Temporal Event Mining of Linked Medical Claims Data

Graham Williams[1], Chris Kelman[2], Rohan Baxter[1], Lifang Gu[1], Simon Hawkins[1], Hongxing He[1], Chris Rainsford[1], and Deanne Vickers[1]

[1] Enterprise Data Mining
CSIRO Mathematical and Information Sciences
GPO Box 664, Canberra, ACT 2601, Australia
Firstname.Lastname@csiro.au
http://datamining.csiro.au
[2] Commonwealth Department of Health and Ageing
Canberra, Australia
Firstname.Lastname@health.gov.au

Abstract.
For universal health care systems such as that found in Australia, the health care budget constitutes one of the largest items for government expenditure. Identifying even small pockets of inefficiencies and wastage can have considerable impact (one percent of $30 billion is still a significant amount) and can result in shifting funds to where they are most needed. We report on the use of a data mining approach to the problem of empirically identifying episodes of health care introduced in [1]. The task is one of detecting, within a temporal sequence relating to a hospital admission, primary care that might be associated with that admission. The results have been used in studies of the relationship between out of hospital and in-hospital care for policy research and, for example, exploring issues relating to preventable hospitalisations.
Keywords: temporal data mining, record linkage, administrative data, health services.

1 Introduction

Data mining has often been used to mine massive administrative data that is collected by organisations for the running of their day-to-day business. This resource is typically not directly suitable for mining and hence the knowledge discovery from data mining process involves considerable effort in transforming the administrative data into forms suitable for analysis.

A characteristic of many administrative data holdings is that they contain often hundreds of millions of transactions relating to perhaps tens of millions of clients. Each transaction has a time and generally also

a location. The time component of such transactions paints a historical picture of a client's interaction with the system — whether that be financial purchases or medical procedures.

Temporal data mining covers data analysis dealing with issues of time. A common requirement is to identify temporal patterns in very large collections of data. Many data mining research efforts have attempted to tackle this problem, with some focusing specifically on identifying change points in time series data (e.g., [2, 3]).

In this paper we consider the issue of temporal data mining in solving the problem of identifying significant events in the temporal sequence. In our case these events correspond to medical episodes of care that are associated with a period of hospitalisation. The task is to identify which of a patient's complex interactions with the whole medical system are actually clinically related to an episode in hospital. The approach presented in this paper continues a research path, reported initially in [1], toward better mining of complex and distributed health data to better understand and identify episodes of care.

2 Domain

2.1 Administering Health Care in Australia

Health care in Australia is administered by both federal and state governments. The federal government provides universal health care coverage, subsidising general practitioner consultations, pathology tests and diagnostic imaging, and indeed most medical service provided outside of hospitals. The state governments fund the hospitals and various ancillary services.

As a consequence of this separation of funding, administrative data is kept in a multitude of different databases and there is no centralised store of this information — there is no complete picture of the administration of health care in Australia. Instead, for each patient a centralised administrative record exists of their primary care events but records of hospital care are spatially distributed.

The Commonwealth Department of Health and Ageing (a federal government department) and Queensland Health (a state government department) have collaborated to link administrative medical data so that, for the first time, a comprehensive picture of health care delivery can be developed. The contribution of this to a better portrayal of the overall delivery of health care and this leading to better policy can not be un-

derstated. Data mining can now be used to explore complete episodes of care in the data.

CSIRO Data Mining has created this unique cleaned and linked administrative health dataset bringing together the *Queensland Hospital Morbidity* data and the Commonwealth *Medicare Benefits Scheme* (MBS) and *Pharmaceutical Benefits Scheme* (PBS) data.

The *Queensland Linked Data Set* (QLDS) [4] links de-identified, administrative, unit-level data, allowing de-identified patients to be tracked through episodes of care as evidenced by their MBS, PBS and Hospital records. This dataset provides a unique view of service utilisation and cost trends and patterns in the overall delivery of Commonwealth and State funded health care.

2.2 Data Preparation

The data used for this study were extracted from the QLDS [4]. This multi-gigabyte dataset covers 30 million hospital separations over 4 years. A *separation* refers to the data associated with an episode in hospital, containing patient admission and discharge information. For each separation all of the corresponding patient's relevant Medicare (MBS) services were obtained and these formed the basis of the service counts and aggregated costs used for applications of this study. There are over 100 million Medicare transactions over the 5 years for this collection of patients. The MBS data covers a period of 6 months prior to the beginning of the hospital data, and 6 months after.

Data preprocessing included filtering the QLDS hospital data to remove hospital separations corresponding to changes in admission status that do not reflect an end to the hospital episode. These records include statistical admissions and discharges, which include changes, for example, from an emergency admission to a longer term admission. Typical of administrative data such changes actually appear as separate records in the database and if not dealt with appropriately can have significant impact on any analyses. Other such administrative changes include hospital transfers.

Data preprocessing of the MBS data included filtering to remove MBS items that were provided within the hospital which are identified in the data as those records with a hospital flag set. These services are rendered in hospital to the patient either as an outpatient, or else as a private patient in a public hospital.

Only MBS items in the following categories are considered: *diagnostic imaging*, *pathology*, and *non-invasive investigative procedures*.

2.3 Pre Admission and Post Discharge Medical Services

The identification of groups of services and costs relating to a particular episode of care is a central issue in health services research [5–9]. An episode of care is *a block of one or more medical services, received by an individual during a period of relatively continuous contact with one or more providers of service, in relation to a particular medical problem or situation* [6]. One example of the application of episodes of care is their use in measuring the costs and services for a particular disease or condition of interest, such as diabetes, asthma or depression.

In our application, the focus of the treatment episode is a hospital admission and the block of related medical services are pathology tests, diagnostic imaging and non-invasive investigative procedures. *A priori* these medical services are likely to be associated with a hospital admission rather than ongoing ambulatory health care. These medical services are not complete, but they contain the most expensive items in pre-admission preparation for a hospital admission and post-discharge care.

We anchor our episode of care around a particular hospital admission. The episode of care begins x days prior to the admission and ends y days after discharge from hospital. All related medical services received in these two intervals are included in the episode of care. Figure 1 captures the situation. For our purposes services rendered whilst in hospital are ignored.

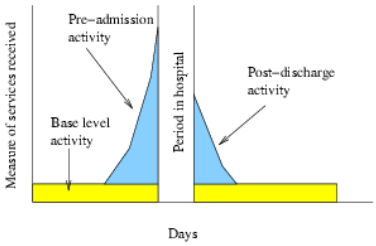


Fig. 1. Example of the pre-admission and post-discharge task.

Previous work using episodes of care has concentrated on one or two narrow clinical areas. This meant that the model for estimating an appropriate episode of care interval could be hand-crafted and clinically assessed. In contrast, our application required episodes of care to be estimated for 666 *Diagnostic Related Groups* (DRGs) covering the full range of hospital admission types.

DRGs were developed to simplify hospital funding by grouping medical conditions that had similar resource requirements during a patient's hospital stay. Conditions are grouped together if they involve either closely related diagnoses or similar surgical procedures. Variables such as age and existence of complications or comorbidities are taken into account. As a result of this, cases within the same DRG are not necessarily equivalent clinically and will likely have different pre and post management requirements. Consequently, the use of DRGs for this analysis will be expected to be affected by the inbuilt variance in patterns of service use within many of the DRGs.

In a previous study we assumed fixed intervals of 90 days for pre-admission and post-discharge for each of the DRGs. These fixed interval assumptions are clearly not clinically valid. For example, the DRG for a broken femur should have a 0-day pre-admission interval (since broken legs are not planned for) and a 40 day post-discharge interval (the average recovery time). A delivery admission DRG for a birth will have a six-month pre-admission time (as women typically obtain pathology tests and diagnostic imaging six-months prior to a birth).

Time and resources do not allow for individual assessment of each of the 666 DRGs. Our solution is a robust and automatic means, based on machine learning techniques, to estimate the intervals for the 666 types of episodes of care.

Others have explored the problem of identifying cut-points (change-points or segmentation). The problem arises in many applications in data mining, artificial intelligence and statistics, including segmenting time series [10], decision tree algorithms and image processing. A range of criteria have been proposed in the literature for determining if some time series data should be segmented into two or more regions [11]. Cohen and Adams [12] describe the segmentation of categorical time series data using a voting experts approach to combine evidence for segmentation boundaries.

Our solution develops data driven estimates for the DRG episodes of care resulting from an ensemble of alternative estimation models (or panel of experts) being combined through a voting scheme [13]. We perform sensitivity analysis to assess the accuracy of the episodes of care estimates. We have also included a preliminary clinical assessment of the episode of care estimates for four DRGs.

3 Multiple Model Temporal Mining

Our data driven technique to time-frame-adjust DRG intervals of workup (pre-admission) and followup (post-discharge), motivated by [5]. The approach employs a *multiple experts* or *ensemble* paradigm [13] where several change-point estimators are employed and an averaged majority voting scheme is used to determine the final change-points, which then define the pre/post intervals.

A number of alternative estimators were investigated and four were chosen to form the *ensemble*: mean and variance optimisation; regression tree; multivariate adaptive regression splines; and multi point splines.

For each DRG a table containing a count of all pre-admission (and separately post-discharge) MBS services in diagnostic imaging, pathology, and non-invasive investigative procedures was constructed. Services were counted on a daily basis over all separations for the particular DRG. These were counted up to 180 days (6 months) pre-admission (and separately 180 days post-discharge). The daily counts were then normalised by dividing by the number of hospital admissions for that DRG.

The choice of 180 days pre/post was made to allow for a background pattern of servicing to be identified and then any intervals of increased servicing could be identified as being associated with the hospital separation. Therefore each of the methods produces its estimated best cutoff time for pre and post hospitalisations, which are optimised by its own criteria.

In the following, we describe each of the four methods and then explain how the *ensemble* voting method is employed to determine the final set of pre/post intervals. Further details, particularly of the choice of model parameters, is included in [1].

Mean and Variance Optimisation: The approach here is to search for a cut point t_c between T_1 and T_{180} ($T_1 < t_c < T_{180}$) that partitions the 180 day period (pre or post) into two parts $[T_1, t_c)$ and $[t_c, T_{180}]$. The search finds the value of t_c which maximises the difference between the mean of the two partitions and minimises the variance in each partition. The original service count data are smoothed using a moving average of 3 days prior and 3 days post and then normalised to values between 0 and 1. Suppose then that μ_1 is the normalised mean of $[T_1, t_c)$ and σ_1 is the standard deviation of this interval. Similarly μ_2 and σ_2 for $[t_c, T_{180}]$. The t_c chosen is that which minimises the term:

$$\frac{1}{|\mu_1 - \mu_2|} + \beta(std_1 + std_2) \tag{1}$$

The parameter β allows fine tuning of the importance of the variance with respect to the mean. By experimentation β was set to 20.

Regression Tree: Regression trees [14] recursively partition data to build a regression model for separate parts of the data. *Rpart*, the regression tree routine provided by R [15], was used, fitting a constant model to the leaves of the tree. The splitting criterion is based on ANOVA (performing an analysis of the variance on the data) whereby the cut-point maximising the reduction in the squared error fit to the actual data is chosen.

Multivariate Adaptive Regression Splines: MARS is a spline fitting regression approach where splines are fitted to distinct intervals of the data [16]. The cut points (called knots in MARS) are searched for through an exhaustive approach, optimising a so called loss of fit criterion. The R package [15] *mda* provides the implementation of the function *mars* used for this analysis.

Multi Point Splines: A series of natural splines were fitted to the data points using the R *spline* function with the *mean* function to approximate the background level of activity before the workup to admission from the first 100 (of 180) days of the data. Once the splines have been fitted a search from the day of admission finds the first spline knot (cut point) that falls below the mean value (ignoring the knot at day 0).

Ensemble: Each of the methods discussed in the previous four sections identifies a pre and post interval for each of the 666 DRGs. A successful approach in statistical modelling and machine learning has been to combine multiple predictive models through a voting mechanism. This approach is adopted here to combine the estimated pre and post cut points from each of the four "experts" into single pre and post cut points for each DRG. Our scheme of combining the resultant time cutoff values $\{t_1, t_2, \ldots, t_N\}$ from each of the N estimators $S = \{a_1, a_2, \ldots, a_N\}$ is as follows:

If we are interested in using a subset, n, of these N estimators to form different combinations, there are k such subsets, $k = {}^{N}_{n}\mathrm{C}$:

$$s_i = \{b_1^i, b_2^i, ...b_n^i\} \subseteq S \quad for \quad i = 1, 2, \ldots, k \tag{2}$$

where $b_1^i, b_2^i, ...b_n^i$ are n different estimators from the N estimators. Assume that the mean and standard deviation of the estimated cutoff values from

the subset s_i are $Mean(i)$ and $STD(i)$. The final optimal cutoff value T_c chosen is the mean value of the subset, which has the smallest standard deviation (Equation 3):

$$T_c = Mean(\arg\min_i(STD(i))) \tag{3}$$

In our case $N = 4$ and $n = 3$ was chosen. Therefore our method selects from the four estimated cut points the three having the least *variance*. The final cut point is then the *mean* value of these three.

4 Results

4.1 Examples

Figure 2 presents some sample DRG plots. Plotted is the data for 180 days prior to a hospital admission for the specified DRG and 180 days after discharge from hospital. The data are the daily count of MBS items in diagnostic imaging, pathology, and non-invasive investigative procedures received by all patients, normalised by dividing the number of separations for the DRG.

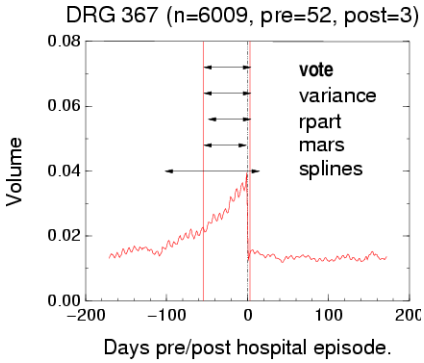
The typical example illustrated in Figure 2(a) shows a pre-admission interval having a marked increase in MBS activity 52 days prior to admission. The post-discharge interval of just 3 days illustrates no or little activity after the episode in hospital. The post-discharge quickly stabilises to the baseline activity.

It can be clearly seen from the plots that different DRGs have widely varying levels of baseline service use — this is consistent with the expectation that DRGs that capture patients who are older and sicker will generally be expected to regularly consume more health services outside hospital than those who are young and fit.

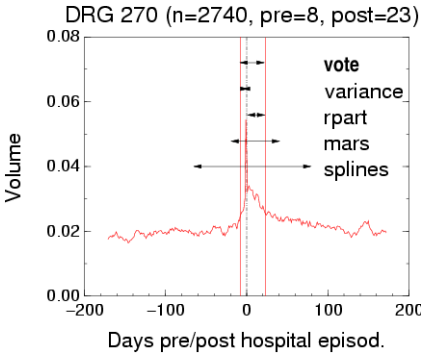
There is, however, a wide diversity of patterns. We illustrate some of the variation here, noting however that these are not the typical cases from among the 666 DRGs.

Figure 2(a), DRG 367 (*Cholecystectomy*) clearly demonstrates the clinical history of gall bladder disease. A prodromal illness with increasing number of investigations precedes the admission by up to three months. Following the operation, investigations virtually cease. The patient is cured. The estimates are 52 and 3 days.

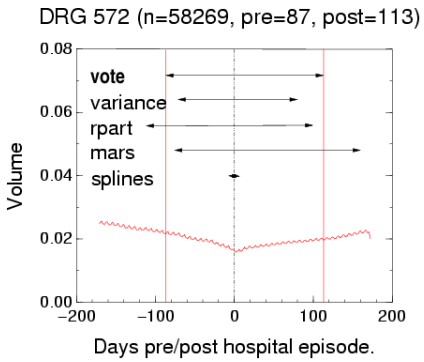
Figure 2(b), DRG 270 (*Unstable Angina*) illustrates an ongoing high service baseline with little increase in service use noted prior to admission



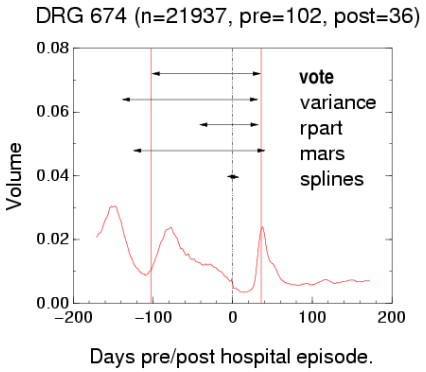
(a) DRG 367: Cholecystectomy without common bile duct exploration.



(b) DRG 270: Unstable angina without complication and/or comorbidity.



(c) DRG 572: Hospital admission for renal dialysis.



(d) DRG 674: Vaginal delivery without complicating diagnosis.

Fig. 2. DRG plots. The x-axis records days before and after hospitalisation with the period in hospital collapsed to 0. The y-axis is a normalised volume of MBS servicing. For subfigure (a) the estimated pre-admission and post-discharge intervals for the four methods, plus the final voted period, are shown (horizontal lines). The plot title indicates the DRG population as 6009 hospital admissions. The *ensemble* methodology, using voting, estimates a pre-admission interval of 52 days and a post-discharge interval of 3 days. The solid vertical lines correspond to the voted period.

until the last week. This is consistent with poorly controlled angina; worsening pain is experienced until investigations are updated and admission recommended. Following discharge, further monitoring is required. The *ensemble* methodology has identified estimates of 8 and 23 days.

A very different and indeed unusual pattern is illustrated in Figure 2(c). For this DRG 572 (*Renal Dialysis*) the pattern is not unexpected. The management pattern for this chronic disease is evident, with renal dialysis repeated every few days for these patients so that no true pre or post admission period exists. The estimates are 87 and 113 days. The algorithm produces an artefact that will need adjustment. Also shown in the graph is the V shaped plot demonstrating the 'double counting' that occurs when short readmission periods are involved.

Another class of patterns with a long lead up time before hospitalisation is typified by Figure 2(d) DRG 674 (*Normal Vaginal Delivery*). Standard obstetric practice for mid and late term pregnancy check-ups is clearly demonstrated (at 5 and 5 moths prior to admission), with a small amount of post-delivery follow-up. Pregnancy, being a process that does not usually involve ill health, does not comply with the usual pattern of disease. It is not surprising that the algorithms fail to capture the full 'treatment' cycle here.

Further examples and results of a study on the sensitivity of the estimated episodes of care have been presented elsewhere [1].

4.2 Preliminary Clinical Evaluation

DRGs involving unplanned admissions can have been supplied by a clinician. They are identified in Table 1.

Unplanned admissions, *a priori*, should have no workup (i.e., it should be 0 days). Thus we can empirically evaluate the approach. *Rpart* performed the best with the mean estimate closest to the true value of 0. *Variance* estimated 5 for all DRGs. This is an artefact caused by the estimator having a minimum of 5 days in a segment. *Mars* estimated 20 for all the DRGs except one due to a bias towards having at least 20 days in a segment! *Vote* did not have the lowest mean estimate, however the estimate of 6 days on average is a vast improvement on the previous methods of fixed intervals of 30 or 60 days for all DRGs.

In summary, unplanned admission DRGs are a good stress test of the estimators because they involve estimating a value at the boundary of the time interval. This analysis revealed bias in estimating the null model case for all our estimators. However the bias is least for *Rpart*.

DRG	Description	Vote	Rpart	Mars	Var	Spline
50	Severe Head Injury	6	7	20	5	7
51	Moderate Head Injury	4	0	20	5	7
52	Minor Head Injury	6	0	20	5	7
471	Fracture, sprain, strain and dislocation of forearm, hand, or foot (age > 74 with CC)	9	9	20	5	14
472	Fracture ... (age > 74 without CC) or (age < 75 with CC)	6	0	20	5	14
473	Fracture ... (age < 75 without CC)	8	0	20	5	29
474	Fracture ... upper arm or lower leg (age > 64 without CC)	4	0	20	5	7
475	Fracture ... (age > 64 with CC)	8	0	20	5	58
476	Fracture ... (age < 65 without CC)	4	0	20	5	7
Mean		6.1	1.8	20.0	5.0	16.7
Std		1.9	3.5	0.0	0.0	17.1

Table 1. DRGs involving unplanned admissions. CC means Complicating Conditions.

5 Discussion and Conclusions

Data mining concepts have been used in searching for patterns of temporal intervals of care relevant to hospitalisation episodes. Data preprocessing required linking different administrative datasets to obtain data relating hospital episodes to out of hospital care. The data were then aggregated to obtain daily counts of patient services prior to and after episodes in hospital resulting in a temporal sequence. Temporal data mining is used to search for a significant change in the temporal sequence both before an episode in hospital and after discharge from hospital. The ensemble approach applies multiple techniques and combines the results through a panel of experts using an averaged voting method. The approach has been demonstrated as effective in identifying the periods of pre-admission and post-discharge services related to particular admissions. The application used to illustrate the approach employs administrative health data but the approach would be useful in many other similar situations.

Numerous assumptions have been made, such as the assumption that a single pre-admission interval and a single post-admission interval is adequate for capturing individual DRG episodes of care. This implies similar medical service utilisation patterns over patients within each DRG. For DRGs with very different age-gender mixes and different co-morbidities this may be an approximation. This initial research has identified the practicality of the data mining approach and an area for future work is to validate the results with expert advice from clinicians and to further explore issues relating to the assumptions.

References

1. Williams, G., Baxter, R., Kelman, C., Rainsford, C., He, H., Gu, L., Vickers, D., Hawkins, S.: Estimating episodes of care using linked medical claims data. In: AI 2002: Advances in Artificial Intelligence, 15th Australian Joint Conference on Artificial Intelligence, Lecture Notes in Computer Science. Volume 2557., Canberra, Australia, Springer-Verlag (2002) 660–671
2. Guralnik, V., Srivastava, J.: Event detection from time series data. In: KDD-99, Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, AAAI Press (1999)
3. Baek, J., Cho, S.: Left shoulder detection in korea composite stock price index using an auto-associative neural network. In: Proceddings of Intelligent Data Engineering and Automated Learning, IDEAL 2000, Springer-Verlag (2000)
4. Williams, G., Vickers, D., Baxter, R., Hawkins, S., Kelman, C., Solon, R., He, H., Gu, L.: Queensland linked data set. Technical Report CMIS 02/21, CSIRO Mathematical and Information Sciences, Canberra (2002) Report on the development, structure and content of the Queensland Linked Data Set, in collaboration with the Commonwealth Department of Health and Ageing and Queensland Health.
5. Kelman, C.: Monitoring health care using national administrative data collections. PhD thesis, National Centre for Epidemiology and Population Health, Australian National University, Canberra (2000) `thesis.anu.edu.au/public/adt-ANU20020620.151547/index.html`, pp 91-94.
6. Solon, J.A., Feeney, S.H., Jones, S.H., Rigg, R.D., Sheps, C.G.: Delineating episodes of medical care. American Journal of Public Health **57** (1967) 401–408
7. Wingert, T.D., Kralewski, J.E., Lindquist, T.J., Knutson, D.J.: Constructing episodes of care from encounter and claims data: some methodological issues. Inquiry **32** (1995) 162–170
8. Lestina, D., Miller, T., Smith, G.: Creating injury episodes using medical claims data. The Journal of Trauma **45** (1998) 565–569
9. Schulman, K.A., Yabroff, K.R., Kong, J., Gold, K.F., Rubenstein, L.E., Epstein, A.J., Glick, H.: A claims data approach to defining an episode of care. Pharmacoepidemilogy and Drug Safety **10** (2001) 417–427
10. Tong, H.: Non-Linear Time Series: A Dynamical System Approach. Oxford University Press, New York (1990)
11. Oliver, J.J., Baxter, R.A., Wallace, C.S.: Minimum message length segmentation. In: Research and Development in Knowledge Discovery and Data Mining: Lecture Notes in Artificial Intelligence, Springer (1998) 223–233
12. Cohen, P., Adams, N.: An algorithm for segmenting categorical time series into meaningful episodes. In: Proceedings of the Fourth International Symposium on Intelligent Data Analysis, Lisbon Portugal (2001)
13. Dietterich, T.G.: Ensemble methods in machine learning. In Kittker, J., Roli, F., eds.: Proceedings of the First International Workshop on Multiple Classifier Systems (MCS00): Lecture Notes in Computer Science. Volume 1857., Cagliari, Italy, Spinger (2000) 1–15 `citeseer.nj.nec.com/dietterich00ensemble.html`.
14. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth, Belmont, CA (1984)
15. Venables, W.N., Smith, D.M., The R Development Team: An Introduction to R. 1.5.0 edn. (2002)
16. Friedman, J.: Multivariate adaptive regression splines. The Annals of Statistics **19** (1991) 1–141