Journal of the American Medical Informatics Association Advance Access published September 15, 2015

Real-time prediction of mortality, readmission, and length of stay using electronic health record data

RECEIVED 9 December 2014
REVISED 23 March 2015
ACCEPTED 24 June 2015

Xiongcai Cai[1], Oscar Perez-Concha[2], Enrico Coiera[2], Fernando Martin-Sanchez[3], Richard Day[4], David Roffe[5], Blanca Gallego[2]

ABSTRACT

Objective To develop a predictive model for real-time predictions of length of stay, mortality, and readmission for hospitalized patients using electronic health records (EHRs).

Materials and Methods A Bayesian Network model was built to estimate the probability of a hospitalized patient being "at home," in the hospital, or dead for each of the next 7 days. The network utilizes patient-specific administrative and laboratory data and is updated each time a new pathology test result becomes available. Electronic health records from 32 634 patients admitted to a Sydney metropolitan hospital via the emergency department from July 2008 through December 2011 were used. The model was tested on 2011 data and trained on the data of earlier years.

Results The model achieved an average daily accuracy of 80% and area under the receiving operating characteristic curve (AUROC) of 0.82. The model's predictive ability was highest within 24 hours from prediction (AUROC = 0.83) and decreased slightly with time. Death was the most predictable outcome with a daily average accuracy of 93% and AUROC of 0.84.

Discussion We developed the first non–disease-specific model that simultaneously predicts remaining days of hospitalization, death, and readmission as part of the same outcome. By providing a future daily probability for each outcome class, we enable the visualization of future patient trajectories. Among these, it is possible to identify trajectories indicating expected discharge, expected continuing hospitalization, expected death, and possible readmission.

Conclusions Bayesian Networks can model EHRs to provide real-time forecasts for patient outcomes, which provide richer information than traditional independent point predictions of length of stay, death, or readmission, and can thus better support decision making.

Keywords: prediction, patient outcome, mortality, readmission, length of stay

BACKGROUND AND SIGNIFICANCE

Rapid identification of hospitalized patients at high risk for an extended length of stay (LOS), readmission, or death has the potential to improve quality of care and reduce avoidable harm and costs. Early and accurate identification of patients at high risk of death can be used to call emergency medical teams to prevent death or, alternatively, to initiate counseling about end-of-life care.[1] Appropriate management of patients at their end of life by the provision of emergency and hospital medical services, particularly the transition from acute to palliate care, is a growing challenge for our health care systems, requiring better education and improved risk-assessment tools.[2] Early and accurate knowledge of LOS can aid hospital administrators in the management of bed occupancy. This is a crucial problem faced by hospitals, which are pressured to shorten the LOS of their patients, potentially increasing their risk of dying after discharge.[3] An accurate estimate of LOS together with risk of readmission and death can also help clinicians with important discharge planning strategies for their patients; these strategies are likely to improve continuity of care, and prevent readmissions and deaths after discharge.[4]

With the implementation of electronic health record (EHR) systems, laboratory test results, surgery data, ward transfers, and other relevant temporal clinical information are available at the point of care. This knowledge can be used to predict mortality, LOS, and readmissions in real time. The most successful current models, achieving a C statistic around 0.9, are those predicting in-hospital mortality.[5–8] Two of these models update their predictions of in-hospital mortality risk as new information about the patient becomes available: (1) Rothman *et al*[6] predicts mortality within 24 hours using the "Rothman index," a heuristically built, continuously updated index of patient conditions based on pathology results, nursing assessments, and vital signs; and (2) Wong *et al*[7] uses a time-dependent Cox regression method to predict patients' daily risk of death during hospitalization. Long-term mortality using large administrative datasets is also estimated with high accuracy.[9,10] In these models, history of health care utilization and services, such as palliative care, are used as proxies for patients' clinical status.

Prediction of LOS is also addressed in a number of studies but mostly in the context of specific diseases. Very few studies attempt to predict LOS across all conditions using EHRs. A notable exception is a study by Liu *et al*[11] that uses automated laboratory and comorbidity measures in a regression model to predict LOS at admission, with R^2 of 0.134 and a root-mean-square error of 170 days. Readmission prediction using routinely collected administrative and clinical data focuses on predicting all-cause readmission within 30 days postdischarge and generally achieves poor to fair results.[12] Only 1 study reported a C statistic above 0.8.[13] In this model, the strongest predictive power came from a comprehensive risk score trained with administrative and claims data of over 5.6 million patients to classify patients into hierarchical condition categories.

Although existing models perform well, particularly for the prediction of in-hospital mortality, most of them have been designed to

Correspondence to Blanca Gallego, Centre of Health Informatics, Australian Institute of Health Innovation, Level 6, 75 Talavera Road, Macquarie University, NSW 2109, Australia; blanca.gallegoluxan@mq.edu.au

For numbered affiliations see end of article.

predict a single outcome within a given time period. This study represents the first model to simultaneously estimate the sequence of future daily probabilities of being in the hospital, "at home" (for which we mean having been discharged alive), or dead over a time period. This provides a more comprehensive, finer-grained forecast of patient status. In addition, these predictions are continuously updated as new information becomes available. These properties make this model a suitable tool to aid in decision making at the point of care.

MATERIALS AND METHODS

Data

Electronic health records from 32 634 patients admitted to a Sydney metropolitan hospital via the emergency department (ED) from July 1, 2008 through December 31, 2011 were collected. Additionally, for each patient, 1-year history and 6-month postadmission records of all hospital admissions, emergency department visits, and deaths within the State of New South Wales (NSW) were extracted from population health datasets—namely, the NSW Admitted Patient Data Collection (APDC), the NSW Emergency Department Data Collection (EDDC), and the NSW Registry of Births, Deaths and Marriages (RBDM). The Centre for Record Linkage independently carried out both data linkages—the linkage between the hospital EHRs and the NSW administrative datasets, and the linkage amongst the NSW administrative datasets.[14] Of the original 32 895 patients from the Sydney metropolitan hospital, 15 could not be linked to the APDC and 246 could not be linked to the EDDC. The linkage amongst the APDC, EDDC, and RBDM was performed using a probabilistic linkage procedure, which guarantees false-positive rates $< 0.5\%$ and false-negative rates $< 0.1\%$.[14] The dataset was split into nonoverlapping training and test sets. The training set contained records of 24 625 patients admitted to the hospital from July 1, 2008 through December 31, 2010. The test set contained the remaining records of 8009 patients admitted to the hospital in 2011.

Each patient was characterized by a set of static variables, including patient demographics (such as age and sex), patient history (such as cumulative LOS in the previous year), and administrative admission information (such as day of the week of admission or mode of arrival to the ED). Dynamic variables included days already in the hospital, ward type, and the value of pathology test results per temporal event. A temporal event was defined as the time when 1 or more pathology test results were made available and valid in the EHR for clinicians to read (see figure 1).

Pathology test results were labeled according to the laboratory-provided reference range as "missing," "normal," or "abnormal." Here, pathology tests were defined by test type as well as panel type. For example, Bicarbonate appears twice—once as part of the Urea, Electrolytes, Creatinine panel and a second time as part of the blood gas HCO_3 panel. A list of these variables and their distribution across temporal events in the training dataset is shown in tables 1 and 2. Hospital admissions were characterized using ward type. Patient wards were correlated to major diagnostic categories and preferred over them because, unlike diagnostic categories and codes, they are readily available to use for prediction in real time. Patient comorbidities were estimated using *International Classification of Diseases, Tenth Revision,* (*ICD-10*) codes from patients' hospital admissions during the previous year. However, since only 31% of patients had a history of hospitalization in the previous year, a comorbidity index using this

Figure 1: An illustration of how the predictive model is updated following the availability of one or more pathology test results in the EHR system. Patient outcomes comprise the probability of staying at hospital, being "at home" (ie, having been discharged alive), or being dead during each of the 7 days following a temporal event.
Abbreviation: CT, computed tomography.

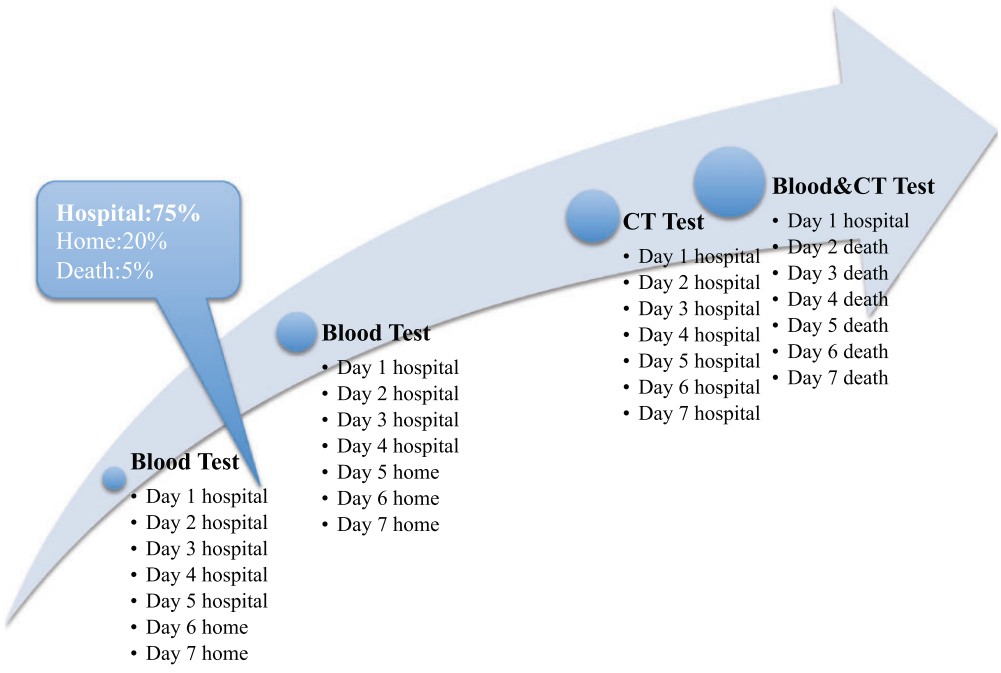


Table 1: Patient and admission characteristics across temporal events in the training dataset

Feature	Temporal	Statistics				
		Mean	SD	25th percentile	Median	75th percentile
Age, y	No	57	21	39	58	74
Cumulative LOS previous year, d	No	6	18	0	0	1
No. of days since previous admission[a]	No	268	139	132	365	365
No. of days since admission	Yes	6.1	12.1	0	2	7
Day since last event	Yes	0.4	1.4	0	0	0
Pathology tests from admission	Yes	29.4	62.4	2	6	2

Gender, %	No	Male	Female
		61	39

		No	Yes
Mental Disorders, %	No	60	40
Cancer, %	No	89	11

Triage Code[b], %	No	1	2	3	4	5	Empty
		17	21	52	7	2	1

Day of Week, %	No	Mon	Tue	Wed	Thu	Fri	Sat	Sun
		14	16	14	13	15	14	13

Time of Day, %	No	00:00-07:59	08:00-15:59	16:00-23:59
		20	44	36

Marital Code, %	No	Married	Divorced	Widow	Others
		36	4	10	50

Mode of Arrival to Emergency, %	No	Ambulance Service	61
		Private Car	17
		Community Public Transport	17
		Others	5

Ward, %	Yes	Emergency Department	22
		Intensive Care/High Dependency Unit	13
		Cardiothoracic Surgery/Transplant	12
		Neurology/Vascular/Stroke Unit	11
		Gastrointestinal Unit	9
		Cardiology	9
		Oncology/Hematology/Immunology/Pharmacology	8
		Others combined (< 7% each)	16

Abbreviation: LOS, length of stay.
[a]Note that past history spans for 1 year and 31% of patients were admitted during the previous year.
[b]Triage codes: 1 = Resuscitation, 2 = Emergency, 3 = Urgent, 4 = Semi-urgent, 5 = Non-urgent.

variable was not informative. Only conditions that could not have emerged during a hospitalization and were thought to be important co-morbidity groups, namely cancer patients and patients with mental health conditions, were included and defined using their corresponding *ICD-10* codes before and during the current hospitalization.

Patient outcomes comprised the probability of being in the hospital, at home, or dead during each of the 7 days following a temporal event. In the training dataset, 2% of patients died and 18% were discharged alive in the first day following a new pathology result. By day 7, 5% of patients had died and 41% had been discharged from the hospital.

Model

A predictive model was built in 5 steps as depicted in figure 2:

1. For each target day after the time of prediction (day 1, 24 hours after admission, to day 7), a feature selection algorithm (described

Table 2: Summary of laboratory test results across temporal events in the training dataset[a]

Pathology	Missing, %	Abnormal, %	Normal, %	Pathology	Missing (%)	Abnormal (%)	Normal (%)
Albumin	57	13	30	Lymphocytes	41	32	27
Alkaline Phosphatase	57	11	32	Magnesium	60	7	34
Alanine Aminotransferase	57	16	27	Mean Corpuscular Hemoglobin	41	18	42
Activated Partial Thromboplastin Time	76	9	15	Mean Corpuscular Hemoglobin Concentration	41	2	58
Aspartate Aminotransferase	57	16	26	Mean Corpuscular Volume	41	9	50
Base Excess	67	11	21	Methaemoglobin	67	5	28
Basophils	41	1	58	Monocytes	41	9	50
Bicarbonate	39	20	41	Neutrophils	41	23	37
Calcium	60	9	31	Oxygen Saturation	67	11	22
Carboxyhemoglobin	67	10	23	Carbon Dioxide Partial Pressure	66	11	23
Chloride	39	6	55	Potential Hydrogen	67	13	20
Creatinine	39	16	44	Platelets	41	13	46
C-Reactive Protein	72	16	12	Oxygen Partial Pressure	67	20	13
Glomerular Filtration Rate	40	15	45	Urine Potassium	39	7	54
Eosinophils	41	3	57	Potassium (UEC)	67	4	29
Fraction of Inspired Oxygen	66	0	34	Prothrombin Time	82	6	12
Gamma Glutamyl Transferase	57	20	23	Red Blood Cell	41	30	30
Glucose	67	14	18	Red Cell Distribution Width	41	24	36
Hemoglobin	41	26	33	Urine Sodium	39	14	47
Bicarbonate	67	13	19	Sodium (UEC)	67	14	19
Hematocrit	41	33	27	Total Bilirubin	57	7	36
Inorganic Phosphate	60	8	33	Total Protein	57	13	30
Ionised Calcium	67	18	15	Urea	39	21	40
Lactate	67	4	28	White Blood Cell	41	19	40

Abbreviation: UEC, urea, electrolytes, creatinine.
[a]Pathology test results were labeled as "missing," "normal," or "abnormal" according to the laboratory-provided reference range.

below) was used to select those variables most highly correlated with the target day values and yet uncorrelated with each other. We will call these the *primary features*.
2. For each primary feature, the same feature selection algorithm was used to select those variables most highly correlated with the primary feature values and yet uncorrelated with each other. We will call these the *secondary features*. This strategy allows missing values from the primary set of features to be inferred from the added secondary set of features, and represents a model-based imputation approach within a Bayesian Network (BN) framework.[15]
3. Primary and secondary features were included as nodes in a BN. Arcs were created from the target days to the primary features and from the primary features to the secondary features.
4. Prior and conditional probabilities were learned from the training dataset. This dataset consisted of a set of temporal events with their corresponding primary and secondary features and patient outcomes.
5. In the test dataset, outcomes for each patient on each target day after each temporal event were predicted using the "learned" BN model.

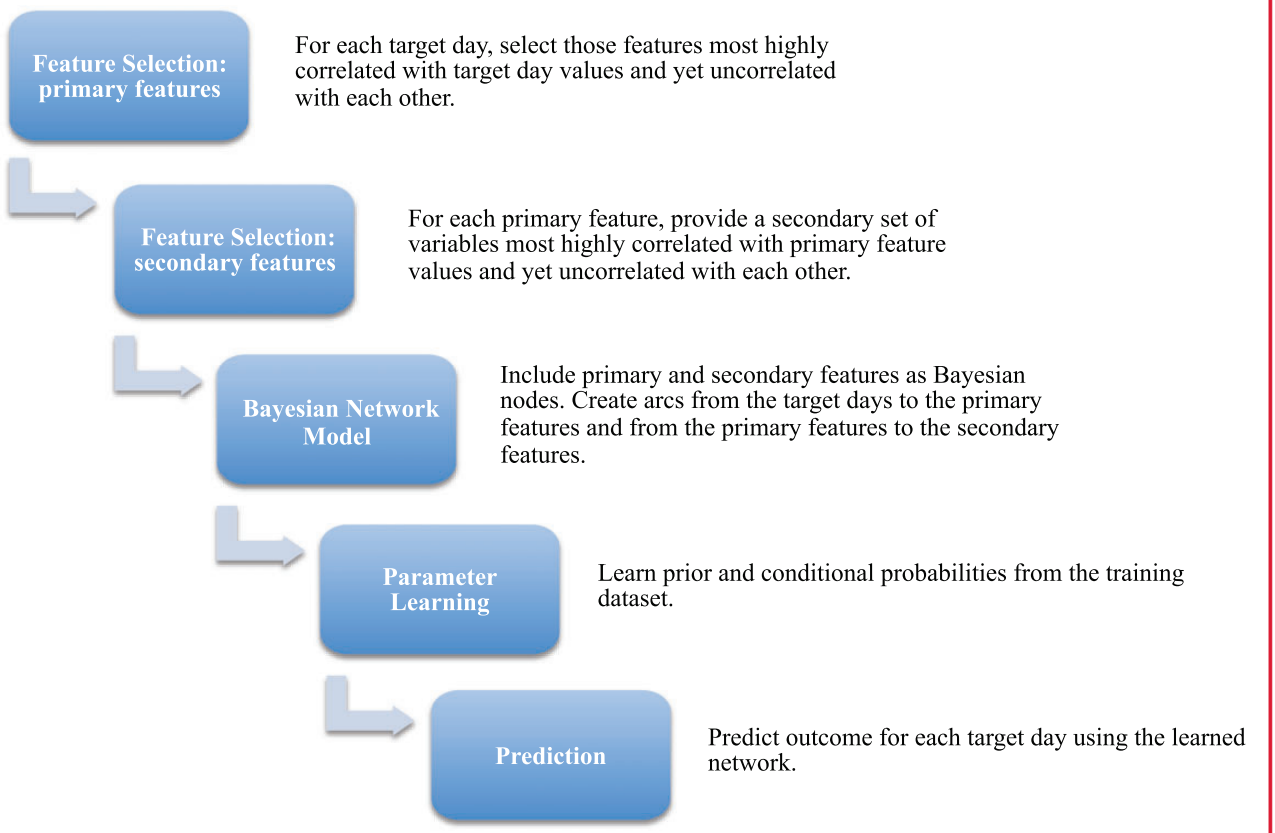
Feature Selection Algorithm

We selected features using a correlation-based feature selection approach and a "best-first search" algorithm.[16] In this approach, a feature V_i is said to be relevant to the class C if there exists some feature values v_i and class value c for which $P(V_i = v_i) > 0$ such that

$$P(C = c | V_i = v_i) \neq P(C = c).$$

Starting with an empty set of features, the space of feature subsets is searched by a "greedy hill-climbing" algorithm[16] augmented with a backtracking facility. The final selected feature subset contains those features most highly correlated with the output classes and yet uncorrelated with each other. Correlation is defined by the Pearson correlation coefficient.[17]

Figure 2: Overview of how the predictive model was built in 5 steps.



Bayesian Network

A BN, $B = (G, P)$ consists of an acyclic directed graph (G), with nodes corresponding to a set of random variables X; P is the joint probability distribution of the variables in X such that

$$P(X) = \prod_{x \in X} P(x|\pi_G(x))$$

where $\pi_G(x)$ denotes the parents of X in G. In our patient risk prediction problem, X corresponds to the input data (features) and output (probability of an outcome). The representation of a joint probability distribution in a BN generally reduces the number of parameters that need to be estimated and allows for more efficient probabilistic inference.[18]

Our BN model included the 7 target days, selected primary features, and selected secondary features as nodes, representing random variables. Arcs representing direct dependencies between 2 variables were created from each of the target days to their corresponding selected primary features, and from these primary features to their corresponding selected secondary features. Prior probability distributions for nodes without parents and probability distributions conditional on parent nodes were learnt from the training dataset and are shown in appendix A. The full network containing all nodes and arcs is shown in figure 3. The BN model was programmed using Structural Modeling, Inference, and Learning Engine, a fully portable library of C++ classes implementing graphical decision-theoretic methods developed at the Decision Systems Laboratory, University of Pittsburgh.[19]

RESULTS

Temporal validation was used to evaluate the ability of the model to predict events for unseen patients within the same population from which the model was derived.[20] Area under the receiving operating characteristic and accuracy for each outcome class in the testing set are shown in table 3. Other model performance indicators have been included in appendix A. The highest predictive power was achieved on day 1 (24 hours after a given prediction time), with a daily average accuracy of 86% and AUROC = 0.83, and decreased slightly with time. Daily average AUROC remained above 0.80 for all days. As with previous models, prediction of death was the most accurate outcome, with average accuracies of 93% and AUROC = 0.84.

One of the advantages of our approach is that, for each patient, the model provides a sequence of future daily probabilities (days 1–7) for each outcome class, as opposed to independent single point predictions of LOS, readmission or death within a prespecified period. In order to illustrate this approach, we selected 4 groups from the testing dataset: patients who die during the week after the time of prediction, patients who are discharged alive, patients who continue to be hospitalized, and patients who are readmitted after discharge. For each group, we randomly picked patients for whom the model correctly classified the patient outcome for all, or most of, the 7 future days. These predictions of future "patient trajectories" are displayed in figure 4.

Figure 4A illustrates a prediction of *expected continuing hospitalization*, where the probability of staying in the hospital dominates all others throughout the 7-day forecasting period. Figure 4B shows a

Figure 3: The Bayesian Network model includes 7 target days (in yellow), selected primary features (in purple), and selected secondary features (in blue) as nodes, representing random variables. Arcs were created from each of the target days to their corresponding selected primary features, and from these primary features to their corresponding selected secondary features. Rectangular nodes represent dynamic variables while elliptical nodes represent static variables.
Abbreviations: U. Sodium, urine sodium; U. Potassium, urine potassium; HCT, hematocrit; WBC, white blood cell count; Hgb, hemoglobin; UEC, urea, electrolytes, creatinine; Alk. Phos., alkaline phosphatase; pH, potential hydrogen; CRP, C-reactive protein; RBC, red blood cell count; APTT, activated partial thromboplastin time; Tot. Protein, total protein; ED arrival, mode of arrival to emergency department; Triage, triage category; Prev. LOS, cumulative length of stay in previous hospitalizations; Inorg Phos, inorganic phosphate; Test Count, number of laboratory tests performed so far during hospitalization; HOS days, days already in the hospital; Hours since HOS, hours since previous hospitalization.

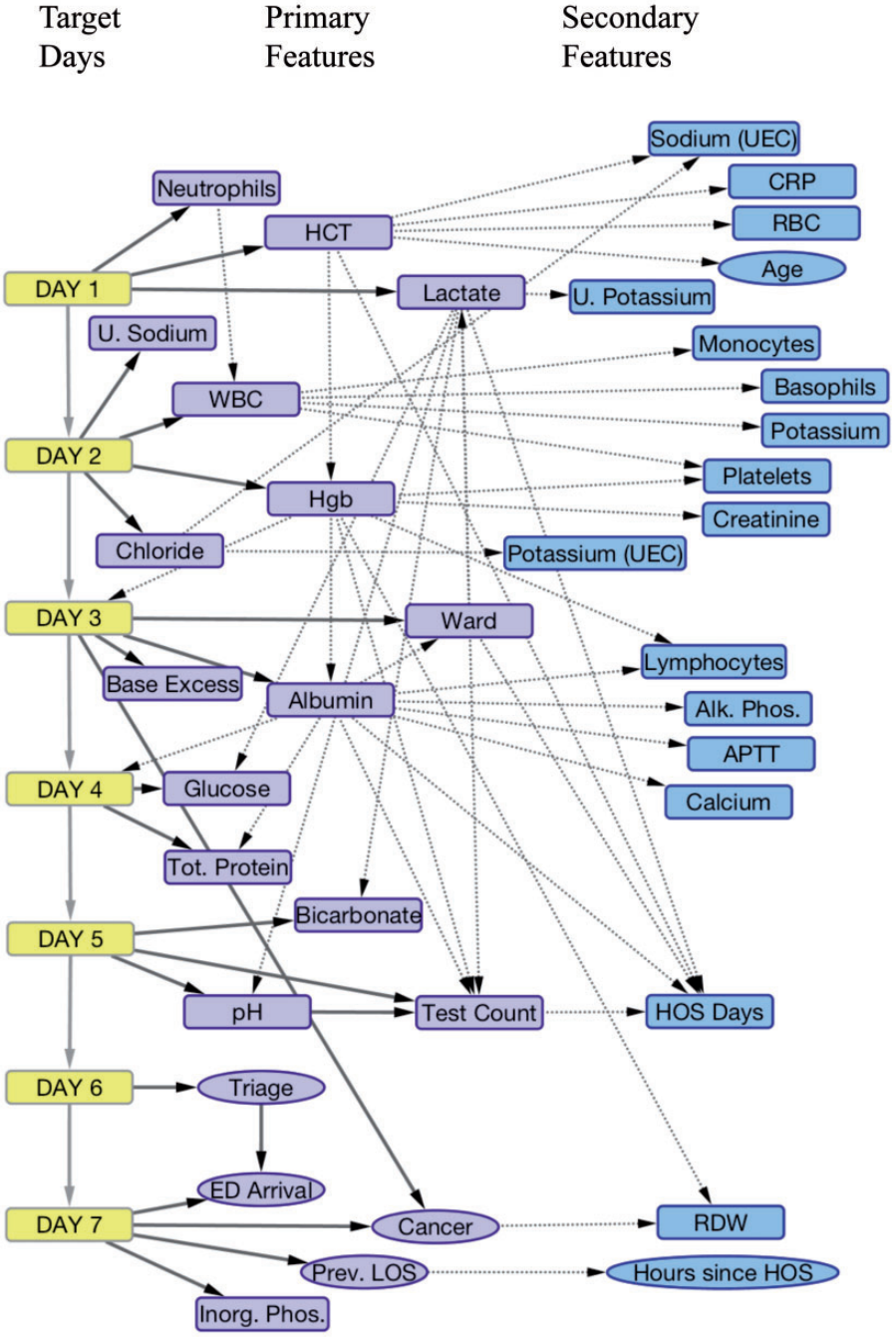


Table 3: Bayesian Network accuracy (ACC) and area under the receiver operating characteristic curve (AUROC) for each target day and each outcome class: Hospital, Home (referring to discharged patient), and Death

		Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Weekly Average
Hospital	AUROC	0.78	0.77	0.76	0.76	0.76	0.75	0.74	0.76
	ACC	0.78	0.74	0.71	0.70	0.69	0.68	0.67	0.71
Home	AUROC	0.84	0.84	0.84	0.84	0.85	0.84	0.84	0.84
	ACC	0.81	0.78	0.77	0.77	0.76	0.76	0.75	0.77
Death	AUROC	0.87	0.86	0.84	0.84	0.84	0.83	0.83	0.84
	ACC	0.97	0.96	0.93	0.92	0.92	0.91	0.90	0.93
Daily Average	AUROC	0.83	0.82	0.82	0.82	0.81	0.81	0.80	0.82
	ACC	0.86	0.83	0.80	0.80	0.79	0.78	0.77	0.80

typical prediction of *expected discharge*, where the probability of dying remains low throughout the week, and the probability of being at home increases steadily from day 1 until it becomes higher than that of being in the hospital at day 5. This indicates that discharge without complications is expected around that day. Figures 4C and D show typical predictions of *expected death*. Figure 4C represents expected immediate death, since the probability of dying in day 1 is already higher than all others and remains so throughout the week. In Figure 4D, the probability of staying in the hospital initially dominates but decreases at the same time as the probability of death increases; and on day 4, the latter surpasses all other probabilities and remains dominant until the end of the forecasting period. This indicates that the patient is expected to die around or after day 4. Figure 4E shows a prediction of *expected readmission,* where the probability of being at home exceeds that of being in the hospital on day 3 but becomes lower again after day 4. This indicates the possibility of a readmission after day 4. In reality the patient for whom this prediction was made, was discharged on day 6 and readmitted on day 7. This type of trajectories, where the daily probabilities of being in the hospital and at home fluctuate around 50%, are common in patients who are discharged and readmitted within a fortnight, indicating that readmission is hard to predict with the variables available in this study.

DISCUSSION

This study presents a validated model for estimating the probability that a hospitalized patient will remain in the hospital, be discharged, readmitted, or dead in each of the next 7 days immediately after a new pathology test result is available in the EHR system. To the best of our knowledge, our model is the first non–disease-specific model that combines the following features: (1) consolidates remaining days of hospitalization, death, and readmission in the same outcome variable; and (2) predicts a sequence of future daily probabilities rather than a single probability over a given time period. As illustrated in figure 4, estimating simultaneously the future daily probability of being in the hospital, at home, or dead over a time period provides a more comprehensive, finer-grained forecast of patient risk.

Similar to the Rothman index,[6] this model has been built to provide continuously updated information of a patient's status independent of disease type or reason for admission. This provides a longitudinal view of the patient, which may help with earlier detection of acute events, discharge planning, and continuity of care. In our model, patients' risk of extended LOS, readmission, or death is updated whenever a new pathology test becomes available. This time for updating has been chosen due to the higher frequency of laboratory tests as compared with other temporal events such as ward movements or surgeries. However, it could be easily extended to incorporate other events. Additionally, in this model, a pathology test that is not updated at the time of a temporal event is considered as missing. A possible alternative may be to consider the last available result or the last available result within a time range.

This model uses both patient history and administrative and clinical information contained in patients' EHRs. Most of this information was available in real time, with the exception of the diagnostic codes used to identify cancer patients and the data on previous hospitalizations at different hospitals. Although we used *ICD-10* codes in this study, identification of cancer patients and other phenotypes is currently possible using data contained in the EHR system.[21] As shown in our experiments, administrative and clinical variables are discriminative variables that are suitable for predicting death, at least within the following week. However, prediction of discharge time and readmission is more challenging, since those outcomes might depend on other variables not available to this study, such as social or economic factors, and may also require larger training sets. We expect that as the accuracy, consistency, completeness, and availability of EHRs rapidly improve, so will the predictive power of these types of models, enabling new decision support tools.

We have used a graphical static probabilistic model in which both the network structure as well as the parameter learning has come from data. Desirable extensions of this work include the incorporation of knowledge-based information in the construction of the network structure, alternative ways of dealing with missing values, and the extension to dynamic BNs.

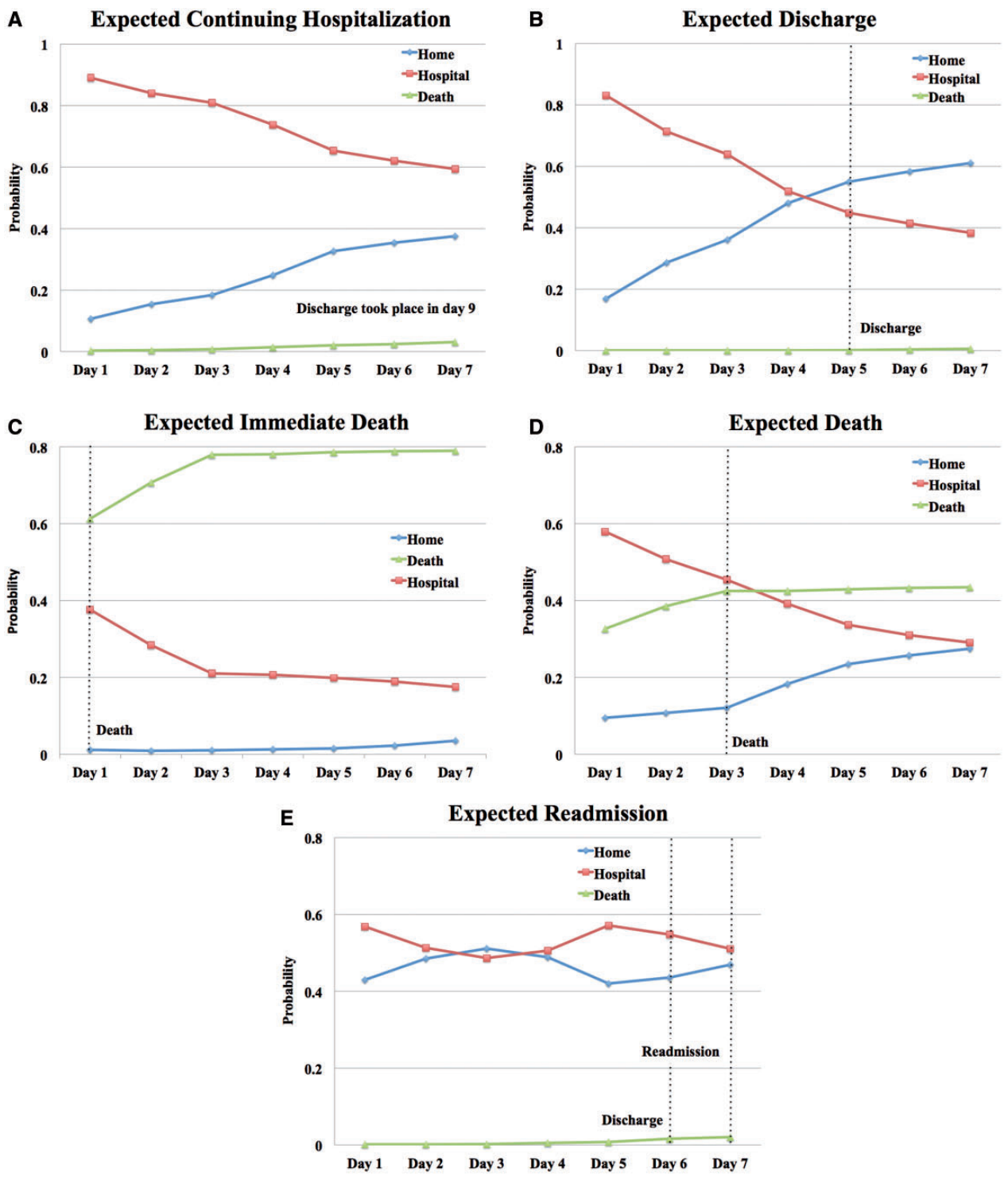
CONCLUSION

We have developed a BN model that simultaneously estimates the future daily probabilities of being in the hospital, at home, or dead for a hospitalized patient over a week after a new pathology test result becomes available. This model has good predictive power and provides a finer-grained longitudinal forecast of patient status to aid in decision making at the point of care.

CONTRIBUTORS

Dr Cai performed the data analysis, developed and tested the predictive model, and contributed to the writing of the manuscript. Dr Perez-Concha contributed to the data analysis and writing of the manuscript. Professors Martin-Sanchez, Coiera, and Day provided advice on the

Figure 4: Model predictions for selected individual patients. Each line represents a future daily probability of being in the hospital (red), "at home" (blue), or dead (green). Days 1 to 7 represent the future consecutive days relative to the time of prediction. The dotted vertical lines indicate true events. Patients and times of prediction have been randomly selected among examples for which the model correctly classifies patient outcomes for all, or most of, the 7 days. Panel A shows a typical prediction of expected continuing hospitalization; Panel B illustrates a prediction of expected discharge. Panels C and D are typical predictions of expected death, and Panel E predicts possible readmission.



model development, provided the clinical context, and contributed to the writing of the manuscript. Roffe provided expertise on the hospital's EHR system. Dr Gallego was the senior researcher leading the project. She led the model development and the writing of the manuscript.

FUNDING

This work was supported by National Health and Medical Research Council, project grant No. 1045548 and program Grant 568612.

COMPETING INTERESTS

None.

ETHICS APPROVAL

Ethics approval was obtained from the NSW Population and Health Services Research Ethics Committee (HREC/13/CIPHS/29) and the hospital's Ethics Committee.

ACKNOWLEDGEMENTS

Data collection, extraction, and analysis. The administrative data used in this study was provided by the NSW Ministry of Health. This data was linked to the hospital's EHR system by the NSW Centre for Record Linkage. The authors would like to thank the hospital's IT personnel for providing the data extraction. Dr Cai was responsible for the data analysis after the extraction.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://jamia.oxfordjournals.org/.

REFERENCES

1. Kenward G, Castle N, Hodgetts T, *et al*. Evaluation of a medical emergency team one year after implementation. *Resuscitation* 2004;61(3):257–263.
2. Forero R, McDonnell G, Gallego B, *et al*. A literature review on care at the end-of-life in the emergency department. *Emerg Med Int* 2012;2012. doi:10.1155/2012/486516.
3. Pouw ME, Peelen L, Moons K, *et al*. Including post-discharge mortality in calculation of hospital standardised mortality ratios: retrospective analysis of hospital episode statistics. *BMJ* 2013;347:f5913.
4. Bauer M, Fitzgerald L, Haesler E, *et al*. Hospital discharge planning for frail older people and their family. Are we delivering best practice? A review of the evidence. *J Clin Nurs* 2009;18(18):2539–2546.
5. Tabak YP, Sun X, Nunez CM, *et al*. Using electronic health record data to develop inpatient mortality predictive model: acute laboratory risk of mortality score (ALaRMS). *J Am Med Inform Assoc* 2014;21(3):455–463.
6. Rothman MJ, Rothman SI, Beals J. Development and validation of a continuous measure of patient condition using the electronic medical record. *J Biomed Inform* 2013;46(5):837–848.
7. Wong J,Taljaard M, Forster AJ, *et al*. Derivation and validation of a model to predict daily risk of death in hospital. *Med Care* 2011;49(8):734–743.
8. van Walraven C, Escobar GJ, Greene JD, *et al*. The Kaiser Permanente inpatient risk adjustment methodology was valid in an external patient population. *J Clin Epidemiol* 2010;63(7):798–803.
9. Coiera E, Wang Y, Magrabi F, *et al*. Predicting the cumulative risk of death during hospitalization by modeling weekend, weekday and diurnal mortality risks. *BMC Health Serv Res* 2014;14(1):226.
10. van Walraven C. The Hospital-patient One-year Mortality Risk score accurately predicted long-term death risk in hospitalized patients. *J Clin Epidemiol* 2014;67(9):1025–1034.
11. Liu V, Kipnis P, Gould MK, *et al*. Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. *Med Care* 2010;48(8):739–744.
12. Kansagara D, Englander H, Salanitro A, *et al*. Risk prediction models for hospital readmission: a systematic review. *JAMA* 2011;306(15):1688–98.
13. Shulan M, Gao K, Moore CD. Predicting 30-day all-cause hospital readmissions. *Health Care Manag Sci* 2013;16(2):167–175.
14. Lawrence G, Dinh I, Taylor L. The Centre for Health Record Linkage: a new resource for health services research and evaluation. *Health Inform Manage J* 2008;37(2):60–62.
15. Pigott TD. A review of methods for missing data. *Educ Res Eval* 2001;7(4):353–83.
16. Hall MA. *Correlation-based Feature Selection for Machine Learning* [PhD dissertation]. The University of Waikato, Hamilton, NewZealand, 1999.
17. Pearson K. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* 1895: 240–242.
18. Pearl J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA: Morgan Kaufmann Publishers Inc; 1988.
19. Druzdzel MJ. SMILE: Structural Modeling, Inference, and Learning Engine and GeNIe: A development environment for graphical decision-theoretic models (Intelligent Systems Demonstration). In: *AAAI '99/IAAI '99 Proceedings of the Sixteenth National Conference on Artificial Intelligence*. Menlo Park, CA: American Association for Artifical Intelligence; 1999: 902–3.
20. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19(4):453–73.
21. Shivade C, Raghavan P, Fosler-Lussier E, *et al*. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014;21(2):221–30. doi: 10.1136/amiajnl-2013-001935.

AUTHOR AFFILIATIONS

[1]School of Computer Science and Engineering, The University of New South Wales, Sydney, Australia
[2]Centre of Health Informatics, AIHI, Macquarie University, Sydney, Australia
[3]Melbourne School of Information, The University of Melbourne, Melbourne, Australia
[4]School of Medical Sciences, The University of New South Wales, Sydney, Australia
[5]Information Technology Service Centre, St Vincent's Hospital, Sydney, Australia