Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200186

Predicting Length of Stay in Hospital Using Electronic Records Available at the Time of Admission

Marta WILK[a,d], D William R MARSH[a,d1], Sarah DE FREITAS[c], John PROWLE[b,d]
[a]*School of Electronic Engineering and Computer Science*
[b]*William Harvey Research Institute, Barts and the London School of Medicine and Dentistry*
[c]*Acute and Renal Medicine, The Royal London Hospital, Barts Health NHS Trust*
[d]*Queen Mary University of London*

Abstract. Predicting a patient's hospital length of stay (LoS) can help manage staffing. In this paper, we explore LoS prediction for a large group of patients admitted non-electively. We use information available at admission, including demographics, acute and long-term diagnoses and physiological tests results. Data were extracted from the electronic health records (EHR), so that the LoS prediction would not require additional data entry. Although the data can be accessed, the system does not present a unified view of the data for one patient: to resolve this we designed a process of cleaning and combining data for each patient. The data was used to fit semi-parametric, parametric and competing outcomes survival models. All models performed similarly, with concordance of approximately 0.7. Calibration results showed underestimation of predicted discharges for patients with high discharge probabilities and overestimation of predicted discharges for those with low discharge probabilities. The main challenges in operationalizing LoS predictions are delays in entering admissions data into EHR and absent data about non-medical factors determining discharges.

Keywords. Length of Stay, Bed management, Survival analysis, Medical data transformation.

1. Introduction

We develop and evaluate a Length of Stay (LoS) predictive model and use the prediction of individual patients' LoS from daily admitted cohorts to estimate levels of bed occupancy in the following days. Given the potential of LoS prediction to optimise hospital operations, many methods have been used to predict LoS including machine learning techniques, such as neural networks, logistic regression and regression trees [1] [2] [3], as well as statistical methods including negative binomial regression [4]. Here, we use survival models to obtain the probabilities of discharge of individual patients from daily admissions. Whereas many earlier studies have focussed on a restricted patient group, our aim is a model that can be applied to a wide range of patients.

[1] Corresponding author: School of Electronic Engineering and Computer Science, Mile End Road, London E1 4NS, Email: d.w.r.marsh@qmul.ac.uk

2. Background

The Acute Care Unit (ACU) in the Royal London Hospital provides short term treatment of patients aged 16 or older, admitted non-electively from Emergency Department (ED) or specialist clinics. After 48 hours patients are typically moved to specialist units or discharged. Patients are admitted to the ACU with a wide range of conditions, including injuries, acute health conditions and acute episodes of chronic conditions.

We used the data of the admissions to the ACU in the period 2013-17, consisting of 35,792 admission records for 20,836 unique patients – 52% women and 48% men, with a median age of 65 years. Median LoS was 4 days, with a mean of 8.4 days. 4.8% admissions ended with the death in hospital. There were 6,246 unique ICD-10 diagnostic codes in the sample. Patients' LoS was defined as the time interval between admission to ACU and discharge from hospital or death.

3. Methods

3.1. Data transformation

The Barts Health NHS Trust stores information about patient encounters in a Cerner Data Warehouse. Although the stored data is comprehensive, there are several challenges for building a data set for predictive modelling.

- Multiple identifiers are used: although there is a unique patient ID, every admission to a separate unit, including moves between units during a stay, is recorded with a unique encounter ID. The encounters that make up a hospital stay are not linked: a way to do this was found to assemble a patient's diagnoses, physiological results and treatment data for one admission.
- Data is entered at different times. Some data are entered into the electronic system in real-time during admissions but other data are recorded on paper and only entered into EHR at the end of patient's admission to the hospital. This can result in some of the data not entered into system or data recorded with errors.
- Codes and descriptions of diagnoses are stored together and there is no commonly used system to group individual diagnoses into clusters of clinically similar categories. A patient's diagnoses and physiological tests are often recorded with ED encounter ID, instead of ACU encounter ID.

We extracted demographic data from the ACU unit using patient and admission IDs and cleaned duplicates and errors. We extracted diagnoses data from all hospital's units using the patients ID in the demographic data. For each admission, multiple diagnoses were recorded. To obtain information about comorbidities, we extracted ICD-10 codes from all admissions before the current one, counting as comorbidities diagnoses that were recorded at least once in an earlier admission, excluding any from the latest hospital stay. However, only 33% of admissions were repeats; other patients were assumed free of serious comorbidities. We used the R *icd* package to classify ICD codes into comorbidities and calculated Charlson scores.

To cluster acute diagnoses, we extracted ICD codes and descriptions from current admissions and used R *erm* package, that groups clinically similar diagnoses. For the codes that could not be matched automatically, manual comparisons and corrections were made, from the ICD-10-CM 2014 to the ICD-10-CM-2019 codes, compatible with *erm* package. This method enabled all the common acute diagnoses to be clustered. Physiological tests results (blood tests and NEWS score), were extracted from all hospital units from time interval *admission date minus one day* and *admission date plus two days,* to account for the tests performed in ED for patients admitted overnight and for repeated tests e.g. when the first sample has been compromised. This procedure resulted in missing values of 3%-5% for the most common tests (e.g. Hemoglobin, RBCDW) and up to 30% for less common ones (e.g. Bilirubin, Albumin). The missing values were imputed with population normal values.

In the last stage, we derived additional variables from the extracted data. For example, for *Kidney_problem* variable, we calculated for repeated admissions median Creatinine for patients' previous admissions and extracted dialyses and kidney transplant codes. We assigned the category 3 for admissions with kidney transplant recorded, 2 for ones with dialyses, 1 with the Creatinine level at a given admission at least 1.5 higher than median Creatinine from previous admissions and 0 if none applied. *Admission_number* and *Readmission_within_year* variables were derived to indicate if an admission was a consecutive one and if a given patient was often readmitted. The final data set consisted of 49 predictors.

3.2. Survival analysis

We developed and evaluated semi-parametric (Cox) and parametric (log-normal) models and Fine and Gray competing outcomes model. We set the evaluation time for 30 days and we defined the 'event' as discharge from hospital. Deaths during admission were right-censored for Cox and parametric models and a competing outcome for the Fine and Gray model. For Cox and parametric models, we fitted splines for continuous variables to account for non-linear relationships with LoS and we used backward variable selection algorithm to select significant predictors (p-value <=0.05 for a global model).

4. Results

The hazards in the global Cox model for the selected variables were not proportional, and stratifying on non-proportional predictors did improve the proportionality result. Full Fine and Grey model did not converge, so this model was developed with the variables selected for Cox model. All models had similar concordance, presented in Table 1.

Table 1 Comparison of models' concordance.

Model	Concordance Index (AUC)
Cox	0.705
Log-normal	0.706
Fine and Gray model	0.703

Given the similar performance, we chose the Cox model for further analysis, as hazard ratios for discharge event could be directly obtained for Cox model regression coefficients, but not from Fine and Gray model [5] [6]. Validation of Cox model with resampling showed there was little overfitting. Predicting discharges by day 4 showed a mean prediction error of 0.033, with underestimation of discharges for the groups with high discharge probability and overestimation of discharges for groups with low discharge probability. Prediction on day 11 showed a slight overestimation of discharges for patients with medium probabilities of discharge, but the mean error decreased to 0.014. Figure 1 shows Cox model calibration performance on days 4 and 11.

4.1. Bed estimation

We calculated means of individual patients' discharge predictions with confidence intervals (CI) to obtain the number of beds released from daily admissions. Figure 2 shows the predicted output for days 2, 3, 10,15 and 30 from admission of 14 patients. In the external sample of 180 admissions, for days 2 and 3, 66% of observed numbers of beds released were within predicted CI, day 10 - 81%, day 15 - 79% and day 30 - 75%.

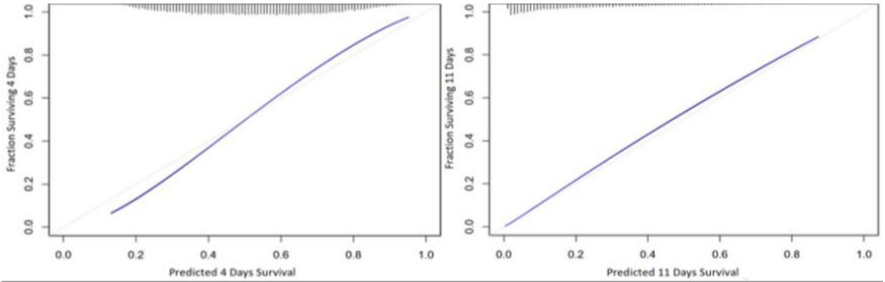

Figure 1 Calibration of predictions on 4th and 11th day - Cox function.

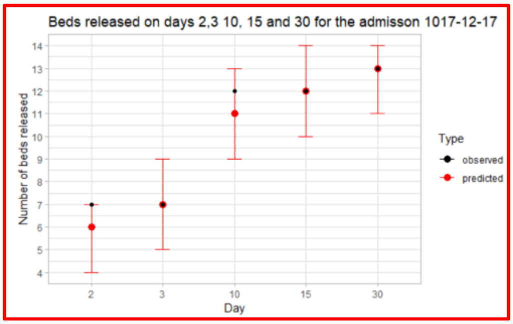

Figure 2 Predicted and observed number of beds released on days 2,3,10,15 and 30.

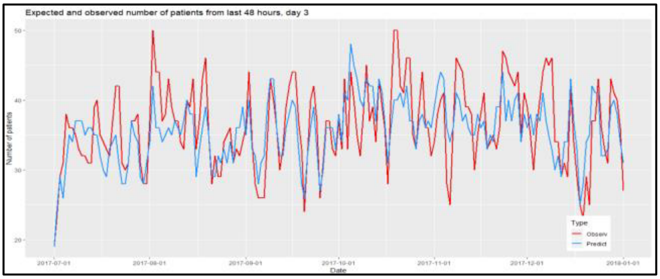

Figure 3 Predicted and observed number of beds occupied from 3 consecutive admissions.

We used the model to predict bed occupancy from several consecutive daily admissions. Figure 3 shows the predicted and observed occupancy after 3 days from the admission on a given day, and the previous 2 days, for 180 admission days. The model predictions match the trends in the observed data. Perhaps because the Cox model does not account for deaths during the admission, the variation in larger in observed than predicted values.

5. Conclusions and future work

We have shown that patient information stored in EHR is sufficient for predicting LoS for a large group of patients with varying diagnoses admitted non-electively. This opens the way to make such predictions routinely without requiring additional data generation.

Building a data sets for modelling is challenging due to the form in which data is stored, but we showed that the necessary linking of records can be achieved so that a uniform dataset can be derived for all patients. We used LoS predictions for patients from daily admissions to estimate the bed occupancy on chosen days post-admission. Despite the suboptimal calibration group of patients with more extreme discharge probabilities, the model performed well for the prediction of beds released from daily admissions.

The next steps in operationalizing the model will be to investigate ways to facilitate real-time entry and access to data. Notably, the ICD-10 codes on which we rely are currently coded manually, but the use of machine learning methods (e.g. [7]) to automate coding has already been investigated and need not be perfectly accurate for LoS prediction. Sourcing comorbidity data for patients admitted to hospital for the first time also needs to be resolved; we plan to investigate the possible use of GP records for this. We will work to extend the prediction to elective admissions and investigate the hospital units to which patients are transferred when they are not discharged from the ACU. We will also investigate ways to incorporate the social factors in LoS analysis, as they can strongly influence a patient's discharge times and cannot be estimated from purely medical records. Finally, to transform the model into a decision tool, senior nurses and bed managers will be consulted on what predictions outputs are useful to support effective bed management.

6. References

[1] Pei-Fang Tsai, Po-ChiaChen, Yen-YouChen, Hao-YuanSong, Hsiu-MeiLin, Fu-ManLin, Qiou-PiengHuang, “Length of Hospital Stay Prediction at the Admission Stage for Cardiology Patients Using Artificial Neural Network,” Journal of Healthcare Engineering, 2016.

[2] Sean Barnes, Eric Hamrock, Matthew Toerper, Sauleh Siddiqui, Scott Levin, “Real-time prediction of inpatient length of stay for discharge prioritization,” Journal of the American Medical Informatics Association, vol. 23, 2016.

[3] Lior Turgeman, Jerrold H. May, Roberta Sciulli, “Insights from a machine learning model for predicting the hospital Length of Stay (LOS) at the time of admission,” Expert Systems with Applications, vol. 78, 2017.

[4] Evelene M Carter, Henry WW Potts, “Predicting length of stay from an electronic patient record system: a primary total knee replacement example,” BMC Medical Informatics and Decision Making, 2014.

[5] E. Christensen, “Multivariate Survival Analysis Using Cox's Regression Model,” Hepatology, vol. 7, no. 6, 1987.

[6] Peter C. Austin, Jason P. Fine, “Practical recommendations for reporting Fine‐Gray model analyses for competing risk data,” Statistics in Medicine, vol. 36, no. 27, 2017.

[7] YunZhi Chen, HuiJuan Lu, LanJuan Li, “Automatic ICD-10 coding algorithm using an improved longest common subsequence based on semantic similarity,” PLOS One, vol. 12, no. 3, 2017.