

Guest Editors' Introduction to Special Issue on Health Data Mining

Jim Warren¹, John Roddick², Geoff Webb³, Graham Williams⁴

¹University of Auckland (formerly, University of South Australia)

²Flinders University, ³Monash University, ⁴Australian Taxation Office

Australia has a relatively advanced health IT infrastructure. For someone close to the issues it may be tempting to focus on the negatives – such as a limited interoperability between hospital and community systems, lack of uptake of standards, and each state doing its own thing. On the positive side, however, Australia has many highlights, including a high rate and quality of Internet connectivity among providers and the public, electronic prescribing in the community as the norm, and substantial national administrative data collections. These latter two features in particular are very good news for those aiming to apply data mining techniques to health data. This excellent available material for data mining is supporting a thriving community of health data mining researchers.

This special issue of *eJHI* presents contributions from Australian health data mining researchers that illustrate a range of techniques, of applications, and of issues.

The objective of data mining is to discover knowledge. In health, the 'knowledge' may characterize a snapshot description of local conditions or describe a fundamental truth of human biology. In "Finding Anomalies in Medicare," Pearson, Murray and Mettenmeyer examine medical services of providers over a quarter to identify those with unexpected behaviours with respect to metrics of the practice. Such anomalies are often readily explained with additional domain knowledge, but may be indicators of fraud. In contrast to the identification of cases

for audit, in "Epidemiological Data Mining of Cardiovascular Bayesian Networks" Twardy et al. look at Bayesian Network (BN) learning with an eye to decision support systems development. In BN learning, one attempts to learn the best causal structure to account for observational data. Twardy et al. compare models inferred through a variety of techniques to machine-learned BNs and to a BN manually engineered from the literature in the context of coronary heart disease risk.

The temporal nature of health data – i.e., the fact that an interesting pattern is often only definable as a sequence of events over time, rather than simply as a set of attribute values at one point in time – presents an ongoing challenge for the techniques of data mining. Chen et al. present "Frequency-based Rare Events Mining in Administrative Health Data" wherein they apply association rule mining to a sequence of services over time. Their method allows the efficient discovery of associations of patterns to occurrences of events of interest, notably adverse drug reactions. Staying with the theme of temporal health data analysis, Warren et al. present "Inferring 'Therapeutic States' of Patients from Community Electronic Prescribing Data." This paper illustrates the identification of areas for clinical audit through the perusal of visual representations of a practice's prescribing on patients over time.

In "Towards Role Based Hypothesis Evaluation for Health Data Mining," Shillabeer and Roddick address the question of 'interestingness of rules'

(or, in their more precise parlance, the 'information strength of hypotheses') discovered through data mining. The choice of criteria for interestingness is a fundamental question confronted in all health data mining work, including each of the first four research efforts presented in this special issue. Shillabeer and Roddick provide a framework for consideration of the relevant hypothesis strength criteria to support the diversity of health professional roles.

Our final paper, "Research within the Privacy Regulations: Problems and Solutions for Database Custodians" by Mulligan, Rogers and Braunack-Mayer, addresses another ubiquitous issue in health data mining – the ethical and practical problems of collecting and analysing personal data. Mulligan et al. review the relevant privacy regulations with particular attention to approaches to consent. Achieving an acceptable approach to privacy remains one of the great practical barriers separating the data mining technologies from the health data.

The selection of papers in this special issue illustrates many things, which in aggregate can be reduced to two themes: (1) there is a lot of good work going on involving a multidisciplinary spectrum of technical, clinical and social perspectives; and (2) in most cases there is still a significant distance to go before the work will deliver real benefits toward the efficiency and quality of healthcare delivery. I hope the reader is stimulated by the opportunities and challenges of health data mining.