*MEDINFO 2023 — The Future Is Accessible*
*J. Bichel-Findlay et al. (Eds.)*

*© 2024 International Medical Informatics Association (IMIA) and IOS Press.*
*This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).*
*doi:10.3233/SHTI231076*

# Graphical Association Analysis for Identifying Variation in Provider Claims for Joint Replacement Surgery

James KEMP[a,1], Christopher BARKER[b], Norm GOOD[c] and Michael BAIN[a]
[a] *University of New South Wales, Australia*
[b] *Australian Government Department of Health, Australia*
[c] *Commonwealth Scientific and Industrial Research Organization, Australia*

ORCiD ID: James Kemp https://orcid.org/0000-0002-1329-6707, Christopher Barker https://orcid.org/0000-0003-2494-8587, Norm Good https://orcid.org/0000-0001-6446-7644, Michael Bain https://orcid.org/0000-0002-4309-6511

**Abstract.** Identifying potentially fraudulent or wasteful medical insurance claims can be difficult due to the large amounts of data and human effort involved. We applied unsupervised machine learning to construct interpretable models which rank variations in medical provider claiming behaviour in the domain of unilateral joint replacement surgery, using data from the Australian Medicare Benefits Schedule. For each of three surgical procedures reference models of claims for each procedure were constructed and compared analytically to models of individual provider claims. Providers were ranked using a score based on fees for typical claims made in addition to those in the reference model. Evaluation of the results indicated that the top-ranked providers were likely to be unusual in their claiming patterns, with typical claims from outlying providers adding up to 192% to the cost of a procedure. The method is efficient, generalizable to other procedures and, being interpretable, integrates well into existing workflows.

**Keywords.** Unsupervised machine learning, data mining, orthopedic procedures, national health insurance, fraud

## 1. Introduction

Fraudulent or wasteful medical insurance claims made by health care providers are costly for insurers. Fraudulent claim detection rates in the Medicare Benefits Schedule (MBS) are below international benchmarks, and better analysis methods could lower the cost of detection as well as increase detection rates [1,2].

MBS data has imbalanced and heterogeneous claim characteristics, and labels distinguishing appropriate and inappropriate claims do not exist. Given these conditions many techniques are unsuitable [3]. Moreover, since the outputs of any method must be examined by human medical experts, the results must have an intuitive, human-interpretable explanation; few methods in the literature include this aspect as a priority. As unsupervised learning is based on identifying subsets of related items in data, it can

[1] Corresponding Author: James Kemp, email: james.kemp@unsw.edu.au.

be viewed as a graph problem. We applied association rule mining using a probabilistic measure to identify edges between related items which were then combined into graphs for cost analysis, provider ranking and visualization. Association rules provide a probabilistic definition of graph edges where distance metrics may be difficult to define [4].

## 2. Methods

All claims for patients undergoing a unilateral joint replacement of the hip, knee, or shoulder were extracted for the years 2010-2014 from a previously publicly released, later retracted, dataset of MBS claims from the Australian Government Department of Health (DoH), containing all MBS claims from a random sample of 10% of patients [5].

A transaction set $T_p$ was constructed for each of the three procedures $p$. Each transaction contained all unique item codes claimed by a single provider for a single patient on a single day. Association rules (ARs) were found by evaluating support and conviction between each item pair $(x, y)$ in the transaction set [6]. We used a conviction threshold of greater than 1 (suggesting a positive association). We trialled a range of support thresholds, selecting 0.05 for the final models as a middle ground between capturing items which might be part of typical claiming behaviour, and allowing potentially wasteful items to be flagged as suspicious. If a transaction contained only one item a null item was temporarily added so that the AR mining algorithm could detect a single item as a typical transaction. ARs were combined to construct directed graphs (digraphs). A set of edges comprising ordered pairs representing the antecedent and consequent, respectively, of each discovered AR was created for each procedure subset in the dataset. The set of all items occurring as either antecedent or consequent of any discovered AR comprised the set of vertices. These graphs were used as reference models, which describe typical behaviour for providers in the procedure. Components of the graph can be viewed as representing roles in the surgery (e.g., surgeon, anaesthetist, etc.), and we assigned a label to each based on the items in the component (see Figure 1).

A digraph model was also created in the same manner for each provider. The transaction set for each provider was a subset of $T_p$ containing transactions specific to that provider. The support threshold for these models was set to 0.5, giving an intuitive indication of the provider claiming an item combination for at least half of their transactions. Providers with fewer than three transactions were excluded on the basis that no typical claim set could be identified. Providers were labelled according to the label of the reference model component which shared the most items with the provider model. A *suspicion score* was assigned to each provider by summing fees for items in the provider model which were absent from the relevant reference model. Providers were ranked by their suspicion scores. Ranking can assist auditors with prioritizing potentially non-compliant activity [3].

### 2.1. Result Validation

In consultation with two subject-matter-experts (SMEs) from the DoH, two processes were used for validation of the ranking. Due to time and resource constraints at the DoH, only claims related to providers of one of the three joint replacement procedure items (those labelled as surgeons by our model) were assessed. Firstly, a metric commonly used at the DoH to investigate compliance in surgical procedures (mean benefit of non-

procedure items, or MBNPI) was used to rank the surgical providers, then rank-biased overlap (RBO) was used to compare to our ranking, with a parameter setting of 0.9 giving 84% weight to the top 10 results in each ranking [7]. Secondly, summaries of the provider models for the ten highest scoring surgeons from each procedure were provided to a senior medical advisor (SMA), along with percentile information for the items and item co-claims as requested by the SMA. The SMA was asked to assess whether each provider warranted further investigation.

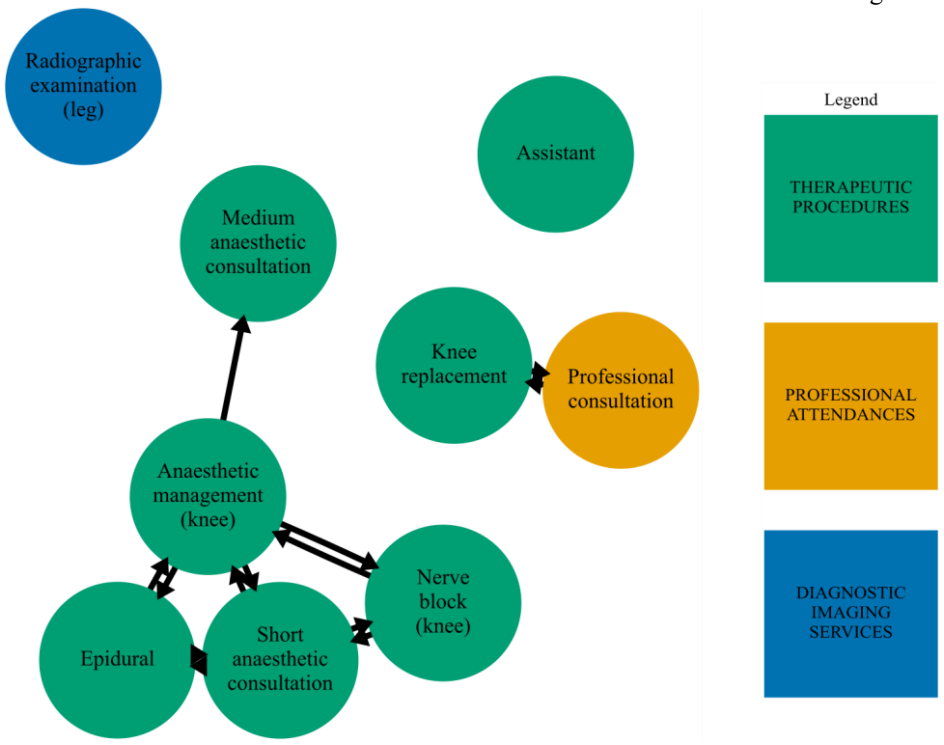


**Figure 1.** An example of a reference model for knee replacement. Vertices represent item codes and edges represent associations between items; item codes have been replaced with short descriptions here. Vertices are coloured by the item's MBS group. Graph components represent different roles within the surgery.

## 3. Results

In the shoulder and knee models the top 10 providers were all labelled as surgeons. In the hip model, 4 of the top 10 providers were labelled as anaesthetists, and the rest as surgeons. Compared with the cost of the joint replacement procedure code, typical claims from outlying surgeons added costs of up to 192%, 64%, and 82% for shoulder, hip, and knee replacements, respectively.

The RBO between the ranking of surgeons from the MBNPI and AA model ranking was 0.750 for shoulders, 0.396 for hips, 0.675 for knees. Examination of the claims from the top ten providers in the MBNPI rankings showed that some providers had wide variation in their claim patterns; that is, the typical claim model identified by the AA method was not an exact match for many of their given surgeries as they were often claiming additional items which varied across the transactions. For other providers, the

mean procedure cost was skewed upwards by one surgery with several expensive item claims, where their typical practice might not be so costly.

The SMA noted that in all cases presented, providers were co-claiming items in unusual patterns compared with their peers. If the behaviour persisted in a complete data sample, it is likely the cases would be considered actionable, although in the 10% sample the potential for cost recovery was too low. The SMA also noted that the discovered patterns of unusual behaviour were a cause for concern, and that the reference model could be useful for decision-makers in identifying a typical pattern of item claims. These comments indicate that the model is working as intended in identifying repeated unusual claims from providers.

We examined the 4 top-scoring anaesthetists in the unfiltered ranking for the hip procedures and determined that they presented similar variation from the norm to the surgeons, in that they too made item claims or co-claims in the top 2-3 percentile compared with their peers.

## 4. Discussion

As shown by the comparison of the top-ranking providers between our method and the existing method, if providers have high amounts of variation when co-claiming items, the lack of strong associations may mean that items which they commonly claim are absent from their model, due to interest measures being based on item co-occurrence rather than occurrence. It is this variation that using the typical models is designed to avoid, on the assumption that variation is inherent in medical procedures and unusual claims are sometimes expected. While providers with a high degree of unrepeated variation may be making wasteful claims, this can be more difficult to determine from claims data alone (i.e., without reference to patient medical records). In this way, the AA method can be considered applicable to repeated instances of a given wasteful behaviour.

Limitations arose from the study and model designs, as well as from the data used. Validation of the models was necessarily limited; due to the human effort involved in assessing the claims, only a small sample of providers could be assessed. The SMEs were able to advice on the variation of the providers from their peers and the process of reviewing provider claims but were unable to say whether reference model item combinations should be expected. As medical insurance claims do not necessarily represent the service provided, claims are assessed in view of accordance with the body of their peers, rather than the perspective of medical relevance.

Future work could explore incorporating additional data, or the use of other pattern mining techniques or rule types. The use of different interest measures or suspicion scores, and/or using quantity of claims as vertex or edge weights, should be investigated. Appropriate parameter thresholds may vary depending on the procedure under investigation, and more validation of appropriate thresholds would be required in any implementation of the method.

## 5. Conclusions

We found AA to be a useful form of unsupervised learning in the health fraud and waste detection context, and the combination of reference and provider models creates an interpretable tool for detecting potentially inappropriate medical insurance claims. The

implementation presented in this study provides an unsupervised method for describing variation in claims around surgical procedures, with the advantages of being human-interpretable, generalisable to other same-day procedures, and able to learn provider roles from data.

## Acknowledgements

This research is supported by an Industry PhD scholarship which includes funding from the Commonwealth Scientific and Industrial Research Organisation, the Department of Health, Australian Government, and an Australian Government Research Training Program (RTP) scholarship. Ethical approval for this study was granted by the University of New South Wales Human Research Ethics Committee Executive. Thanks to Dr. Amy Virdi from the Australian Government Department of Health for her contributions to the empirical results of this work, and to Dr. Sebastiano Barbieri and Prof. Louisa Jorm from the UNSW Centre for Big Data Research in Health, for their joint contributions to project conception and methodology, and Prof. Jorm's contributions to funding and data acquisition. Source code is available at https://github.com/jpkemp/anomaly_detection_framework.

## References

[1] Couffinhal A, Frankowski A. Wasting with intention: Fraud, abuse, corruption and other integrity violations in the health sector. 2017: 265-01,

[2] Australian Government. Budget strategy and outlook: budget paper no. 1 2017–18.: Australian Government; 2017.

[3] Ekin T, Ieva F, Ruggeri F, Soyer R. Statistical Medical Fraud Assessment: Exposition to an Emerging Field. Int Stat Rev. 2018 Dec;86(3):379-02, doi: 10.1111/insr.12269.

[4] Huang Z, Li J, Su H, Watts GS, Chen H. Large-scale regulatory network analysis from microarray data: modified Bayesian network learning and association rule mining. Decis Support Syst. 2007 Aug;43(4):1207-25, doi: https://doi.org/10.1016/j.dss.2006.02.002.

[5] Australian Government Department of Health. Public Release of Linkable 10% sample of Medicare Benefits Scheme (Medicare) and Pharmaceutical Benefits Scheme (PBS) Data: Australian Government Department of Health; 2016 [updated 11-Aug-2016; cited 2020 01-Apr-2020]. data.release@health.gov.au]. Available from: http://www.pbs.gov.au/info/news/2016/08/public-release-of-linkable-10-percent-mbs-and-pbs-data.

[6] Tan PN. Introduction to data mining. Second edition. ed. Steinbach M, Karpatne A, Kumar V, editors. New York, NY: New York, NY : Pearson Education, Inc.; 2019.

[7] Webber W, Moffat A, Zobel J. A similarity measure for indefinite rankings. ACM Trans Inf Syst. 2010 Nov;28(4):1-38, doi: 10.1145/1852102.1852106.