Leveraging Big Data Analytics to Reduce Healthcare Costs

Uma Srinivasan and Bavani Arunasalam, *Capital Markets Cooperative Research Centre, Australia*

Two novel applications for analyzing health insurance claims leverage big data to detect fraud, abuse, waste, and errors. Claim anomalies detected using these applications help private health insurers identify hidden cost overruns that transaction processing systems can't detect.

In Australia, private health insurers (PHIs) process claims using systems that have built-in validation techniques to detect invalid billing items. Most of these systems tend to focus on each claim individually, without considering the other claims involved in an "episode of admitted patient care"—that is, the time interval between a hospital admission and departure. Furthermore, these systems can't adjudicate on matters that might constitute abuse and waste. Currently, there's no consistent and transparent way for PHIs to show their clients the "value" they receive for paying additional charges and premiums for private health care services. It's thus essential for insurers and other healthcare funders to have sophisticated analytical systems that can identify cost overruns that might constitute fraud, abuse, waste, and errors, so they can provide a set of uniform comparative measures for determining quality of care provided.

A range of analytical solutions exist, but few provide an operational interface that lets claims adjudicators and healthcare advisers view all current and past claims to help them understand why claims or providers have been flagged as requiring investigation. We've thus developed two applications that provide not just effective analytics but also rational explanations for alerts that help facilitate appropriate action. The applications use large volumes of complex codified and free text data extracted from claims and hospital discharge data to identify claims (or claiming patterns) that represent fraud, abuse, waste, and errors. For a typical medium-sized insurer, our applications analyze 25 million claims related to 1.2 million members.

1520-9202/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

The first application, CMC-I+Plus, provides advanced performance analytics using claims-scoring and predictive-modeling techniques applied to hospital and medical claims data. The second application, CMC Health Insurance Business Intelligence Services (CMC-HIBIS), applies a combination of leading-edge business rules and business intelligence technologies to hospital, medical, and ancillary claims data to identify fraud, waste, abuse, and errors and to generate alerts, along with relevant explanations, that the claims-processing and risk-compliance staff can understand.

Healthcare in Australia

In most developed countries, the healthcare sector deals with very large volumes of electronic health data related to patient services. Most primary data is created and stored by health services providers, including general practitioner doctors, specialists and surgeons, public and private hospitals and clinics, support services providers (such as pathology and x-ray technicians), and health professionals (such as physiotherapists and optometrists). However, some data is passed to the funders—that is, government agencies and insurers.

In Australia, healthcare funding for a population of 22 million is provided by government and private entities. Current cost estimates exceed US$120 billion per annum.[1] The government entities (which fund approximately $100 billion of the cost) include the national insurer, Medicare, the related Pharmaceutical Benefits Scheme (PBS), and the State and Territory Governments in relation to public hospitals and clinics. PHIs and accident compensation insurers represent the next largest group of healthcare funders. They allow customers to avoid waiting lists, provide them with choice of hospitals and doctors, and provide ancillary coverage for items not covered by Medicare (such as dental services).

Citizens provide significant funding for their own healthcare through direct payments to providers and through out-of-pocket contributions when public and private health insurance schemes don't cover the full cost. Every time a person receives a health service—other than one provided in a public hospital—the service provider sends a claim to the appropriate funder. The claim usually provides details about the service and its cost. Episodes involving hospitalization often generate claims from both the hospital and one or more doctors. Electronic health data is thus widely distributed across service providers and funders.

All Australians are entitled to Medicare service and have a Medicare number, which is used for billing purposes and to claim government-subsidized Medicare services. By law, the Medicare number can't be used to link an individual's health data across services. To pay for Medicare services, individuals entitled to Medicare contribute a percentage of their salary to compulsory Medicare through the Australian tax system. Private health insurance is optional, and the premiums are paid by the individual directly to the insurer. To reduce the burden on the public health system, the government offers rebates to encourage more people to purchase private health insurance.

Australian PHIs cover approximately 45 percent of the population and process data related to over three million acute hospital episodes, 27 million medical services, and 73 million ancillary services.[2] Most of this data is in the form of hospital, medical, and ancillary claims, submitted by service providers. Individual insurers allocate their own unique member ID for their members. Therefore, all claims and hospital data submitted by service providers to the insurer have this unique member ID specified in the claim.

Most people in Australia use a mixture of public and private healthcare. An individual uses the Medicare number to use the government public health services and the PHI member ID while using private health insurance services. These data sets currently aren't linked, even though most public and private hospitals, physicians, and laboratories all have electronic health data. The new person-controlled electronic health record (PCEHR) initiative (www.ehealth.gov.au), currently being rolled out, is designed to enable the sharing of electronic health data scattered across the system with the express permission of the individual.

To understand the full significance of the knowledge embedded in health insurance claims data, we must understand the myriad forms of complex relationships that emerge when claims data is positioned in a much broader context of overall healthcare. Governments and PHIs have realized that detecting errors and fraud in

Table 1. The different coding schemes used by the Australian health sector.

Coding scheme	Purpose	Users
Commonwealth Medical Benefits Schedule (CMBS)	Provides clinical procedure codes used to bill for services performed (the cost of these services is listed under the government Medicare Benefits Schedule—see www.mbsonline.gov.au)	Doctors and hospitals
International Classification of Diseases (ICD-10)	Provides diagnosis and procedure codes documented in hospital medical records and specified in the hospital case-mix protocol (HCP) data that is sent by the hospitals to the private health insurers (PHIs)	Public hospitals (for disease classification and reporting) and private hospitals (for case-mix reporting)
Diagnosis related groups (DRG)	Provides a clinically meaningful way of relating the types of cases treated in a hospital to the resources required by (and therefore the costs incurred by) the hospital	Hospitals claiming money from funders (where the remuneration is based on case payments)
Health insurance claims and payments system (HICAPS—a generally used point-of-service electronic claiming solution)	Provides treatment codes used for electronic claims by ancillary service providers	Ancillary service providers such as physiotherapists, dentists, or optometrists
Prosthetic codes	Helps the Therapeutic Goods Authority approve prosthetics that can be used by the Australian healthcare industry	Hospitals, Medical providers
Type of Occurrence Classification System (TOOCS) workers' compensation coding system (Australia)	Provides codes for injuries and illness used by Workers' compensation agencies to code disease processes (largely based on ICD 9 and ICD 10, with extensions as required to indicate the nature and type of accident and injuries)	Workers' Compensation schemes and insurance agents

healthcare claims, as well as identifying episodes that constitute abuse and waste, is essential to ensuring the economic sustainability of the health system and optimal outcomes for citizens. A vital part of achieving both objectives depends on the funders having advanced analytics.[3] The impact of business intelligence and big data analytics on the industry in general and the health sector in particular is well documented.[4]

Healthcare Datasets and Coding Schemes

Australian PHIs and the accident compensation insurance sector use coding schemes to process bills and claims and create regulatory and mandatory reports. Table 1 shows the different coding schemes and how they're used in the Australian health sector. We didn't include the Systematized Nomenclature of Medicine (Snomed) coding system, because it's still in the early stages of terminology trials and has yet to be used in a billing or hospital clinical record system.

In Australia, PHIs receive rich data sets from providers. The hospital and medical claims processed by a PHI contain data that specifies the type of service provided and the cost of that service. The service is specified as a Commonwealth Medical Benefits Schedule (CMBS) code (www.mbsonline.gov.au), as stipulated by the Australian Government. However there's no cap or upper bound on the fee charged by a physician, resulting in out-of-pocket expenses for the patient. In general, insurers will link CMBS codes to a service charge indicating the cost of that service.

The hospital case-mix protocol (HCP) reports[5] submitted by hospitals contain clinical data that includes the ICD-10 diagnosis codes (www.icd10data.com), Diagnosis Related Group (DRG) codes,[6] the length of the hospital stay, the type of prosthesis used, and details about related patient comorbidities. The HCP data related to a specific admission contains the member ID and is received by the PHI only after the patient is discharged from the hospital. Combining business data in the claim forms with clinical data in the HCP reports gives us the capability to calculate several nationally accepted indicators[3] developed by Australian Institute of Health and Welfare (AIHW),[7] and analyze key aspects of health economics such as efficiencies, health outcomes, and clinical performance indicators related to the quality of care.

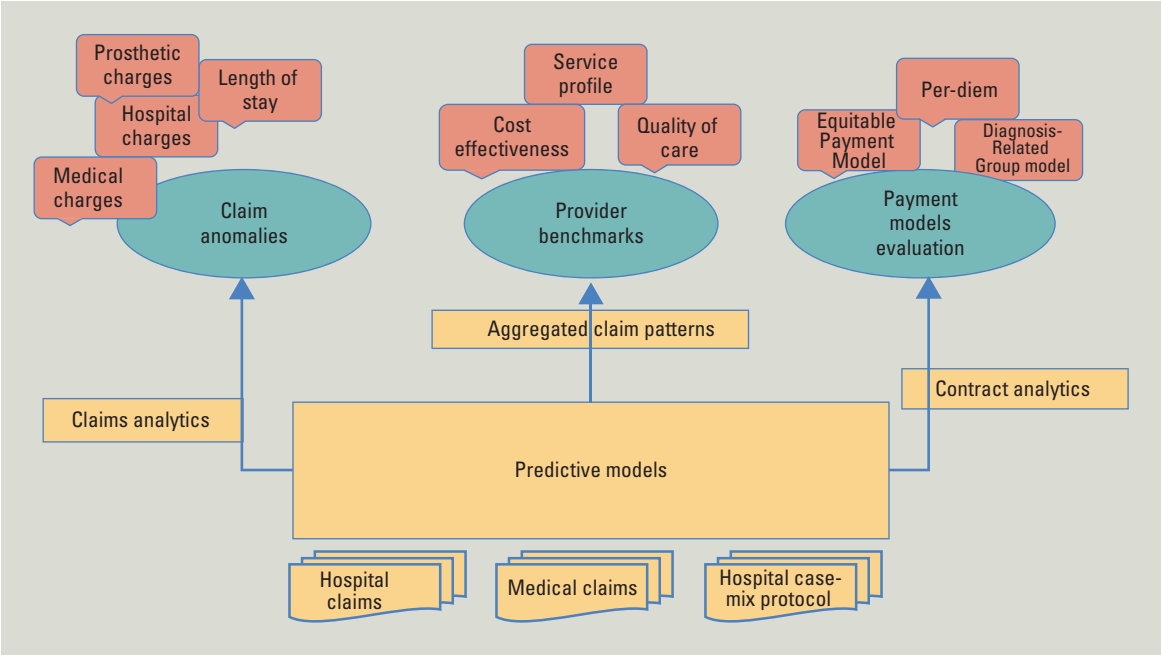

Figure 1. An overview of CMC-I+Plus. The application provides comprehensive claims-based intelligence to help detect and investigate potential anomalies in health claims and related data.

For any given hospital admission, we deal with three sets of claims data. The first is *medical claims*, sent by doctors who performed the service while the patient was in hospital. The next set is *hospital claims*, sent by the hospital's billing department. A claim can include several claim lines, each of which indicates a specific service cost, an accommodation cost, a prosthetic cost, a laboratory investigation cost, a pharmacy drug cost, and so on. The final set of claims data is the *HCP data*, which presents a consolidated summary of the patient's clinical care during that particular episode or admission.

The claim line is the basic unit of data used in our analysis, and there are two broad types of claims: clinical claims and ancillary claims. This broad division is based on the way the health insurance products are structured. Clinical claims include claims from both hospitals and doctors. Ancillary claims include claims from service providers such as dentists, optometrists, physiotherapists, occupational therapists, speech-pathologist, and dieticians.

CMC-I+Plus

I+Plus is an advanced analytical solution that provides comprehensive claims-based intelligence to help detect and investigate potential anomalies in health claims and related data (see Figure 1). These anomalies could be related to costs (such as hospital accommodation costs, costs of prosthetics, or medical fees and charges) or quality of care (such as increased length of stay, unexplained infections, unplanned readmissions, or excessive services or supplies). An important aspect of I+Plus is the ability to both generate and report on these anomalies at a detailed admission level, and aggregate these anomalies at a higher level of abstraction to provide comparative performances of service providers and treatments, from both financial and quality-of-care perspectives.

The core of the I+Plus system is a set of predictive models developed using advanced statistical and data mining techniques that run through large volumes of historical healthcare data. These predictive models help predict hospital, medical, and prosthetic costs and the length of stay in the hospital for each individual hospital admission. Consequently, the models identify anomalous admissions that significantly deviate from the predicted values.

Because healthcare data arrives in different forms from different sources, selecting the appropriate set of features (key elements) from this integrated data set is the key to accurately modeling behavior. I+Plus predictive models are based on two major types of features: those that are directly extracted from the claims data and

those that are computed using a combination of domain knowledge and data mining and statistical techniques. (A detailed description of these techniques is outside the scope of this article.) The features include member demographics, such as age and gender, and provider-related features, such as the type of hospital and the location. They also include admission-related features, such as the procedures performed (ICD-10 and CMBS codes), comorbidities diagnosed, and additional services performed (such as imaging and pathology tests conducted during the hospital admission).

The I+Plus system architecture readily integrates data from all these different sources and automatically selects only the key features that have a significant impact on the target variable we're seeking to predict. Consequently, the predictions made by the I+Plus models are fine-tuned to consider features of the admission, such as the primary treatment code, additional procedures carried out, comorbidities, patient age and gender, and type of hospital (public versus private).

Furthermore, the I+Plus system consists of distinct predictive models for different types of treatments such as hip-related procedures and cardiology procedures. This further improves the model's accuracy, because different features can have varied levels of impact on different treatments. I+Plus follows the treatment classification structure used in the Medicare Benefits Schedule, with models developed at the most appropriate level (node) in the hierarchy (for example, "single knee replacement" appears in the "orthopedic" category, which falls under "surgical operations.") At each node, models are created to predict hospital, medical, and prostheses charges and the length of stay. The result is thousands of models tailored to each primary procedure type and each target variable. When a new claim arrives, based on the principal treatment, the system automatically chooses the most appropriate model and predicts the expected values of the target variables.

Our predictive models are mainly based on regression analysis. The reliability of a regression model is given by a metric called *R-squared*, which provides a measure of accuracy of the predictive models. In our system, we only use models with R-squared higher than 0.6 to ensure reliable prediction. Results from the I+Plus models can also be used at an admission level in aggregated forms at a provider, member, and treatment level.

Admission-Level Analytics

I+Plus identifies and highlights anomalous admissions, such as hospital charges that are significantly higher than expected without any indication of complications or adverse events. It also looks for instances of over-servicing, such as when the number of pathology tests or imaging services performed is significantly higher than numbers noted for similar admissions for the same treatment.

The I+Plus user interface facilitates a detailed investigation of such anomalous admissions by bringing all the available information for a given admission onto one page and providing the user with a complete view of the admission with which to conduct a detailed investigation. The system also lets users compare a selected (anomalous) admission, on a feature-by-feature basis, with a similar admission that aligns with the benchmark.

Aggregated-Level Analytics

The I+Plus system provides aggregated results so users can compare the performance of providers from different perspectives, such as cost effectiveness and quality of care, and to detect atypical patterns at the provider, member, and treatment level. Examples of such atypical patterns could be a provider performing an investigative procedure that's usually not performed by other providers for the same treatment or a provider using a specific, high-cost medication not used by other providers.

Contract Analytics

I+Plus provides comprehensive intelligence to support provider contract negotiations and provider performance benchmarking. The interface lets the user compare the claiming patterns of different providers and provides detailed information for decision making.

CMC-HIBIS

The CMC-HIBIS application is based on a comprehensive and ever-expanding rule base that generates alerts for suspected fraud, waste, abuse, and errors (see Figure 2). The solution

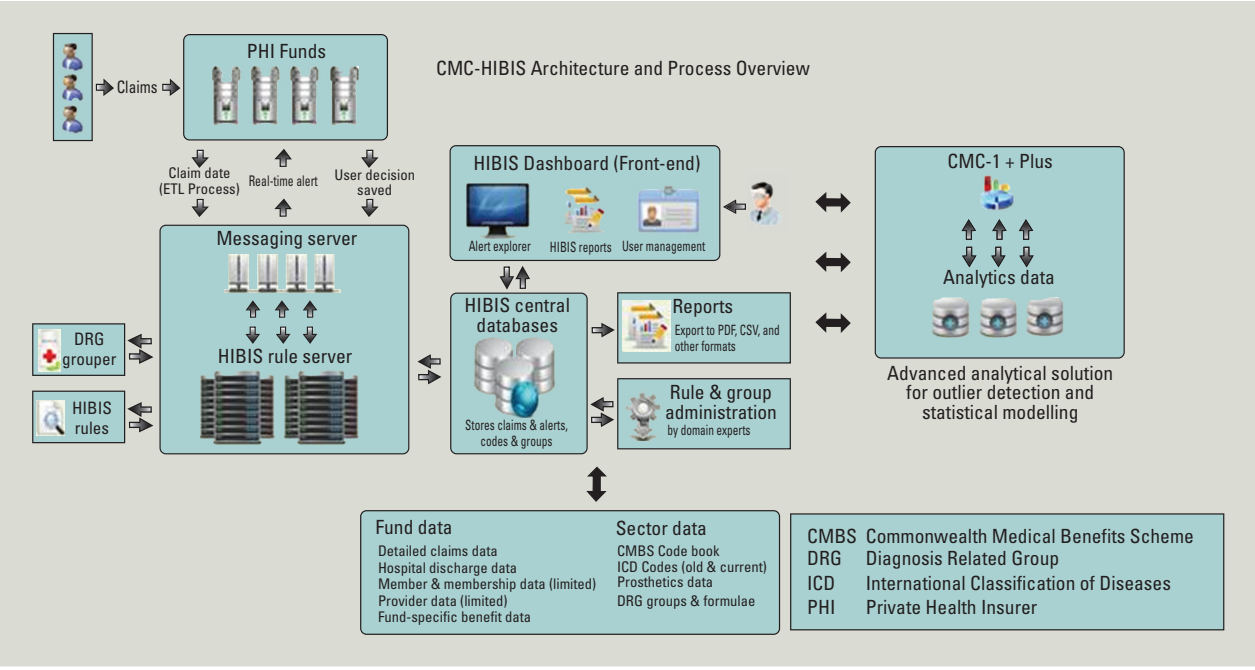

Figure 2. A schematic diagram of CMC Health Insurance Business Intelligence Services (CMC-HIBIS) and CMC-I+Plus. Components of the CMC-HIBIS application use a comprehensive and ever expanding rule base managed by the HIBIS rule server that generates alerts that identify suspicious claims. The HIBIS dashboard is the front-end that enables claims analysts to review the alerts. CMC-I+Plus is the predictive-model-based analytical system for detecting and predicting potential anomalies and benchmark services and providers.

enables the claims-processing and compliance staff to review the claims and, where appropriate, refuse or reduce payment, or, where payment has already been made, secure refunds.

Experts with deep knowledge in focused areas, such as clinical claims processing, billing, fraud investigations, and clinical coding capabilities, have worked together with healthcare domain experts in developing and specifying the rules and alerts.

The Alert Explorer is an important component of CMC-HIBIS, and is the visualization tool that lets insurer staff audit contentious claims based on the type of alerts raised. When it's positioned in the context of the workflow in claims and compliance departments, the Alert Explorer provides an extensive drill-down capability and allows for the recording of actions taken (such as rejecting a claim or overriding an alert) and results achieved (such as receiving a refund).

The Alert Explorer lets users

- search for and view alerts showing potential fraud, abuse, waste, and errors;
- filter the view of alerts based on specific search criteria;
- review the details of the claims relating the alert to the member's medical history;
- review the details and summary of alerts related to a specific provider; and
- manage the entire workflow of the alert-claim recovery process.

Here we provide some illustrative examples of alerts.

Our first example is an alert raised when a hospital charges for excessive stents not used in the procedure. In this claim, four cardiac coronary stents were charged, which is inconsistent with the surgeon and discharge data. The cost overrun due to this one erroneous claim is over $16,000, because each stent costs around $4,000.

Our next example is an alert raised when a total knee replacement was charged twice in one week. The first procedure should insert the prosthesis in both knees, but the second procedure, if performed within a week, should normally be a revision procedure, which costs a lot less than a repeat of the prosthetic insertion procedure. The cost overrun in this case is over $5,000.

Another example is an alert raised when a procedure performed is unrelated to the principal

diagnosis. Because this is a more generalized alert that can be raised under many incompatible diagnosis and procedure combinations, its overruns depend on the diagnosis and procedure codes causing the anomaly. The cost overrun in this case can range from $5,000 to $8,000.

The final example is a theatre fee that exceeds the corresponding claim by the surgeon. (Theatre fees are just one component of the overall hospital costs—they're based on the complexity of the surgical procedure.) Cost overruns for this alert are approximately $5,000, depending on the procedure performed.

More details about the CMC-HIBIS system appear elsewhere (www.cmc-is.com/index.php/main/solutions#hibis).[8]

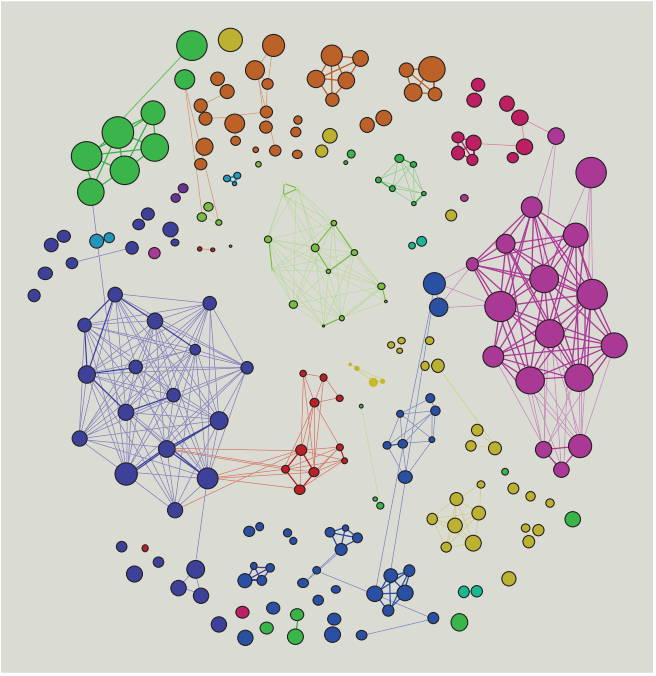

Figure 3. The pattern of sharing patients by a community of physiotherapy providers. Each cluster of nodes indicates a close-knit community of providers who share common patients. A large size node with a thick edge connecting it to one specific node indicates that a provider with many patients is sharing a large number of his or her patients with one specific provider.

Creating effective analytics that produce actionable intelligence is particularly difficult for healthcare claims, given the enormous complexity and variety of data and the many human factors that can lead to anomalies. However, over large populations and time frames, effectively managing and analyzing this quintessential big data could deliver significant financial and healthcare benefits.

Our results indicate that claim anomalies detected using these applications are enabling private health insurance funds to recover hidden cost overruns that aren't easily detected using traditional transaction processing systems. Due to confidentiality agreements, we can't share the value of refunds, but our clients are generally receiving a good return on their investment. In addition, the I+Plus system offers metrics and filters to help insurers negotiate contracts with service providers, monitor the quality of care, and benchmark provider performance, thus contributing to well-defined and acceptable standards of care in the health sector.

Our I+Plus work has given us a good understanding of healthcare claiming patterns of hospital and medical providers. Our continuing research moves on from that understanding to explore complex relationships within and across health claims over time. We've started using social network analysis techniques to study provider-member relationships and behavior, analyzing provider communities, which frequently share a large volume of patients, to identify over-servicing and other behavioral factors associated with the health insurance sector.

Figure 3 shows a network of physiotherapy providers who share common patients. The nodes represent the providers, and the edges represent the number of patients shared in one week. The size of the nodes indicates the number of patients handled by that provider, and a thicker edge indicates a higher number of patients shared between the providers in one week. Large clusters indicate a close-knit community of practitioners frequently sharing patients. The idea is to explore whether there are any unusual behavior patterns exhibited by a group of providers frequently sharing large numbers of patients. This area of work based on network analytics has yet to be evaluated by our user community.

Another area we are exploring is text mining. Healthcare data contains various text descriptions that include useful information regarding the treatments and procedures that can be used to detect over-servicing and other useful information.

Acknowledgements

We wish to acknowledge the contribution of the entire CMC-HIBIS and CMC-I+PLUS development teams and many key staff from our clients who have translated the research ideas into successful applications. We particularly wish to thank David Jonas, CEO, CMC Insurance Solutions, for his support and encouragement.

References

1. "Health Expenditure Australia 2011–12," Australian Inst. Health and Welfare, 25 Sept. 2013; www.aihw.gov.au/publication-detail/?id=60129544658.
2. *Private Health Insurance Australia: Quarterly Statistics*, Australian Government, June 2013; http://phiac.gov.au/wp-content/uploads/2013/08/Qtr-Stats-Jun13.pdf.
3. S. Barret, "Insurance Fraud and Abuse: A Very Serious Problem," *Quackwatch*, 10 Jan. 2006; www.quackwatch.org/02ConsumerProtection/insfraud.html.
4. H. Chen, R.H.L. Chiang, and V.C. Storey, "Business Intelligence and Analytics: From Big Data to Big Impact," MIS Quarterly, vol. 36, no. 4, 2012; http://ai.arizona.edu/mis510/other/MISQ%20BI%20Special%20Issue%20Introduction%20Chen-Chiang-Storey%20December%202012.pdf.
5. "Hospital Casemix Protocol (HCP)," Australian Government—Department of Health, May 2013; www.health.gov.au/internet/main/publishing.nsf/Content/health-casemix-data-collections-about-HCP.
6. "Round 12 (2007-08) Cost Report—Public Version 5.1, Private Version 5.1 and Private Day Hospital Facilities (Standalone) Version 5.1," Australian Government—Department of Health, Dec. 2012; www.health.gov.au/internet/main/publishing.nsf/Content/Round_12-cost-reports.
7. "A Set of Performance Indicators for the Health and Aged Care System," AIHW Report, June 2008; www.aihw.gov.au/indicators/index.cfm.
8. "Health Insurance Business Intelligence and Claims Leakage," white paper, CMC Insurance Solutions, July 2012; www.cmc-is.com/index.php/main/research.

Uma Srinivasan *is a lead scientist in Health informatics at Capital Markets Cooperative Research Centre, Australia. Her experience in designing software solutions for hospitals and health departments has shaped her research focus on information solutions that enable efficient and high quality healthcare. Srinivasan holds a PhD in computer science from University of New South Wales, Australia. Contact her at uma.srinivasan@cmc-is.com.*

Bavani Arunasalam *is a lead scientist in data mining and analytics at Capital Markets Cooperative Research Centre, Australia. She is an Honorary Associate of the University of Sydney and has extensive experience in risk identification, fraud detection, and predictive modeling using advanced data mining, text mining, and statistical techniques. Arunasalam received her PhD in computer science from Sydney University. Contact her at bavani.arunasalam@cmc-is.com.*