

Time-to-Event Predictive Modeling for Chronic Conditions Using Electronic Health Records

Yu-Kai Lin, Hsinchun Chen, and Randall A. Brown, *University of Arizona*

Shu-Hsing Li, *National Taiwan University*

Hung-Jen Yang, *Min-Sheng General Hospital*

An electronic health records-based time-to-event model identifies clinically validated risk factors associated with chronic care. Experiments suggest that EHR-based predictive modeling can effectively support decision making for chronic care patients.

One hundred and forty-one million Americans—almost half the US population—were living with one or more chronic conditions in 2010, and the patient population is expected to increase at a speed of more than 10 million new cases per decade. Given the increasing number people of living with

chronic illness and the escalating cost of chronic care, the need to facilitate clinical decision making for chronic care has never been higher. However, existing healthcare systems are oriented toward acute problems and are inadequate in managing chronic conditions. To enable effective chronic care, it is critical to be able to capture and represent a patient's disease progression pattern over time so that timely and personalized interventions can be made.

Electronic health records (EHRs) are a reliable source of longitudinal observations for monitoring the progression of chronic conditions in clinical practice. Recent years have seen surging interests in EHR data analytics for clinical decision support and knowledge discovery. Although significant progress has

been made to move the current practice in this direction, prognostic modeling frameworks and tools tailored for longitudinal EHR data analysis to support chronic care management remain inadequate.

Time-to-event modeling (also known as survival analysis) is a statistical technique for representing and predicting the length of time to an event occurrence based on an individual's traits.^{1,2} Time-to-event analysis considers not only whether an event will occur, but also the length of time to its occurrence. We use the phrase “time-to-event analysis” instead of “survival analysis” because it's more descriptive of the method and because survival isn't our focus. Indeed, caring for patients with chronic conditions involves a wide array of events other than

mortality. For instance, hospitalization and the development of severe complications are critically important events in chronic care. These events indicate the severity of a patient's condition and how well the condition has been managed. Time-to-event analysis could inform physicians about a patient's chronic condition and help the physician anticipate complications and plan interventions to reduce the patient's risk of an event.

We offer a general framework of time-to-event predictive modeling for chronic conditions based on EHRs. The proposed framework addresses various challenges in constructing a predictive model from EHRs and enables the collection of a rich set of clinically meaningful features. Furthermore, the integration of data abstraction techniques in our modeling framework demonstrates improved accuracy in time-to-event predictions.

Developing Models for Chronic Care

Many studies of prognostic modeling are conducted in the context of prospective cohort studies, in which investigators routinely follow up with patients.³ (See the related sidebar for further research in this area.) In this case, standard statistical methods for data analytics exist. These methods, however, can't be directly applied to EHR data, especially when we're analyzing the progression of chronic diseases. One challenge is the irregularly spaced data in EHRs. Many statistical methods require balanced panel data and/or equidistant time series to analyze temporal phenomena. Although such data are typically available for prospective cohort studies, in most other clinical scenarios and in EHRs, patients typically visit hospitals irregularly.⁴ That is, they don't visit hospitals routinely, and the interval between clinical visits is an arbitrary length of

time. In addition, missing values are a prevailing phenomenon in EHR data. Patients normally don't take all tests and examinations when they visit hospitals. Often, we only observe some phenotype information from patients in each of their visits, resulting in missing values for the others. Finally, EHR data are inherently highly dimensional and spread across multiple aspects of healthcare. Not all the collected data are important or relevant. Features need to be carefully selected or constructed before data analysis to achieve the best predictive performance.

Our time-to-event modeling framework differs from prior predictive modeling studies in three ways:

- We emphasize the use of large-scale, observational EHR data for chronic disease time-to-event predictive modeling.
- We formulate an innovative guideline-based feature selection approach to capture a wide array of clinically meaningful factors in our models. This approach is consistent with the spirit and practice of evidence-based medicine and enables a clinically rigorous selection of features.
- We integrate data abstraction techniques into our modeling procedure to reduce data dimensionality and enhance prediction accuracy. Although data abstraction is a common procedure in modern medical informatics research, to our knowledge it hasn't been used in time-to-event modeling.

We chose diabetes as our research case. According to the World Health Organization, the worldwide population of diabetic patients is projected to grow from 171 million in 2000 to 366 million in 2030. This study investigates the event of diabetes-induced hospitalization, which is a strong indicator that

the diabetic patient's health is being poorly managed. In our time-to-event models, we estimate and predict the length of time from the onset diagnosis of diabetes to the first diabetes-related hospitalization. Through the proposed analytical framework, we aim to identify the factors associated with the progression of diabetes and predict individual patient's long-term risk to diabetes-induced hospitalization.

Methods

Figure 1 gives an overview of our research framework, which addresses the challenges of irregularly spaced data, missing values, and high data dimensionality in EHRs. In addition, we integrate data abstraction techniques and extended Cox models with time-dependent covariates in our framework to improve prediction accuracy.

Guideline-Based Feature Selection

Clinical practice guidelines are developed to summarize the state of the art clinical research and provide recommendations for optimal management of a clinical condition. Guideline recommendations provide standards of care, delineating how clinicians should screen, evaluate, and treat a clinical condition. As such, in building time-to-event models for chronic conditions, clinical guidelines are an invaluable resource and supply critical features for data analysis.

We perform a formal knowledge engineering procedure to extract and encode concepts in guidelines, and map the concepts to EHR data. For the current research case of diabetes, we use the American Association of Clinical Endocrinologists Diabetes Care Guidelines. During the guideline-encoding process, we repeatedly consulted two clinicians to clarify clinical concepts and validate the results.

We extracted and encoded approximately 100 concepts from the guidelines.

Related Work in Prognostic Modeling

Studies of prognostic modeling prevail in the medicine and health informatics domains. Table A summarizes the areas and methods used in select work on chronic conditions.

As Table A shows, multiple methods can be used in prognostic modeling. The Cox model is the most commonly used statistical method in these studies, but it was often applied on cohort databases in which patient data were purposely collected for research. As such, missing values and irregularly spaced observations are often not major issues in these databases. However, this isn't the case in EHRs, as noted in the main article.

When missing values are a concern, data imputation is often considered. Aditya Khosla and his colleagues compared several single imputation methods, such as column mean and linear regression.² These methods are naïve and prone to bias because they neglect the variance in the data-generation process.

Table A also indicates that prior studies of clinical predictive modeling, especially in conventional medical research, typically focus on a limited number of features. These features are selected only when clinical evidence supports their causal relations with the event or outcome variable. This evidence-based feature selection approach is well-received because it's in line with the philosophy of evidence-based medicine. However, the resulting feature

set is very confined, typically about a dozen features. To date, it remains unclear how to extend this evidence-based approach to select a larger number of features that remain clinically justifiable.

References

1. M.B. Sesen et al., "Survival Prediction and Treatment Recommendation with Bayesian Techniques in Lung Cancer," *AMIA Annual Symp. Proc.*, vol. 2012, 2012, pp. 838–847.
2. A. Khosla et al., "An Integrated Machine Learning Approach to Stroke Prediction," *Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 2010, pp. 183–191.
3. D.R. Cox, "Regression Models and Life-Tables," *J. Royal Statistical Soc. Series B (Methodological)*, vol. 34, no. 2, 1972, pp. 187–220.
4. J. Hippisley-Cox et al., "Predicting Risk of Type 2 Diabetes in England and Wales: Prospective Derivation and Validation of QDScore," *British Medical J.*, vol. 338, 2009; <http://dx.doi.org/10.1136/bmj.b880>.
5. B.H. Cho et al., "Application of Irregular and Unbalanced Data to Predict Diabetic Nephropathy Using Visualization and Feature Selection Methods," *Artificial Intelligence in Medicine*, vol. 42, no. 1, 2008, pp. 37–53.
6. J. Hippisley-Cox et al., "Predicting Cardiovascular Risk in England and Wales: Prospective Derivation and Validation of QRISK2," *British Medical J.*, vol. 336, no. 7659, 2008, pp. 1475–1482.

Table A. Summary of related prior studies.

Study	Event/outcome	Data source	Feature selection	No. of features	Modeling technique	Address missing values
M. Berkan Sesen and colleagues ¹	Lung cancer one-year survival	CD	EB	9	NB, BN	No. Use complete data
Aditya Khosla and colleagues ²	Stroke risk in 5 years	CD	SMLB	200	Cox model, ³ SVM	Yes. By single imputation methods
Julia Hippisley-Cox and colleagues ⁴	Risk of type II diabetes in 10 years	CD	EB	10	Cox model	Yes. By multiple imputation
Baek Hwan Cho and colleagues ⁵	Onset of diabetic nephropathy	EHRs	SMLB	184	LR, SVM	Yes. By temporal abstraction
Julia Hippisley-Cox and colleagues ⁶	Risk of cardiovascular diseases in 10 years	CD	EB	14	Cox model	No. Use complete data

BN = Bayesian network; CD = cohort database; EB = evidence-based; EHR = electronic health record; LR = logistic regression; NB = naïve Bayes; SMLB = statistical or machine-learning based; and SVM = support vector machine.

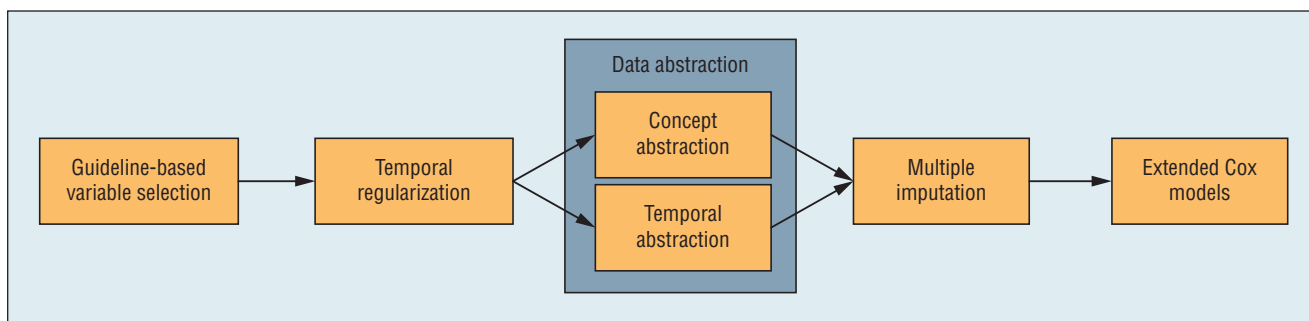


Figure 1. Research framework. The five-component time-to-event framework addresses the issues of irregularly spaced data, missing values, and high data dimensionality in electronic health records (EHRs).

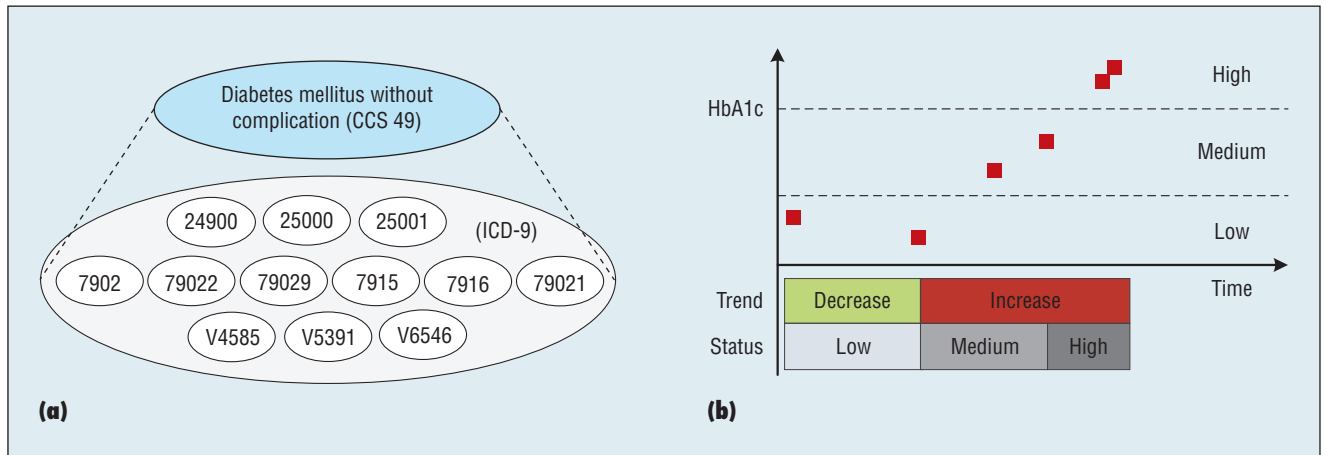


Figure 2. Two types of data abstraction: (a) concept abstraction and (b) temporal abstraction. CCS stands for Clinical Classifications Software and ICD stands for International Classification of Diseases.

The list of concepts covers evaluations, diagnoses, and treatments for diabetes as well as various diabetic complications, including cardiovascular diseases, diabetic nephropathy, retinopathy, neuropathy, and peripheral diseases. We then mapped these concepts to the corresponding items in EHRs, resulting in about 400 International Classification of Diseases (ICD)-9 diagnostic codes, 150 treatments, and 20 lab tests and physical evaluations.

Temporal Regularization

Irregularly spaced data often cause difficulties for temporal data analysis. One way to address this problem is to construct equidistant temporal grids from observations. This temporal regularization approach is common in studies that have unequally spaced time series observations. Following this approach, we divide a patient's history into monthly intervals. If there are multiple observations for the same feature in an interval, we take their average as a representative value for the feature in that interval. If there's no observation in an interval, we consider it as a missing value and address it through multiple imputation.

Data Abstraction

The essence of data abstraction is to suppress details and highlight higher-order

meanings, which makes it particularly salient to complex problem solving. Generally, data abstraction encompasses concept and temporal abstraction. Figure 2 illustrates the two types.

In concept abstraction, each raw datum is independently mapped to a generalized concept. We map ICD-9 diagnosis codes to higher-order categories in the Clinical Classifications Software (CCS), and represent medications by their family names. After this concept abstraction procedure, we reduce diagnosis features to 41 CCS categories, and treatment features to 11 treatment families. In temporal abstraction, numerical time-series information is represented by symbolic codes to reflect temporal states (such as high, medium, or low) and trends (such as decrease or increase). We perform both state and trend abstraction for our numerical features. We perform state abstraction by discretizing values into three bins (high, medium, and low), each with the same number of observations. Our trend abstraction is either upward or downward, depending on whether an observed value is followed by a greater or lesser value.

Table 1 summarizes our final feature set with four major feature categories. The base features are gender, age, and other common lab or physical tests. The next two categories—

concept abstracted diagnosis features (DX features) and concept abstracted treatment features (TX features)—are binary variables that take the value 0 before the first appearance of the diagnosis/treatment and the value 1 thereafter. Finally, the temporal abstracted base features encode the states and trends of the lab/physical test results over time.

Multiple Imputation

Multiple imputation is an effective technique for missing data processing and enables a less-biased data analysis than single imputation methods (for example, filling in the column mean). It achieves this by reflecting the uncertainty of the missing values in multiple imputed datasets. Specifically, we generate five imputed datasets from the original dataset. In each imputed dataset, a missing value is filled in by a plausible value randomly sampled from a conditional distribution of the observables. The uncertainty of a missing value is represented by its multiple realizations in different imputed datasets.

We execute our multiple imputation procedure using the Amelia II program.⁵ Amelia II makes inferences on the missing values using an expectation-maximization algorithm combined with a bootstrapping procedure. Compared to other alternative multiple

Table 1. The final feature set for time-to-event modeling.

Category	Category code	Variables
Base	Base	Gender, age, smoking, HbA1c, fasting glucose, low-density lipoprotein (LDL) cholesterol, triglyceride, systolic blood pressure, blood urea nitrogen (BUN), creatinine, body weight
Concept abstracted diagnosis features	DX	Clinical Classifications Software (CCS) classes 3, 49, 50, 51, 53, 58, 59, 60, 87, 89, 91, 95, 98, 99, 100, 101, 104, 106, 107, 108, 109, 110, 112, 114, 115, 116, 156, 157, 158, 161, 162, 163, 199, 236, 237, 248, 651, 657, 660, 663, and 670
Concept abstracted treatment features	TX	angiotensin-converting-enzyme (ACE) inhibitors, angiotensin II receptor blockers (ARBs), amputation, antihypertensive therapy, antiplatelet therapy, dipeptidyl peptidase-4 (DPP4) inhibitors, insulin, lipid-lowering therapy, metformin, sulfonyleureas, thiazolidinediones
Temporal abstracted (TA) base features	TA	States and trends of base features (HbA1c, fasting glucose, LDL cholesterol, triglyceride, systolic blood pressure, BUN, creatinine, body weight)

Table 2. Significant risk factors for hospitalization events due to diabetes.

Risk factors	Feature category	Hazard ratio	Lower CI*	Upper CI bound
Open wounds of extremities (CCS 236)	DX	12.898	1.495	121.187
Acute and unspecified renal failure (CCS 157)	DX	11.243	1.569	81.705
Insulin treatment	TX	6.082	3.780	9.787
Smoking	Base	2.750	1.745	4.336
Antiplatelet therapy	TX	2.145	1.200	3.841
Upward trend of body weight	TA	1.788	1.238	2.582
Upward trend of fasting glucose	TA	1.642	1.098	2.459
HbA1c	Base	1.112	1.004	1.233
LDL cholesterol	Base	1.007	1.000	1.013
Fasting glucose	Base	1.003	1.002	1.004
Sulfonyleurea treatment	TX	0.663	0.442	0.994
Low-level state of fasting glucose	TA	0.545	0.304	0.976

* CI = 95 percent confidence interval.

imputation programs, Amelia II is simpler, faster, and produces more robust results that are similar to more sophisticated programs that rely on Markov chain Monte Carlo simulations. We build models and make time-to-event predictions on each of the imputed datasets separately. The reported experimental results are the averaged performance over the five datasets.

Extended Cox Model

The Cox proportional hazards model is a popular tool for time-to-event analysis.¹ The Cox model makes the proportional hazards assumption. An extended Cox model allows covariates to be time-dependent, which alleviates the potential issue of nonproportional hazards and enables a more flexible modeling

framework.² The extended Cox model is given by

$$h(t, \mathbf{X}(t)) = h_0(t) \exp \left[\sum_{i=1}^{P_1} \beta_i X_i + \sum_{j=1}^{P_2} \delta_j X_j(t) \right],$$

where $h(t, \mathbf{X}(t))$ is the hazard value at time t , $h_0(t)$ is an arbitrary baseline hazard function, and \mathbf{X} is a covariate matrix containing P_1 time-independent covariates and P_2 time-dependent covariates. One of the advantages of Cox models in survival and time-to-event analysis is that we don't need to specify the baseline hazard function $h_0(t)$.

Given the constraint on time-invariant covariates, the regular Cox model typically estimates time to event based on information solely from the initial

observations. However, chronic conditions persist and evolve over time. When the research interest is the progression of chronic conditions, initial observations alone are often insufficient to construct a robust and accurate time-to-event model. In contrast, the extended Cox model allows HbA1c, creatinine, and other physiological features to vary with time. For similar reasons, we chose the extended Cox model rather than other popular machine-learning algorithms, such as support vector machines or decision trees, because these algorithms don't have an elegant approach to deal with data censoring and can't incorporate time-varying covariates. Finally, we don't consider other parametric time-to-event models, such as the accelerated failure time models because they require making additional distributional assumptions on the baseline hazard $h_0(t)$.

We construct three time-dependent Cox models. The *baseline* model uses only the base features, which include essential phenotype information on a patient's demographic background and values of key tests (see Table 1). The *extended* model includes DX and TX features along with the base features. Finally, the *full* model further incorporates temporal abstracted (TA) features.

Results and Discussion

We obtained de-identified EHRs from a major 600-bed hospital in Taiwan. In our experiment, 1,860 patients

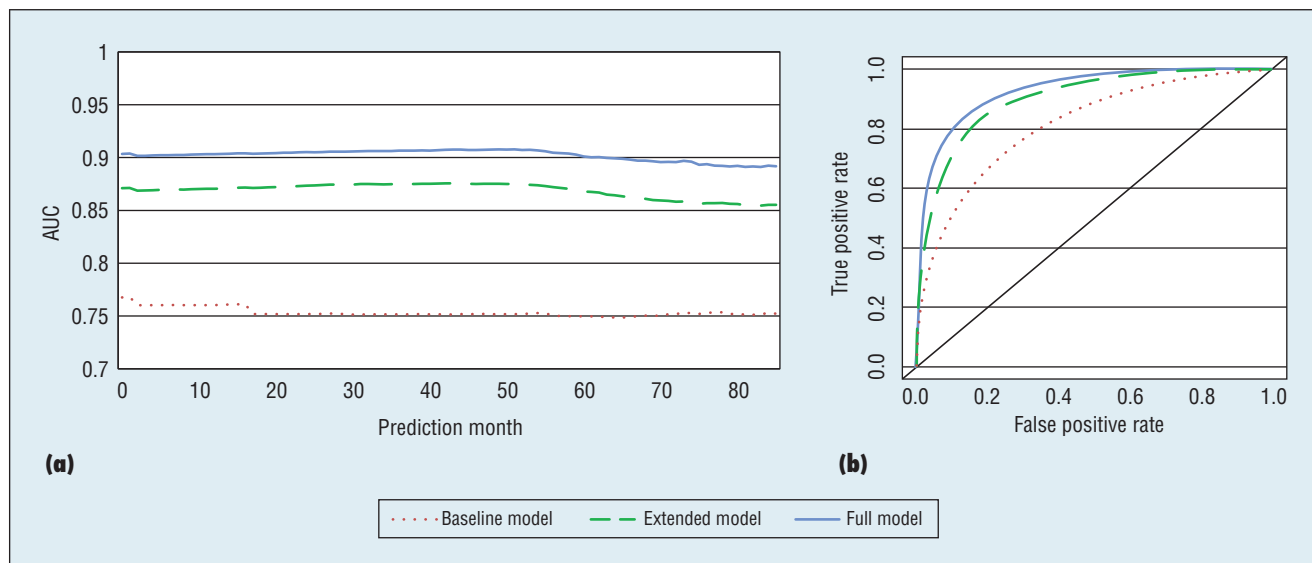


Figure 3. Performance comparison: (a) the receiver operating characteristic (ROC) of the area under an ROC curve (AUC) values over different prediction points, and as a representative case; (b) the time-dependent ROC curve at the 42nd month.

satisfied our selection criteria of having an onset diagnosis of diabetes from 2003 to 2012. Among them, 155 had experienced the event (hospitalization due to diabetes). One defining characteristic of survival and time-to-event data is *right censoring*. Right censoring occurs when a study ends or when a patient is lost to follow-up. As such, we don't know the actual time to an event for right-censored observations. In this study, we consider observations to be right censored at the last visit of an event-free patient.

We evaluate the models by how well they discriminate between patients with and without the event. We derive receiver operating characteristic (ROC) curves and compute the area under an ROC curve (AUC) to assess a model's discriminative ability. Because a patient's history is divided into monthly intervals, we perform prognostic prediction on each month until the last event time (at the 85th month). Our time-dependent ROC analysis for predictive accuracy strictly follows the method outlined by Patrick Heagerty and Yingye Zheng.⁶ Figure 3a shows the AUC values of the three models over

time. The average AUC scores for the baseline, extended, and full models are 0.75, 0.87, and 0.90, respectively. As the figure demonstrates, the full (extended) model consistently outperforms the extended (baseline) model on each of the 85 monthly-based predictions. Figure 3b shows the ROC curves of the three models for the predictions made at the 42nd month, which is the midpoint among all the prediction months. The ROC curves for the predictions of the other months present similar patterns.

Table 2 shows the statistically significant covariates in our full model (p -value ≤ 0.05 in at least two imputed datasets). Hazard ratio is a typical measure of effect in time-to-event analysis. When a hazard ratio is significantly greater (less) than one, the risk factor is deemed positively (negatively) associated with the event. Although we couldn't find suitable prior clinical studies as an external validation source, overall the list of risk factors in Table 2 seems to be clinically reasonable as validated by our medical collaborators. Open wounds on extremities are often the result of diabetic foot

syndrome, which occurs in severe diabetic cases. Renal failure is often the result of poorly controlled diabetes. Insulin treatments are prescribed when oral medications can no longer control the patient's rising glucose level. Smoking and antiplatelet therapy are associated with more frequent hospitalizations, likely representing the increased risk of stroke, heart attack, and peripheral vascular disease. It's noteworthy that several temporal abstracted features are on the list of significant risk factors. The upward trends of body weight and fasting glucose signify high hazard ratios. On the other hand, a low-level fasting glucose is associated with fewer hospitalization events.

While our modeling framework attained a remarkable performance in time-to-event predictions (an AUC of 0.90 in our full model), two limitations of this study point to future research. First, we didn't deal with clinical texts. Clinical texts contain rich descriptions of a patient's current and historical conditions. However, the inclusion of clinical text in temporal modeling necessitates

THE AUTHORS

Yu-Kai Lin is a doctoral candidate in the Management Information Systems Department and a research associate in the Artificial Intelligence (AI) Lab at the University of Arizona. His research interests include health informatics, machine learning, and economic analysis of health IT. Lin has an MBA in technology management from National Tsing Hua University. Contact him at yklin@email.arizona.edu.

Hsinchun Chen is the University of Arizona Regents' Professor and Thomas R. Brown Chair in Management and Technology in the Management Information Systems Department and the funding director of the AI Lab. His research interests include Web computing, search engines, digital libraries, intelligence analysis, biomedical informatics, data/text/Web mining, and knowledge management. Chen has a PhD in information systems from New York University. He's a Fellow of IEEE and AAAS. Contact him at hchen@eller.arizona.edu.

Randall A. Brown is an associate professor of medicine with the University of Arizona Health Network and a research affiliate with the University of Arizona Management Information Systems Department. His research interests include biomedical informatics, stem cell therapy, business intelligence, and portfolio management. Brown has an MD from Rush Medical College in Chicago. Contact him at randallb9@gmail.com.

Shu-Hsing Li is the vice president for finance affairs and a professor of accounting at the National Taiwan University. His research interests include multinational transfer pricing, text mining and capital market research, and health care intelligence and cost management. Li has a PhD in accounting from New York University. Contact him at shli@ntu.edu.tw.

Hung-Jen Yang is currently the president and CEO of Min-Sheng General Hospital in Taiwan. Yang has a master's of public health from Harvard University and an MBA from Claremont Graduate University, respectively. Contact him at fred@hccahealth.com.

We acknowledge Min-Sheng hospital for supporting the research data and our clinical collaborators Dr. Craig Stump and Dr. Hung-Yuan Li for providing medical consultations and comments.


References

1. D.R. Cox, "Regression Models and Life-Tables," *J. Royal Statistical Soc. Series B (Methodological)*, vol. 34, no. 2, 1972, pp. 187–220.
2. J.D. Kalbfleisch and R.L. Prentice, *Statistical Analysis of Failure Time Data*, 2nd ed., Wiley-Interscience, 2002.
3. J. Hippisley-Cox et al., "Predicting Risk of Type 2 Diabetes in England and Wales: Prospective Derivation and Validation of QDScore," *British Medical J.*, vol. 338, 2009; <http://dx.doi.org/10.1136/bmj.b880>.
4. B.H. Cho et al., "Application of Irregular and Unbalanced Data to Predict Diabetic Nephropathy Using Visualization and Feature Selection Methods," *Artificial Intelligence in Medicine*, vol. 42, no. 1, 2008, pp. 37–53.
5. J. Honaker, G. King, and M. Blackwell, "Amelia II: A Program for Missing Data," *J. Statistical Software*, vol. 45, no. i07, 2011; www.jstatsoft.org/v45/i07.
6. P.J. Heagerty and Y. Zheng, "Survival Model Predictive Accuracy and ROC Curves," *Biometrics*, vol. 61, no. 1, 2005, pp. 92–105.
7. Y.-K. Lin, H. Chen, and R.A. Brown, "MedTime: A Temporal Information Extraction System for Clinical Narratives," *J. Biomedical Informatics*, vol. 46, Dec. 2013, pp. S20–S28.

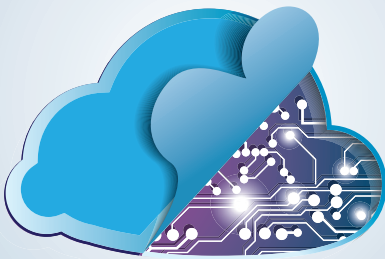
techniques for temporal text processing. Current systems achieve only about 70 percent accuracy in normalizing temporal expressions.⁷ We therefore limited this study to structured information in EHRs. Second, our guideline-based feature selection involved extensive manual work in encoding guideline concepts and mapping them to EHR data elements. As such, this approach isn't as scalable as automatic feature selection schemes. Future research might also consider combining or comparing guideline-based feature selection with other statistical feature-selection methods. □

Acknowledgments

Patient data are de-identified and HIPAA compliant. We obtained Institutional Review Board approval on this study. This research is supported by US National Science Foundation grant CBET-0730908, Defense Threat Reduction Agency Grant HD-TRA10910058, and Taiwan National Science Council grant NSC101-3114-Y-002-003.

 Selected CS articles and columns are also available for free at <http://ComputingNow.computer.org>.


New in 2014



IEEE CLOUD COMPUTING

IEEE Computer Society's newest magazine tackles the emerging technology of cloud computing. Subscribe today!

computer.org/cloudcomputing

IEEE  computer society 