Nonconfidential Patient Types in Emergency Clinical Decision Support

Mark Chignell, Mahsa Rouzbahman, Ryan Kealey, and Reza Samavi | University of Toronto
Erin Yu | Canadian Imperial Bank of Commerce
Tammy Sieminowski | Bridgepoint Hospital in Toronto

Tools that show similar patients' diagnoses and treatment trajectories might provide useful clinical decision support for emergency physicians who use a case-based reasoning approach. However, privacy concerns that arise with indirect use of electronic health records must be addressed.

The ability to explore similar patients' diagnoses, treatments, and outcomes might help physicians with differential diagnosis and treatment planning. However, because of privacy concerns, getting access to individual patient records is difficult within a hospital and seems almost impossible for researchers outside the firewall.

There's a trade-off between powerful personalization of services and the cost of disclosing personal information.[1] Legitimate privacy concerns hamper research innovations because researchers can't use relevant datasets unless certain de-identification and restricted access policies are enforced. The US Health Information Portability and Accountability Act's (HIPAA's) safe harbor standard requires removal of 18 variables including name, telephone number, and mailing address. However, following these de-identification policies might not substantially reduce the risk of re-identification. Several proposals (such as "Methods for the De-identification of Electronic Health Records for Genomic Research"[2]) have focused on algorithms that can mitigate re-identification concerns.

To utilize relevant knowledge locked up in large confidential datasets without violating privacy, we created a method that exports this data in a form that could be used ethically in general recommendation and personalization systems as well as in specialized emergency clinical decision support systems. Our approach follows the de-identification process mandated by privacy law but mitigates de-identification risks via a novel data summarization method—instead of removing more identifiable attributes, we remove individuality by grouping individuals into types.

Clinical Decision Support Using Similar Patients

Our previous research found that emergency physicians had little time to search for clinical evidence when making decisions.[3] Traditional sources of clinical evidence aren't well-suited to the demands of real-time clinical decision making,[4] and physicians frequently resort to a general-purpose search engine, such as Google, when looking for clinical information.[5] As Benjamin Hughes and his colleagues noted, "Although credibility of information was the most cited concern, tools such as Wikipedia or Google are used 3 times more in our

 Copublished by the IEEE Computer and Reliability Societies

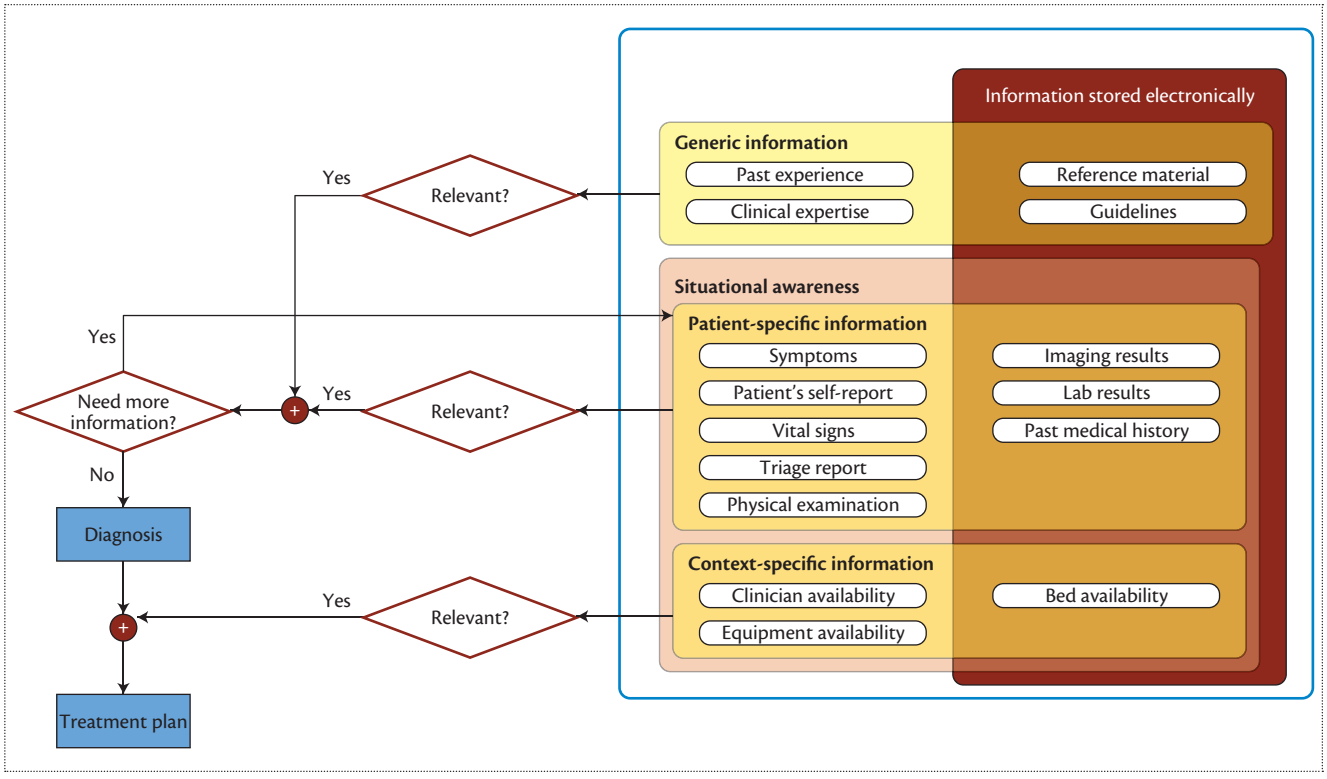

Figure 1. Model of diagnostic decision making in the emergency department.[11] The physician's mental model is composed of a combination of generic information and situation-specific information about the patient and the current context (for example, bed and equipment availability). Diagnosis and treatment are guided by the physician's evolving mental model.

sample than PubMed, the 'official' best evidence tool introduced in medical school."[6] Physicians' decisions are influenced by previous cases they've seen; this case-based reasoning is particularly true for older doctors, who are generally less comfortable using search technologies.[7] Data presentations of similar patients can support clinical decision making[8] and can be done via a background search process that doesn't require additional cognitive effort from the physician.[9]

Although promising, this approach faces significant privacy problems. Again, access to individual patient data is restricted even in the hospitals that store the data, and is extremely problematic when data is exported to outside researchers. So, rather than show individual patient records that raise privacy concerns with each access and viewing, we propose abstracted views of electronic health records (EHRs)—as patient types—that let physicians explore hypotheses about the current patient using nonconfidential information.[10]

To understand how emergency physicians currently make decisions and what kind of support they have, Erin Yu carried out a series of semistructured interviews and observational studies with emergency department physicians at a major Canadian hospital.[11] She collected and analyzed qualitative data on their day-to-day tasks, patient information management, and problem areas. She observed them in their natural work environment in an unobtrusive manner and occasionally asked probe questions about the tasks performed. Yu observed that notifications about abrupt changes in patient condition or a new patient's arrival were delivered by a nurse in person or by a page. Physicians continually checked electronic medical records systems to see whether new information on their patients had arrived. To collect all the information they needed about a particular patient, physicians were required to log in to multiple systems and locate specific records. They also experienced constant interruptions, such as page calls, phone calls, consult requests, patients demanding prompt care, arrival of sicker patients, and test results, which create a burden on working memory and increase the possibility of error. Given the cognitive complexity of the emergency physician's role, it was concluded that any clinical decision support system must be easy to use and cognitively undemanding.

Figure 1 shows information types relevant to emergency physicians: general clinical knowledge gained from formal training, literature, and past experience; patient-specific information gathered from the patient's chart, other clinicians, physical examinations, and conversations with the patient; and various test results.

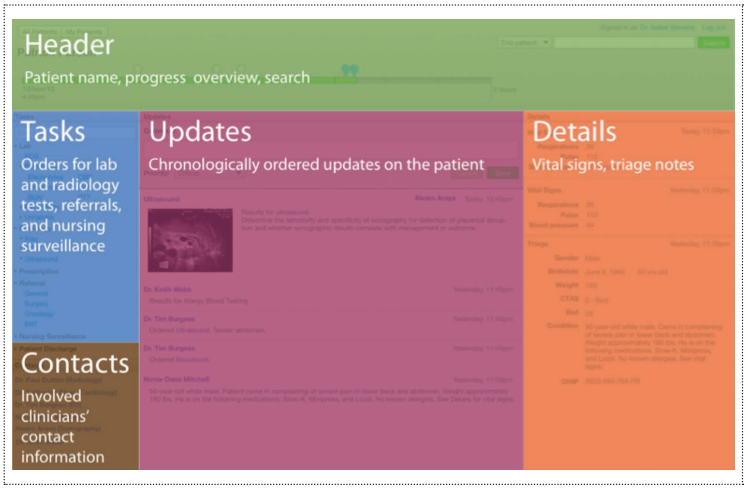

Figure 2. Sample summarized patient record layout. In this case, the summarized record describes typical tasks, updates, and details for a particular patient type.

Physicians combine general and patient-specific clinical information on a case-by-case basis to arrive at a diagnostic decision.

Based on our observations of, and discussions with, physicians at work, we believe that data used to support clinical decision making is best represented in the form of a patient record. Figure 2 shows the general layout for a summarized patient record representing a patient type or *cluster*. An additional section could be added to show diagnoses and outcomes for a particular cluster type.

EHR repository searches are challenging because the data is complex and queries are difficult to formulate. Matching a current patient to a set of abstracted types can be done automatically, avoiding the need for difficult manual searching. But how can we create abstracted data types that capture meaningful patterns in the data?

Data Summarization Methodology

Abstracting patient types requires data mining and machine learning with techniques such as clustering and classifying.[12] A variety of cluster analysis methods can be used in data mining and abstraction.[13] To summarize patient data, we select relevant features in the database or repository, find similar patients through iterative data clustering, summarize the clusters, and interpret the clusters in terms of the patient types that they represent. Ideally, clustering should involve both numerical and textual descriptors to find the most salient cluster descriptions. Because text labels can be particularly problematic from a privacy perspective, we use automated indexing methods to identify and screen out terms that are highly indicative of, or associated with, a particular person or group.[14] We call this method *anti-indexing*—a form of de-identification—because instead of labeling documents or records with key identifiers, we explicitly remove those key identifiers.

Our data-mining approach begins with feature discovery using methods such as principle components analysis and factor analysis. Once we derive the features—for example, sets of drugs that are used to combat particular conditions or lab tests that are ordered to distinguish between sets of competing disease alternatives—domain experts critique them, and we carry out further rounds of feature discovery as required.

We then convert feature values to z-scores, forming a standard normal distribution to ensure that clustering results aren't influenced by differences in scale (for example, meters versus centimeters). We use a variety of algorithms, including K-means, hierarchical, and fuzzy clustering, to find the best clusters of similar patients, which again are evaluated by physician experts. Depending on the results, further rounds of clustering might be required.

To find *superclusters*—parent clusters that contain a set of subclusters—we perform factor analysis on the matrix of cluster centroids (that is, each cluster is treated as a single observation or vector of data for the factor analysis). Each extracted factor represents a possible supercluster. We identify the unique features for each cluster or supercluster through a series of differencing operations. For a supercluster, we subtract the entire sample's centroid from the supercluster's centroid. For clusters within superclusters, we subtract the parent supercluster's centroid from the cluster's centroid. For clusters that aren't within a supercluster, we subtract the centroid of all data that isn't within a supercluster from the cluster's centroid. This process results in a set of differenced centroids—in z-score units—one for each cluster or supercluster. Different z-score cut-off points can be used; in our research, we used an absolute differenced z-score of 2.5 or higher as indicative of discriminating features for a cluster (corresponding to a difference that has less than a 1 percent chance of occurring if the feature is not actually distinctive for that cluster).

Our method currently produces a two-level structure of superclusters and clusters. For larger datasets, a multilevel hierarchy of clusters might exist. In this case, our method can be applied recursively, subjecting a large cluster's data to a similar analysis and breaking it down into its own set of clusters. In this way, our method can be generalized to very large datasets with multilevel clusters. Once the cluster formation process is complete, clusters are summarized in terms of their means and standard deviations on the features, along with the matrix of intercorrelations (within the cluster) for the features. This information can then be used in later regression analyses that predict the value of unknown features for a new patient who matches a given cluster.

Then, in the final step, the clusters are summarized in terms of the mean values on their discriminating features. The clusters are then interpreted and labeled by physicians based on the summarized information for each cluster.

Identification of Patient Types in the MIMIC II Database

To demonstrate our approach, we used the MIMIC II database, available from www.physionet.org. Beth Israel Deaconess Medical Center in Boston provided data from its ICU information systems, hospital archives, and other external sources. We developed a set of abstracted data types for a sample of approximately 25,000 adult ICU patients. The data covers approximately 28,000 hospital admissions and more than 40,000 ICU stays. The database contains 38 different tables; among those, we chose to use the medication, lab, chart, demographics, International Statistical Classification of Diseases and Related Health Problems (ICD) codes, and ICU stay tables for the first round of data analysis.

To create a single data table containing one row per patient, we removed infrequently occurring variables referring to fewer than 2 percent of cases. We also removed outliers (for example, medication doses more than five standard deviations above the mean are likely data entry errors) and replaced them with the mean and collapsed multiple hospital and ICU admissions per patient into a single summary record, averaging their data. Prior to analysis, we normalized each of the variables, using *z*-scores, to ensure scale comparability. For feature extraction, we used three methods—factor analysis, discriminant analysis, and hierarchical clustering—and we carried out the analyses using IBM SPSS (statistical package for the social sciences).

Using factor analysis with Varimax rotation, we extracted meaningful linear combinations of variables from medication variables, lab variables, chart variables, and ICD9 code variables. We used Cronbach's alpha to assess each extracted factor's internal reliability and filtered out those variables that lowered the internal reliability. We then summed (with equal weighting) the remaining variables and divided by the number of variables to create a scale[15] corresponding to the original factor. Because we found only a limited number of extracted features for medication using factor analysis, we also used discriminant analysis to search for additional features. We then used hierarchical clustering to look for clusters of relevant diagnosis (ICD9) features based on patients' diagnostic codes. However, we didn't find strongly interpretable clustering among the ICD codes and thus decided to group them based on their general diagnostic categories in the ICD classification system.

We formed scales for medications, lab tests, and chart values, then two physicians—a family physician and a general practitioner—reviewed them for meaningfulness. We obtained one interpretable medication scale, with a Cronbach's alpha of 0.9, which included the following medications: integrelin, vasopressin, epinephrine-k, lasix, labetolol, and milrinone. From lab test variables, we extracted 10 features with a Cronbach's alpha higher than 0.7. We also extracted two scales from the chart entry variables, each with a Cronbach's alpha higher than 0.8. After identifying the medication, lab test, and chart entry scales, we added them to the data matrix to develop the patient type clusters. To avoid data redundancy, whenever we added a feature, we removed the original variables that we had combined to form the feature.

We carried out *K*-means clustering on the resulting data matrix, varying the number of extracted clusters between 3 and 50. We found that a 40-cluster solution had the best combination of a large number of clusters and a sizeable group of people in each cluster. With this solution, we focused on 27 clusters that each had at least 100 patients. Eight of the clusters contained more than 1,000 patients; the largest had more than 3,000. After factor analysis of the clusters, we found that seven of the original clusters formed a supercluster. We used difference vectors to interpret the clusters. Here, we summarize the results for two subclusters within the supercluster. A more detailed report of the results is available from the authors.

The first subcluster was strongly related to infectious and parasitic diseases, and diseases of the blood and blood-forming organs, and was less related to sense organ and endocrine diseases, nutritional and metabolic diseases, and immunity disorders. The cluster had relatively high levels of respiratory alarm, Ferritin, and white blood cells. A physician on our team identified this cluster as a possible respiratory cluster.

The second subcluster was strongly related to sense organ diseases and weakly related to most other ICD categories in the supercluster. Relative to the parent supercluster, this cluster also had high levels of Nitro-Glycerine-k and respiratory alarm and low values of oxygen saturation alarm and arterial blood pressure. The physician interpreted this subcluster as a cardiac cluster.

The results of our initial exploratory data clustering show that we obtained meaningful clusters that capture some clinically relevant combinations of features. We judged seven of eight supercluster features as likely clinically relevant in an ICU patient. In the two subclusters we highlight here, we judged more than 70 percent of features listed as likely to be clinically relevant (5 of 7 and 7 of 9, respectively) and identified potential interpretations of these clusters as a respiratory cluster and a cardiac cluster.

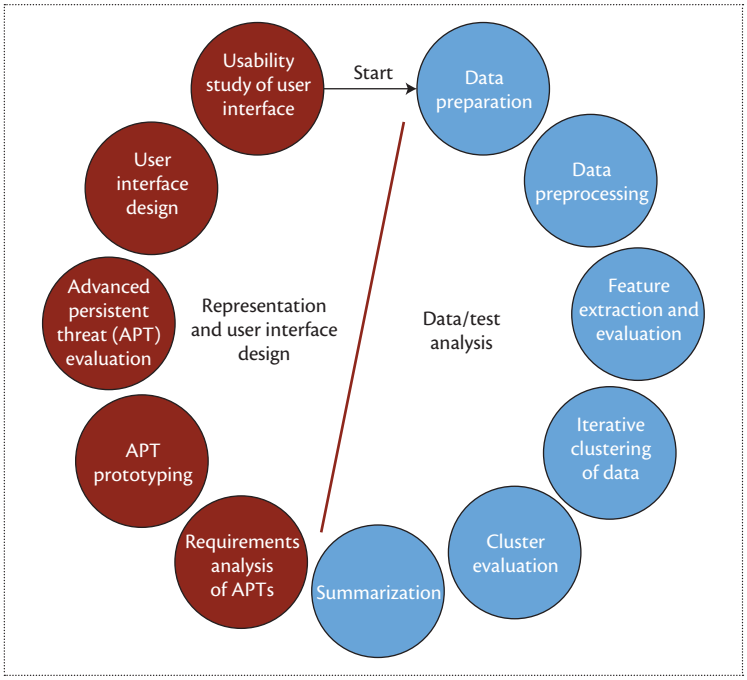

Figure 3. Process for developing a clinical decision support tool. The data mining of summarized patient types is on the right, and the user interface design is on the left.

Limitations

We relied on the MIMIC II database of ICU data and were limited by the variables reported in this database as well as by the types of patients involved in ICU admissions. We weren't able to consult with ICU physicians at Beth Israel Deaconess Hospital to get a better sense of how the data was collected and how the patterns that we observed should be interpreted.

Finally, we used only numerical data in our case study. In future research, we'd like to mine the discharge summaries and other textual data in health records to get a better sense of the reasoning used in making treatment decisions as well as develop features based on outcomes.

Application Development

Clinical decision support systems can utilize abstracted patient types in at least two distinct ways. First, physicians might browse similar patient types to get general ideas about how to diagnose and treat their current patient. Second, physicians can get recommendations (for example, tests to order or treatments to provide) for features of particular patient types. In the first usage, collecting summary information, such as mean values and their confidence intervals, for the features describing each patient type is sufficient. In the second usage, the feature summaries for each patient type are augmented with the matrix of intercorrelations among features—each patient type has its own unique set of correlations.

Again, given emergency physicians' demanding working environment, a clinical decision support application must be easy to use and the information within it easy to assimilate. Figure 3 shows the overall development process, with the data mining of summarized patient types on the right and the user interface design on the left.

Privacy Implications

In a real healthcare setting, when a patient's health data is collected, used, or shared, preserving patient privacy is paramount and is mandated in legislation, such as HIPAA. Current best practice mandates a two-tier policy approach:

- public use, in which a substantial number of attributes are suppressed and the released dataset is sufficiently protected against privacy threats; and
- restricted access research, in which the suppression criteria are more lax and some dates and geographic data are preserved in the datasets.

To ensure compliance with applicable privacy rules, personally identifiable information can be suppressed in our method's data preprocessing step (see Figure 3).

Privacy-preserving algorithms for abstracted patient types could be based on cut-off values relating to the number of patients in the cluster and the number of features in the cluster summary. However, if clusters contain summaries of more than 100 individuals, re-identification isn't the risk but rather inferences about values of unknown features given a set of known features about a patient.

Imagine we have a cluster of 100 patients, summarized by means on 10 features, with associated confidence intervals (for example, the upper and lower values of the 95 percent confidence interval). What's the risk that a determined entity could find out more about a particular individual based on this summary data? Re-identifying individuals in this type of summarized data is difficult unless someone knows a great deal about the patient and is trying to predict the value of a particular feature.

Consider the case of an insurance company that has access to a set of patient type summaries and wants to use it to estimate the probability that individual policy holders have HIV. Assume there are many patient type clusters, but only one of them has a summary feature for HIV diagnosis. Say that 90 percent of the patients in this cluster are HIV positive. What can we say about the probability that a patient has HIV, given that he or she has a roughly matching profile on the other nine features? The simple answer is that we can't properly

estimate this probability because we don't know how many people without HIV—but with otherwise similar feature profiles—were allocated to different (non-HIV) clusters. It's possible that the drugs associated with HIV treatment might be features summarized in the patient type. However, to match the patient to the cluster, the insurance company would have to know that the patient was taking these drugs, which would already imply that the policy holder was HIV positive.

By providing the matrix of intercorrelations between features within a cluster, we make it easier to predict the value of unknown features based on an individual's particular feature profile. This is, in fact, what recommender systems are designed to do. There's some risk that an organization with access to a set of summarized patient types might be able to make probabilistic inferences about one or more feature values for a patient, but only if it already has a sufficient amount of confidential information about that patient—and there will likely be considerable uncertainty surrounding those inferences. However, in view of this residual privacy risk, summarized information about patient types should be exported only to trusted institutions (for instance, other hospitals and medical organizations) and a legal framework should cover its use.

When used appropriately, summarized patient types can retain the important relationships within patient data without greatly increasing the risk of privacy violations. The increased risk should be weighed against potential benefits. In our case study, we were able to find potentially interesting patient groupings that demonstrate the value of our data-mining approach. The development and evaluation of clinical decision support tools based on summarized patient types is a topic of ongoing research. Improved typologies and clinical decision support tools should be possible with sustained research on in-house health data repositories, in close consultation with in-house medical data experts. Future research should also consider using technology from the security community to further mitigate privacy concerns.[16]

Our summarization methods aren't intended to replace methods such as de-identification and collection of informed consent but rather to extend the use of confidential data to recommendation and personalization applications in which appropriately formed clusters of patient types are the most parsimonious and useful form of data. In this case, abstraction not only facilitates application development but also turns confidential data into a more informative summarization of the original data.

We hope this research inspires medical organizations to carry out similar health data mining to identify key abstracted patient data types. In an in-house setting, the clustering process can be supplemented with a reassignment process in which patients closest to each cluster's centroid are reviewed and then filtered to find a coherent patient group that shares a meaningful collection of diagnoses, treatments, and outcomes. Once in-house data analysts find coherent core groups for each cluster concept, they can use *k*th nearest neighbor assignment to assign other patients to these cluster seeds. In this way, they can develop coherent clusters that are large enough to be nonconfidential when summarized. Thus, meaningful distributional data can be used to characterize each cluster, with accurate intercorrelations calculated between features within each cluster. ■

Acknowledgments

A research award from Google and a discovery grant from the National Science and Engineering Research Council of Canada, both held by Mark Chignell, funded this research.

References

1. A. Kobsa, "Privacy-Enhanced Personalization," *Comm. ACM*, vol. 50, no. 8, 2007, pp. 24–33.
2. K. El Emam, "Methods for the De-identification of Electronic Health Records for Genomic Research," *Genome Medicine*, vol. 3, no. 4, 2011, p. 25.
3. E. Yu et al., "Smarter Healthcare: An Emergency Physician View of the Problem," *The Smart Internet: Current Research and Future Applications*, M. Chignell et al., eds, LNCS 6400, Springer, 2010, pp. 9–26.
4. H. Takeshita, D. Davis, and S.E. Straus, "Clinical Evidence at the Point of Care in Acute Medicine: A Handheld Usability Case Study," *Proc. Human Factors and Ergonomics Society Annual Meeting*, vol. 46, no. 16, 2002, pp. 1409–1413.
5. H. Tang and J.H.K. Ng, "Googling for a Diagnosis—the Use of Google as a Diagnostic Aid: Internet Based Study," *BMJ*, no. 333, 2006, pp. 1143–1145.
6. B. Hughes et al., "Junior Physician's Use of Web 2.0 for Information Seeking and Medical Education: A Qualitative Study," *Int'l J. Medical Informatics*, vol. 78, no. 10, 2009, pp. 645–655.
7. N.K. Choudhry et al., "Systematic Review: The Relationship between Clinical Experience and Quality of Health Care," *Annals of Internal Medicine*, vol. 142, no. 4, 2005, pp. 260–273.
8. L.W.C. Chan, "Machine Learning of Patient Similarity," *Proc. Int'l Conf. Bioinformatics and Biomedicine Workshops*, IEEE, 2010, pp. 467–470
9. S.S. Ebadollahi et al., "Predicting Patient's Trajectory of Physiological Data Using Temporal Trends in Similar Patients: A System for Near-Term Prognosis," *Proc. AMIA*, 2010, pp. 192–196; www.ncbi.nlm.nih.gov/pmc/articles/PMC3041306/.
10. K. Natarajan et al., "An Analysis of Clinical Queries in an

Electronic Health Record Search Utility," *Int'l J. Medical Informatics*, vol. 79, no. 7, 2010, pp. 515–522.
11. E. Yu, "Design and Evaluation of an Improved Patient Information Management System for Emergency Department Physicians," master's thesis, Dept. Mechanical and Industrial Engineering, Univ. of Toronto, 2011.
12. R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, Wiley, 2nd ed., 2001.
13. B.S. Everitt et al., *Cluster Analysis*, Wiley, 2011.
14. G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983.
15. P.E. Spector, *Summated Rating Scale Construction: An Introduction*, Sage, 1992.
16. J. Vida and C. Clifton, "Privacy-Preserving Data Mining: Why, How, and When," *IEEE Security & Privacy*, vol. 2, no. 6, 2004, pp. 19–27.

Mark Chignell is a professor of mechanical and industrial engineering, and director of the knowledge media design institute, at the University of Toronto. He's interested in making people smarter and more effective through better design of user interfaces and applications. Chignell received a PhD in psychology from the University of Canterbury in New Zealand. Contact him at chignell@mie.utoronto.ca.

Mahsa Rouzbahman is a PhD candidate in mechanical and industrial engineering at the University of Toronto. Her main fields of interest are human factors research, user interface design for healthcare environments, clinical decision support systems, and data mining of medical records. Rouzbahman received an MS in industrial engineering from the University of Tehran. Contact her at mrouz@mie.utoronto.ca.

Ryan Kealey is a doctoral candidate in mechanical and industrial engineering at the University of Toronto and works as a usability and statistical consultant for Vocalage. His research interests include optimizing healthcare tools' design and development for patients and caregivers. Kealey received an MS in psychology from McMaster University. Contact him at ryan.kealey@utoronto.ca.

Reza Samavi is a postdoctoral research associate at the University of Toronto. His research interests include information privacy and accountability from several perspectives, including data management, ontologies, and usable privacy. Samavi received a PhD in information engineering from the University of Toronto. He's a member of IAPP and IEEE. Contact him at samavi@mie.utoronto.ca.

Erin Yu is a user experience lead at the Canadian Imperial Bank of Commerce, where she designs mobile applications. Her research interests include designing a patient management system for emergency department physicians and transforming complex data and controls into simple and user-friendly interfaces. Yu received an MS in mechanical and industrial engineering from the University of Toronto. Contact her at erin.yu@gmail.com.

Tammy Sieminowski is an attending physician at Bridgepoint Hospital in Toronto. She's interested in improving patient care and health system efficiency by bringing engineering practices and principles to healthcare and translating healthcare for engineers. Sieminowski received her medical degree from the University of Toronto. Contact her at t.sieminowski@utoronto.ca.