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

Xiongcai Cai¹, Oscar Perez-Concha², Enrico Coiera², Fernando Martin-Sanchez³, Richard Day⁴, David Roffe⁵, Blanca Gallego²

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. Appropriate management of patients at their end of life by the provision of emergency and hospital medical services, particularly the transition from acute to palliative care, is a growing challenge for our health care systems, requiring better education and improved risk-assessment tools. 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. 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.

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. Two of these models update their predictions of in-hospital mortality risk as new information about the patient becomes available: (1) Rothman et al. 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. 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. 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. that uses automated laboratory and comorbidity measures in a regression model to predict LOS at admission, with R² 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. Only 1 study reported a C statistic above 0.8. 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...
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. 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%. 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 8,009 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—one as part of the Urea, Electrolytes, Creatinine panel and a second time as part of the blood gas HCO3 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
variable was not informative. Only conditions that could not have emerged during a hospitalization and were thought to be important 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
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.

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. 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 $P(C = c|V_i = v_i) > 0$ 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 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.

---

**Table 2: Summary of laboratory test results across temporal events in the training dataset**

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

Abbreviation: UEC, urea, electrolytes, creatinine.

Pathology test results were labeled as “missing,” “normal,” or “abnormal” according to the laboratory-provided reference range.
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.

Our BN model included the 7 target days, selected primary features, and selected secondary features as nodes, representing random variables. Arrows 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.

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. 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.
mitted within a fortnight, indicating that readmission is hard to predict. Around 50%, are common in patients who are discharged and readmitted. Daily probabilities of being in the hospital and at home fluctuate on day 6 and readmitted on day 7. This type of trajectories, where the probability 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, 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. 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 development of the model. Concha contributed to the data analysis and 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 development of the model.

---

**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**

<table>
<thead>
<tr>
<th></th>
<th>Day 1</th>
<th>Day 2</th>
<th>Day 3</th>
<th>Day 4</th>
<th>Day 5</th>
<th>Day 6</th>
<th>Day 7</th>
<th>Weekly Average</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hospital</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUROC</td>
<td>0.78</td>
<td>0.77</td>
<td>0.76</td>
<td>0.76</td>
<td>0.76</td>
<td>0.75</td>
<td>0.74</td>
<td>0.76</td>
</tr>
<tr>
<td>ACC</td>
<td>0.78</td>
<td>0.74</td>
<td>0.71</td>
<td>0.70</td>
<td>0.69</td>
<td>0.68</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td><strong>Home</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUROC</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.85</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
</tr>
<tr>
<td>ACC</td>
<td>0.81</td>
<td>0.78</td>
<td>0.77</td>
<td>0.77</td>
<td>0.76</td>
<td>0.76</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td><strong>Death</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUROC</td>
<td>0.87</td>
<td>0.86</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
<td>0.83</td>
<td>0.83</td>
<td>0.84</td>
</tr>
<tr>
<td>ACC</td>
<td>0.97</td>
<td>0.96</td>
<td>0.93</td>
<td>0.92</td>
<td>0.92</td>
<td>0.91</td>
<td>0.90</td>
<td>0.93</td>
</tr>
<tr>
<td><strong>Daily Average</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUROC</td>
<td>0.83</td>
<td>0.82</td>
<td>0.82</td>
<td>0.82</td>
<td>0.81</td>
<td>0.81</td>
<td>0.80</td>
<td>0.82</td>
</tr>
<tr>
<td>ACC</td>
<td>0.86</td>
<td>0.83</td>
<td>0.80</td>
<td>0.80</td>
<td>0.79</td>
<td>0.78</td>
<td>0.77</td>
<td>0.80</td>
</tr>
</tbody>
</table>
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/CIPH/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

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