displaying the relationships between risk and the first and second Troponins, for the chest-pain management protocol. People found it "the most appropriate and easiest" way to present the relationships and information, as one could clearly see the "size" of the issues from presenting the information in this manner.

In general, the health-policy workers often need to present two or three way tables. It is suggested that the mosaic plot and its variants have a high applicability for these tables. The main limitation of the mosaic plot is that basic software such as Excel does not have the capability and not many people are aware of this style of presentation method.

3.4 Heatmaps: visualising the connection between severity and morbidity

As mentioned in Section 2, the QHAPDC database consists of two parts: socio-demographic data and clinical data, which includes morbidity coded data. How to use morbidity data is an important but challenging problem. For example, all past INFORMS data mining contests centered on mining the morbidity data to predict health care outcomes [8, 9]. Here we introduce a use of a heatmap to visualise morbidity to effectively identify risk factors and protective factors for both mortality and hospital long-stay.

Heatmaps have been primarily used in genomics to visualise DNA microarrays [5]. As a DNA microarray can be regarded as a 2-dimensional matrix, its rows and columns can be permutated to highlight clusters. Similarly, morbidity data coded in ICDs can be thought as a matrix M:

1. Each patient is a row;
2. Each ICD code is a column;

[8] In ICD-10-CA (Canadian enhancement of ICD-10), STEMI has the code R94.30 (see [7, page 26]).

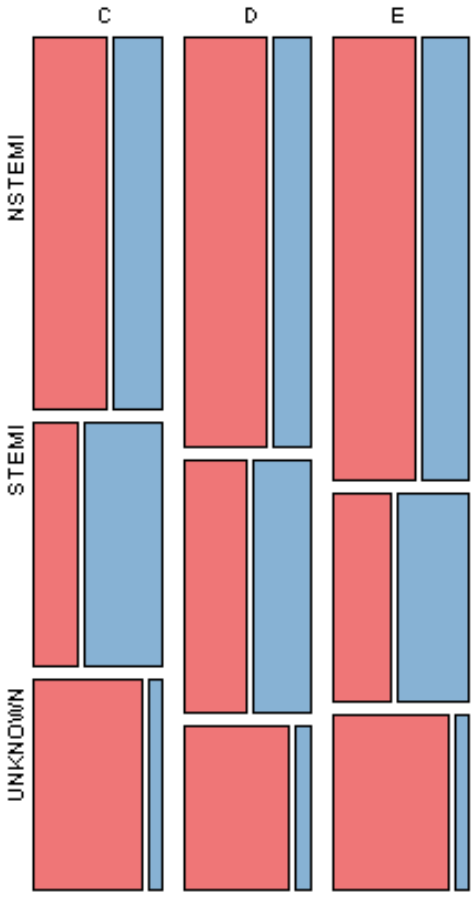


Figure 5: Mosaic plot showing the relation between ST-segment elevated MI (STEMI) and percutaneous coronary intervention (PCI). Three hospitals of similar sizes (C, D, and E) are compared. Blue shading in the plot corresponds to patients who received PCI. The clinical pathways suggest that STEMI patients should receive PCI, whereas non-STEMI patients may not need PCI. The plot shows that STEMI patients are more likely to receive PCI, but that relation cannot be taken as absolute in practice.

3. A matrix cell $M(i, j)$ takes value 1 if patient i was diagnosed with morbidity j, and it takes value 0 otherwise.

Therefore, we consider visualising morbidity data with a heatmap as a natural choice.

In a heatmap, ICDs can be clustered just as in a DNA microarray. Patients could also be clustered with respect to their morbidity information. To identify risk and protective factors, however, we sorted the patients (rows) with the following measure of severity[9]:

$$\text{SEVERITY}(x) = \begin{cases} C - \text{LOS}(x) & \text{if } x \text{ died in hospital} \\ \text{LOS}(x) & \text{otherwise} \end{cases} \quad (1)$$

where $\text{LOS}(x)$ is the length of hospital stay (in hours) of patient x, $C = 2 \max_x \text{LOS}(x)$.

Figure 6 shows such a heatmap. Records are ordered from top to bottom with increasing severity. It shows that a cardiogenic shock has strong correlation with AMI mortality.

In Figure 6, we have included only diagnosis codes; a heatmap could also be generated with both diagnosis and procedure codes to identify correlations (1) between diagnoses and procedures, and (2) between procedures and reduced/prolonged hospital stay. Also the severity measure could be adjusted by age and gender.

3.5 Time plot: visualising temporal trends and seasonal patterns

A *time plot*, where observations are plotted over time [2], displays temporal variation of the data. It helps us

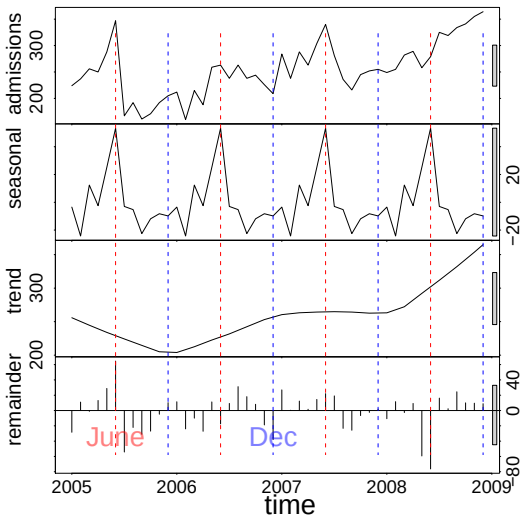
1. recognize trends and seasonal patterns in the data, which is potentially useful for resource allocation and planning, and
2. check the consistency of health care quality and spot outliers that warrant further investigation.

Regarding the second point, a time plot can complement the VLADS system (see Section 1) to provide visualisations that are easier to understand by people with no knowledge of statistical process control.

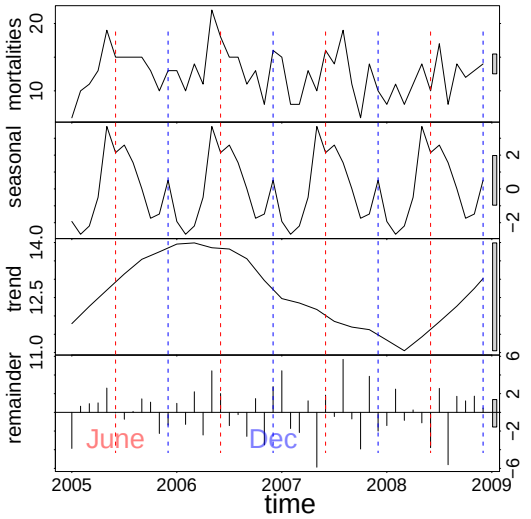
Here we give a simple example of comparing the admission and mortality of AMI patients. Figure 7 plots the number of admissions and deaths each month from January 2005 to December 2008. We use the method described in [3] to decompose the two time series into seasonal, trend and irregular components. Figure 7(a) shows more patients were admitted during the month of June each year, a winter month in the southern hemisphere. It also shows a trend of increasing numbers of admissions since 2006. The seasonal plot in Figure 7(b) shows that most AMI deaths occurred during the winter months in Australia. Comparing the trend plots in the two figures, we see that in 2005, the number of deaths increased while the number of admissions decreased. But in the years 2006 and 2007, according to the available data, the mortality rate decreased monotonically.

3.6 Disease map: visualising spatial variation

As a complementary technique to time plots, which are useful for visualising temporal variation of the data, disease maps are useful for visualising spatial variation. A *Disease map* is a popular type of visualisation for public health data (see for example [12, 22, 23]).



(a) number of admissions



(b) in-hospital mortality

Figure 7: Time plots showing seasonal, trend and irregular decomposition of admission and mortality using loess smoothing [3].

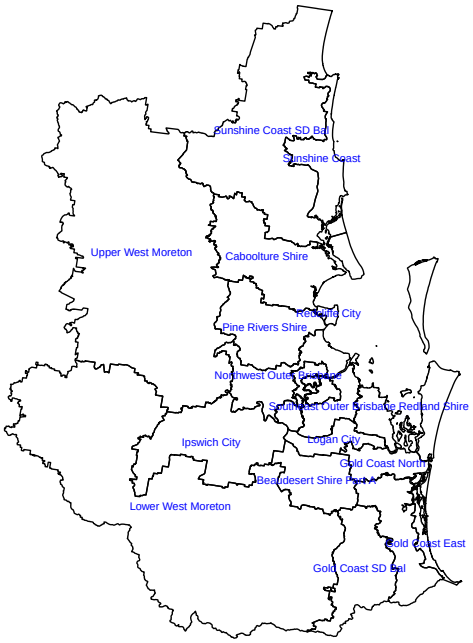


Figure 8: Statistical subdivisions of South East Queensland in 2006. Map data is courtesy of the Australian Bureau of Statistics.

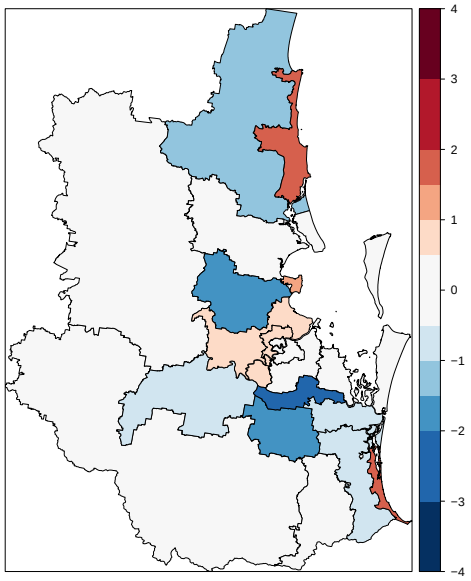


Figure 9: Residual map for the population aged 85 and over in South-East Queensland. A quasi-Poisson model was fitted to the estimated number of residents aged 85 and over in each statistical subdivision. The standardized residuals are plotted to highlight extreme values. The map shows that dense coastal population centres have a higher percentage of seniors.

For a geographic division, let X be the number of observed cases, and E be the expected number of cases (for example by adjusting for its population and age distribution). Then a disease map will highlight the discrepancy between X and E, for example, by plotting the standardized mortality/morbidity ratio (SMR) X/E for each geographic division. Both AMI incidence and mortality rate in Queensland can be mapped in this way to understand their geographic variation, which itself reflects the distribution of indigenous population or older demographic in particular areas.

Due to the sensitive nature of mortality data, however, maps of AMI incidence and mortality rate are not shown here. To still demonstrate the technique, we use the population data from the 2006 Australian Census [19]. Figure 8 shows a map of statistical subdivisions in South East Queensland. The region accommodates more than 66% of Queensland population [18]. Figure 9 shows a residual map for the percentage of the population aged 85 or over. From the map, we see that in the year 2006, both the Sunshine Coast and the Gold Coast East have higher proportions of older population.

The health-policy workers consider the thematic map an "excellent" way to visualize the allocation of resources, and to assess areas of need, provided the appropriate data and technique is used in displaying the information. The availability of mapping layers, however, limits the applicability of this technique.

4 Discussion

This paper has demonstrated that visualisation can provide important clinical insights into various aspects of the routinely collected health data. Our results show that visualisation is helpful in every stage of the data life-cycle, from collection and validation (see Section 3.1), to reporting (see Section 3.2), to knowledge discovery (see Section 3.4), and to anomaly detection (see Section 3.5 and 3.6). Visualisation complements more formal statistical analysis with its flexibility, and generates final products easily understood by data managers and clinicians, who have not necessarily received formal statistical training.

The overarching goal of our project is to expand the audience for healthcare data. This paper makes the following contributions to health-care industry in general and to health informatics & knowledge management in particular:

1. It assesses the applicability and value of various visualisation techniques in health and medical data.
2. It proposes a new way to visualise diagnosis coding with respect to case severity.
3. It constitutes an important step toward a visualisation tool-kit for healthcare workers.

To further develop the toolkit, we are extending the visualisation techniques and applying these techniques to a wider range of data, including a Cesarean section dataset and an Orthopaedic hip-replacement dataset from Queensland Health. At the same time, we have been collaborating with clinicians from Queensland Health to access the utility of these visualisation techniques and to develop types of visualisations that are informative in clinical settings. We also continue to work with CPIC to understand the nature and quality of the Queensland Health data, and to further develop the insights that can be gained with these unique resources.

[9] This severity measure is rather coarse. Currently we are undertaking research for a more accurate estimate of severity [13].

Acknowledgment

This work is supported by an ARC Linkage Grant to MG, DO, JC, and JW (LO 0776417) and the Clinical Practice Improvement Centre, Queensland Health. We thank Michael Milford and Chris Nolan for help on graphics manipulation, and Karen Borchardt for help on locating the ABS census data.

References

[1] Clinical Practice Improvement Centre. *VLADs for dummies*. Wiley, Queensland Health edition, 2008.

[2] Chris Chatfield. *The Analysis of Time Series: An Introduction, Sixth Edition*. Chapman & Hall/CRC, 2003.

[3] Robert B. Cleveland, William S. Cleveland, Jean E. Mcrae, and Irma Terpenning. STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1):3–73, 1990.

[4] Sue Cornes. *Queensland Hospital Admitted Patient Data Collection (QHAPDC) Manual*. Health Information Centre, Queensland Health, 2005–06 edition, 2005.

[5] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(25):14863–14868, 1998.

[6] The National Centre for Classification in Health (Australia). ICD-10-AM. `http://nis-web.fhs.usyd.edu.au/ncch_new/2.aspx`.

[7] Canadian Institute for Health Information. Canadian coding standards for ICD-10-CA and CCI. `http://secure.cihi.ca/cihiweb/dispPage.jsp?cw_page=RC_382_E`.

[8] The Institute for Operations Research and the Management Sciences. INFORMS data mining contest 2008. `http://informsdataminingcontest.googlepages.com/`, 2008.

[9] The Institute for Operations Research and the Management Sciences. 2009 INFORMS data mining contest. `http://www.informsdmcontest2009.org/`, 2009.

[10] Queensland Health. Queensland health implementation standard: Clinical pathways. `http://www.health.qld.gov.au/cpic/pdf/clin_path_imp_strd.pdf`.

[11] Queensland Health. Annual public hospitals performance report 2006-07. `http://www.health.qld.gov.au/performance/performance_report.asp`, November 2008.

[12] Nada Lavrac, Marko Bohanec, Aleksander Pur, Bojan Cestnik, Marko Debeljak, and Andrej Kobler. Data mining and visualization for decision support and modeling of public health-care resources. *Journal of Biomedical Informatics*, 40(4):438 – 447, 2007.

[13] Wei Luo, Marcus Gallagher, and Janet Wiles. Bayesian modelling of inpatient severity. In preparation.

[14] P. Magnus and K. Sadkowsky. *Mortality over the twentieth century in Australia : trends and patterns in major causes of death*. Australian Institute of Health and Welfare, 2006. `http://www.aihw.gov.au/publications/phe/motca/motca.pdf`.

[15] Australian Institute of Health and Welfare. Separation, patient day and average length of stay statistics by Australian Refined Diagnosis Related Group (AR-DRG) version 5.0/5.1, Australia, 1998-99 to 2006-07. `www.aihw.gov.au`.

[16] Australian Institute of Health and Welfare. *Australia's Health 2008*. Australian Institute of Health and Welfare, 2008.

[17] Department of Health and Australia Ageing. MDC-Partition-DRG structure. `http://www.health.gov.au/internet/main/publishing.nsf/Content/2A68FBBD47DC69D0CA25753E00032FC2/$File/MDC-Partition-DRG%20Structure.pdf`.

[18] Department of Infasture and Queensland Government Planning. South East Queensland. `http://www.dip.qld.gov.au/seq`.

[19] Australian Bureau of Statistics. 3235.0 - population by age and sex, Australia, 2006. `http://www.abs.gov.au/ausstats/abs@.nsf/Products/3235.0~2006~Main+Features~Queensland?OpenDocument`, August 2008.

[20] Australian Bureau of Statistics. Australian demographic statistics, dec 2008. `http://www.abs.gov.au/ausstats/abs@.nsf/mf/3101.0`, April 2009. Catalogue number 3101.0.

[21] World Health Organization. International statistical classification of diseases and related health problems, 10th revision, version for 2007. `http://apps.who.int/classifications/apps/icd/icd10online`.

[22] Bambang Parmanto, Maria Paramita, Wayan Sugiantara, Gede Pramana, Matthew Scotch, and Donald Burke. Spatial and multidimensional visualization of Indonesia's village health statistics. *International Journal of Health Geographics*, 7(1):30+, June 2008.

[23] Mohsen Rezaeian. How to visualize public health data? Part one: Box plot and map. *Middle East Journal of Family Medicine*, 6(10):19–24, December 2008.

[24] Mohsen Rezaeian. How to visualize public health data? Part two: Direct and indirect standardization methods. *Middle East Journal of Family Medicine*, 7(1):42–44, January 2009.

[25] Oliver Schulte, Gustavo Frigo, Russell Greiner, Wei Luo, and Hassan Khosravi. A new hybrid method for Bayesian network learning with dependency constraints. In *CIDM*, pages 53–60, 2009.

[26] Martin Theus and Simon Urbanek. *Interactive Graphics for Data Analysis: Principles and Examples (Computer Science and Data Analysis)*. Chapman & Hall/CRC, 2008.

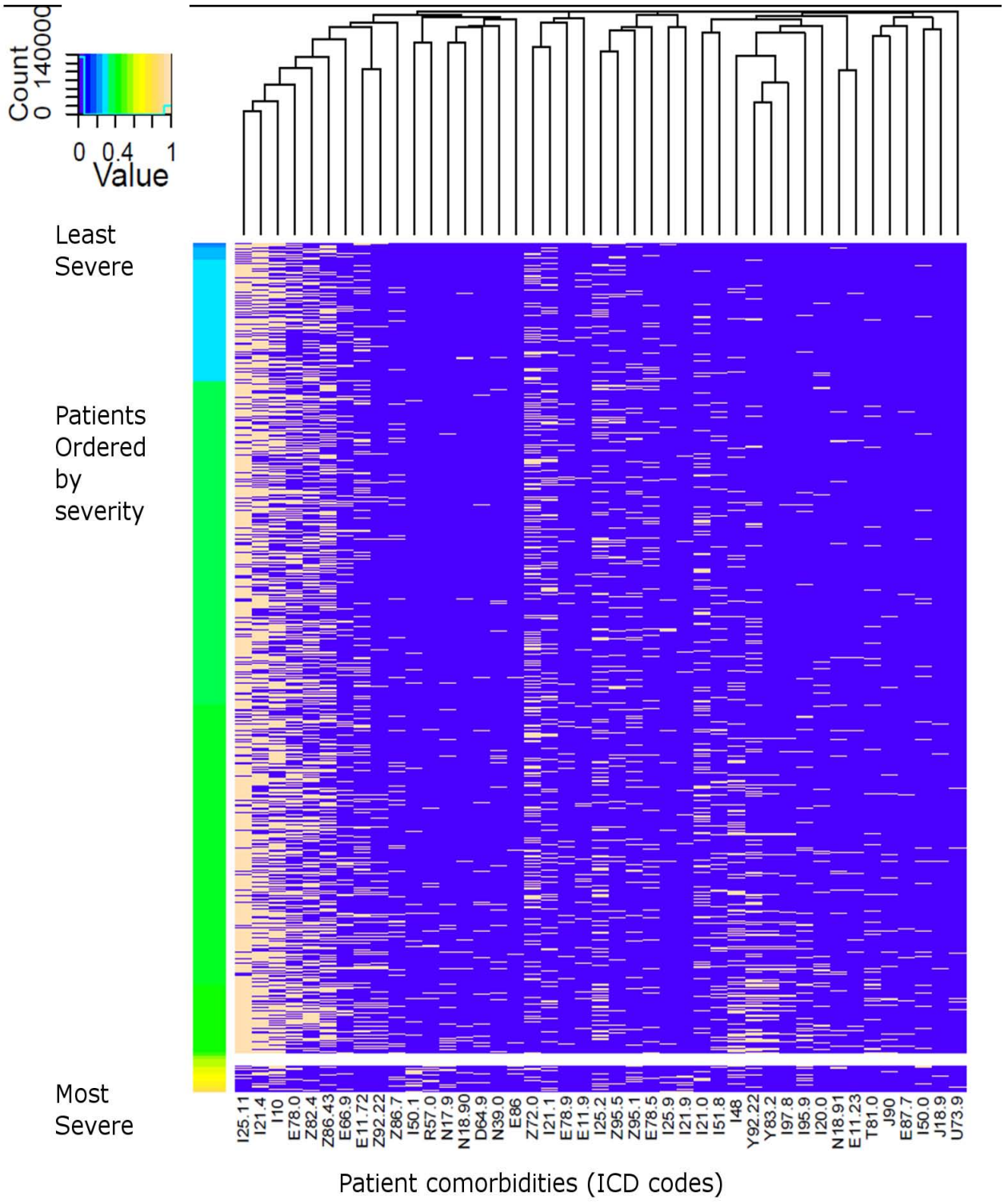


Figure 6: Heatmap showing morbidity codes from a single hospital. Each row represents a patient's morbidity codes; Rows are ordered from top to bottom with increasing severity. A horizontal gap near the bottom separates patients who survived (above the white line) from those who died in hospital (below the line). The vertical side bar on the left shows each patient's severity estimate (normalized to a number between 0 and 1, and then colour-coded with the colour key at the top left corner). Each column represents the distribution of an ICD code across patients. (Columns have been automatically clustered based on similarity of their distribution. The dendrogram at the top shows the clustering.) The plot shows that most patients had the code `I21.4` (acute subendocardial myocardial infarction), which is interpreted as non-ST-segmented Elevated MI in the hospital under study. It also shows that conditions `R57.0` (cardiogenic shock) and `I50.1` (left ventricular failure) are strongly correlated with mortality, while conditions `T81.0` (haemorrhage and haematoma complicating a procedure, not elsewhere classified) and `Z92.22` (personal history of long-term (current) use insulin) often occur among patients with long hospital stays, but not so often among patients who died.