Predicting Number of Hospitalization Days Based on Health Insurance Claims Data using Bagged Regression Trees

Yang Xie, *Student Member, IEEE*, Günter Schreier, *Senior Member, IEEE*, David C.W. Chang, Sandra Neubauer, Stephen J. Redmond, *Senior Member, IEEE*, Nigel H. Lovell, *Fellow, IEEE*

***Abstract*— Healthcare administrators worldwide are striving to both lower the cost of care whilst improving the quality of care given. Therefore, better clinical and administrative decision making is needed to improve these issues. Anticipating outcomes such as number of hospitalization days could contribute to addressing this problem. In this paper, a method was developed, using large-scale health insurance claims data, to predict the number of hospitalization days in a population. We utilized a regression decision tree algorithm, along with insurance claim data from 300,000 individuals over three years, to provide predictions of number of days in hospital in the third year, based on medical admissions and claims data from the first two years. Our method performs well in the general population. For the population aged 65 years and over, the predictive model significantly improves predictions over a baseline method (predicting a constant number of days for each patient), and achieved a specificity of 70.20% and sensitivity of 75.69% in classifying these subjects into two categories of 'no hospitalization' and 'at least one day in hospital'.**

I. Introduction

Administrators of healthcare systems worldwide are striving to both lower the cost of care whilst improving the quality of care given. To simultaneously improve healthcare quality while saving on unnecessary medical expenditure, better clinical and administrative decision making is needed. It has been shown that novel and deep insights gained by analyzing clinical data can make a significant contribution to the process of diagnosis, choice of treatment, and prognostic predictions [1].

Since the more recent rise of Big Data, the availability of large heterogeneous medical datasets has prompted the use of data mining techniques to discover important information contained within these complex data structures. Harper [2] and Yoo *et al.* [1] have written thorough reviews of popular data mining algorithms used in the context of healthcare and biomedicine. One of the most important applications, from both clinical and administrative viewpoints, is to predict individual patient outcomes. Anticipating outcomes such as length of stay (LoS) in hospital, hospital readmission, health costs, and onset of a particular disease, are among the most coveted of prediction capabilities [3], [4].

This research was supported under the Australian Research Council's Linkage Projects funding scheme (project number LP0883728) and HCF Foundation.

Y. Xie, D.C.W. Chang, S.J. Redmond and N.H. Lovell are with the Graduate School of Biomedical Engineering, UNSW Australia, Sydney, New South Wales, 2052. n.lovell@unsw.edu.au

G. Schreier and S. Neubauer are with AIT Austrian Institute of Technology GmbH, 8020 Graz, Austria. guenter.schreier@ait.ac.at

Limited previous research has demonstrated promising results in predicting LoS days for patients in special disease groups using physiological measurement data. Silberbach *et al.* [5] predict LoS for congenital heart disease surgery; Suter-Widmer *et al.* [6] investigated the possibility of predicting hospital stays in patients with community-acquired pneumonia. On the other hand, Harper [2] applied classification algorithms to more general datasets (i.e., an intensive care unit (ICU) dataset and a hospital management system dataset). Multiple targets (i.e., LoS in ICU, LoS in hospital) were predicted based on these general datasets. However, sample sizes were small (with 582 records) or medium (with 17,974 records), which could be a limit in accurately characterizing the population. The most comparable work was conducted by Bertsimas *et al.* [7]. An excellent predictive model was built on a big dataset of 800,000 individuals to predict the possibility of a customer falling into five different health cost groups. They point out that medical information contributes more to predictions of high-cost members than lost-cost members.

The aim of this paper is to build a model that predicts the number of hospitalization days for a general population, using large-scale health insurance claims data. Through this analysis, it is intended to capture the uncertainty and variability among a patient population by predicting individual patient outcomes. In addition, since customers of 65 years and older are more frequent users of medical resources, we are particularly interested in the performance of such a predictive model on this group.

II. Methods

A. Data Set and Summary Statistics

The dataset was obtained from 261,558 de-identified customers of the Hospitals Contribution Fund of Australia (HCF), one of Australia's largest combined registered private health fund and life insurance organizations. Three consecutive years of data, from 2010 to 2012, were provided for analysis. These data contain tables specifying the patient demographics, hospital admission, and medical services provided:

Customer demographics table includes information related to customers themselves, such as gender, age, the type of HCF product they are subscribed to, the date they joined the fund, and several other customer-level information items. In nearly every case a customer will have one entry in this table. Duplicates are aggregated to form a single unique customer entity.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

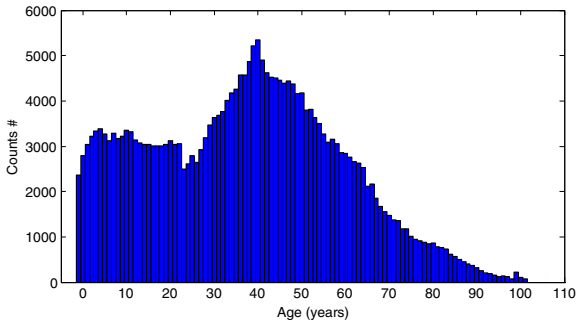

Fig. 1. Customer count segregated by age.

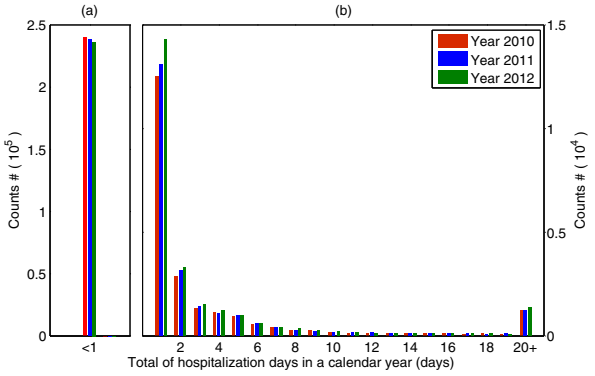

Fig. 2. Hospitalization days histogram for each of the three years of HCF data. According to this figure, the lengths of most hospitalizations is one day.

Hospital admission table includes information related to admissions. Each hospital admission has an entry here. Data fields include illness code, admission date, discharge date, number of hospitalization days, and several other admission-level information items. A customer could have multiple admissions within each year.

Hospital service claim table includes information related to claims. Each claim is related to a hospital admission. An admission could have multiple claims, such as hospital claims, medical claims, prosthesis claims.

Some illustrative figures reflect broad characteristics of the whole dataset. In Fig.1, the distribution of different age groups is shown. In Fig. 2, the distribution of hospitalization days is shown. The lengths of most hospitalizations are one day. Note however that both a same-day hospital admission and a one-day overnight hospital admission are marked as one day in this dataset. In Fig. 3, the average number of hospital days per patient on a yearly basis are shown, segregated by age.

It needs to be pointed out that only the services which are covered by HCF are included in this database – in Australia, Medicare (a publicly funded universal healthcare scheme) concurrently covers most non-hospital services, and these data are not available.

B. Accuracy Calculation and Baseline

The main performance indicator is referred to as the root-mean-square-error (*RMSE*), and is the root-mean-square of the difference between the logarithm of the estimated number

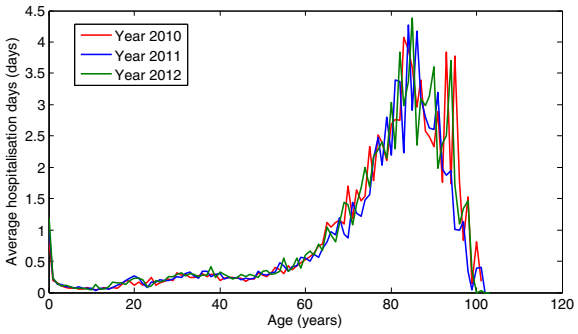

Fig. 3. Average hospitalization days per person segregated by age for each the three years of HCF data.

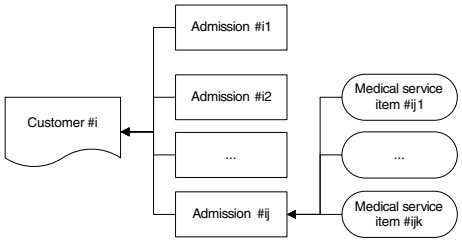

Fig. 4. Three-level structure of dataset: patient demographics, hospital admissions, and medical services

of days in hospital and the logarithm of the true number of days. The logarithm is offset by +1 to avoid a logarithm of zero. The logarithm is used to reduce the importance assigned to those with many hospital days.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[\log(p_i+1) - \log(a_i+1)]^2} \quad (1)$$

Here p_i are the predicted days in hospital for each patient, and a_i are the actual days in hospital. This accuracy may be compared to the accuracy which can be obtained by predicting a constant number of days for each patient. All customers were categorized into three groups. Group 1 includes the whole population. Group 2 includes customers younger than 65 years, while Group 3 is composed of customers aged 65 years old and over. The best baseline accuracies (*RMSE*) for constant hospitalization days prediction of Group 1, 2 and 3, were 0.440, 0.352 and 0.785, respectively. This demonstrates what is shown in Fig. 3, customers of 65 years and older (Group 3) on average tend to have more hospitalization days than customers in Group 2.

C. Data Manipulation

As mentioned in Section II-A, the data are organized in three levels: patient demographics, hospital admissions, and medical services. The data structure is shown in Fig. 4. Since our aim is to predict the number of days spent in hospital for each customer in the subsequent year, information at the admission and medical service levels has been aggregated to the customer level for modeling purposes. The aggregation is conducted at the feature extraction stage, as described in Section II-D.

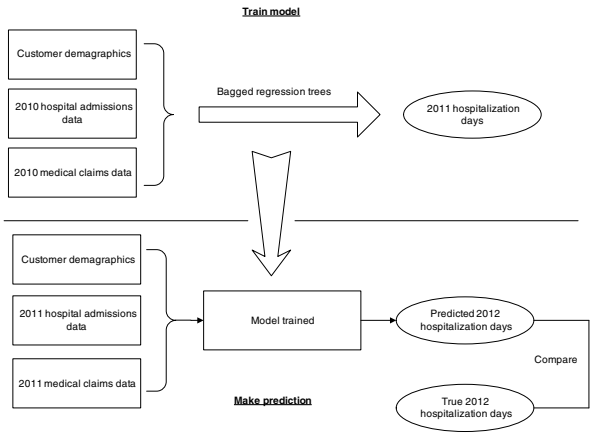

Fig. 5. One-year-model approach. Customer demographics, admission and medical claim data for year 2010, along with 2011 days in hospital were used to train model. Later, at the prediction stage, customer demographics, hospital admission and medical claim data for year 2011 were used to predict the hospitalization days in 2012.

The prediction model is created using the following 'one-year-model' methodology, as shown in Fig. 5. The number of hospitalization days were predicted for the third year, based on admission and medical claim data from the first two years, along with customer demographics. As shown in Fig. 5, admission and medical claim data in the first year was used for training while the data from the second year was used as features for making predictions.

D. Feature Extraction

Table I lists the source data from which features were extracted. The column 'feature level' indicates which level the source variable is from.

The source variables contain various formats of data, such as dates, numeric, text; therefore, source variables needed to be pre-processed before they can be used for modeling. Most of the variables are either numeric or categorical. Numeric variables were kept in their numeric format. For all the categorical variables, values were replaced by integers for the purpose of making the whole feature matrix numeric and conserving computing memory. For some of the categorical variables, nominal features were also extracted by generating a separate binary column for each of the categories in the variable, as a category indicator. All variables with time formats were transformed into a date number, making them numeric. Descriptive statistical methods were applied when aggregating features extracted from the admission and medical claim levels into the customer level. The descriptive statistics methods used included taking the sum, mean, standard deviation, median, maximum, minimum values, and other statistical properties, when aggregation occurred.

E. Predictive Modeling

A predictive model was built using bagged regression trees. Every tree in the ensemble is grown on an independently drawn bootstrap replica of data. Observations not included in this replica are out-of-bag for this tree. To compute prediction of an ensemble of trees for unseen data, it takes an average of predictions from individual trees. Those out-of-bag observations are used as a validation data set to estimate the prediction error, which could help avoid severe over-fitting. Here a MATLAB function named 'TreeBagger' was used to implement the algorithm. The number of trees in the bagging ensemble was 50.

TABLE I

A LIST OF THE SOURCE VARIABLES, FROM WHICH FEATURES WERE EXTRACTED. '*' INDICATES A KEY ATTRIBUTE. ALL KEY VARIABLES HERE ARE DE-IDENTIFIED IDS AND WERE NOT USED AS FEATURES.

Feature level	Source variable name	Description
Customer	*CUSTOMER_DEID	Deidentified ID of customer
	BIRTH_DTE	Date of birth of customer
	CLIENT_STATUS	Status of client. Currently active customers or customers who has terminated their memberships.
	DTE_JOIN_FUND	Date the customer joined the fund
	DTE_ON	Date the current product came into effect
	MEMBER_TYPE	The type of membership
	PCODE	Post code
	PRODUCT	Type of the product
	GENDER	Gender
	TITLE_CODE	Title of customer
	RELATIONSHIP	Dependant relationship
Admission	*ADMISSION_NO	De-identified ID of hospital admission
	DTE_ADMITTED	Date the patient commenced an episode of care
	DRG	Diagnosis related group
	HOSP_TYPE	Public or private hospital
	ICD10_PRINC_DIAG	ICD-10-AM code for the diagnosis or condition chiefly responsible for occasioning the hospital admission
	ILLNESS_CODE	Compound coding for the primary illness responsible for the hospital admission
	DTE_SEPARATED	Date the patient completed an episode of care
	PARTICIPATED_IND	Whether HCF have an agreement with the hospital or not
	SECOND_TIER_CAT	Categorization of hospitals
	SPECIALTY_DESC	Description of specialty
	STAY_STATUS	Hospital stay status (i.e., same-day or overnight accommodation)
	TREATMENT_TYPE	Type of treatment
Claim	*CLAIM_NO	De-identified ID of medical service claim
	AMT_CHARGED	Total amount charged for the service item
	DTE_SERVICE	Date of service delivered
	SURGICAL	The patient classification code used to determine hospital accommodation benefits
	HOSP_ITEM_CLASS	Type of hospital service item
	ITEM_CODE	The item number
	PROVIDER_TYPE	Type of provider
	Provider_Number	Number of provider

III. RESULTS

Table II lists the performance metrics for the proposed method. Shown are the results on the training dataset and testing dataset separately, for the three population groups described in Section II-B. The *RMSE* as calculated using the performance measure described in Eq. (1), by comparing the predicted hospital days against the actual hospital days. A Spearman correlation coefficient (ρ) is calculated also by comparing the prediction against the actual values.

Customers can be categorized into two categories, without hospital days (0 hospital days) and with hospital days (at least 1 hospital day). Therefore, binary analysis is applicable to the result. By setting a threshold in the predicted hospital days, statistics including accuracy (Acc.), specificity (Spec.), sensitivity (Sens.), Cohen's kappa (κ), and area under ROC curve (AUC) metrics were calculated. The optimal results were obtained with a threshold of 0.3 days applied to the continuous output estimate from the bagged tree regression model. When the predicted value was smaller than 0.3 days, we consider it as a prediction of no hospital days and vice versa. The binary analysis results are also displayed in Table II.

Fig. 6 shows scatter-plots of the regression results for the group aged 65 years and over. The subplot on the left is for training and while the right subplot is for testing.

TABLE II

PERFORMANCE METRICS FOR THE PROPOSED METHOD. SHOWN ARE THE RESULTS FOR THE TRAINING DATASET AND TESTING DATASET.

Result	Train[1]	Test [1]	Train[2]	Test[2]	Train[3]	Test[3]
RMSE	0.252	0.402	0.215	0.333	0.428	0.711
ρ	0.494	0.316	0.455	0.256	0.678	0.452
Spec. (%)	95.50	90.69	97.60	93.71	78.95	70.20
Sens. (%)	86.75	41.41	81.98	26.85	98.97	75.69
Acc. (%)	94.72	85.89	96.46	88.31	82.98	71.34
κ	0.717	0.286	0.753	0.207	0.595	0.345
AUC	0.981	0.796	0.986	0.761	0.970	0.809

[1] Group 1: all customers.
[2] Group 2: customers less than 65 years old.
[3] Group 3: customers more than 65 years old.

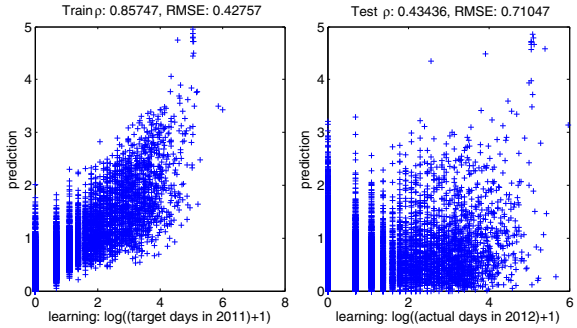

Fig. 6. Scatter-plots for regression results for customers of 65 years and older.

IV. DISCUSSION AND CONCLUSION

A method for predicting future hospitalization days has been developed using features extracted from customer demographics, past hospital admission and medical claim data. The model developed using a training dataset was later evaluated using an unseen testing dataset. In summary, we observe that this method improves predictions over the baseline method (as described in Section II-B) by reducing the *RMSE* measure. As displayed in Table II, on the whole population (Group 1), the *RMSE* measure is lowered from the baseline accuracy 0.440 to 0.402, while for Group 2 (customers below 65 years old), it is improved from a baseline accuracy of 0.352 to 0.333. For those aged 65 and above, this measure gets better from a baseline accuracy of 0.785 down to 0.711.

Table II also shows that the model achieved a high specificity (93.71%) and a relatively low sensitivity (26.85%) on the test set of customers under 65 years old. There could be a variety of reasons causing the low sensitivity. One of the possible explanations could be that young customers in general have less hospitalization days than older customers, as demonstrated in Fig. 3. Moreover, young customers seldom have hospital days or are hospitalized for relatively acute conditions which are unpredictable or unexpected; therefore by its nature, it is difficult to make predictions for younger people.

In contrast, on the test set of customers of 65 years or older, a moderate specificity (70.20%) and sensitivity (75.69%) were obtained. The specificity dropped in comparison to that of the other two groups. Although from the viewpoint of a health insurer, a drop in specificity may not be of critical concern. If the estimated hospitalization does not occur, there is no cost incurred. However, the underlying reasons are worthy of investigation and could be social and/or economic. For instance, going into hospital can be daunting for elderly people. They may avoid attending hospital for fear. Therefore, although the model estimates that the customer would be hospitalized, this prediction does not eventuate in reality. Reasons for this warrant further investigation.

Since people aged 65 years and older are heavy users of medical resources, it is more interesting to know the performance of this method on older customers. Comparing the results of three groups, a conclusion could be made that this model performs 'best' for the senior population, achieving the highest correlation ρ, Cohen's kappa κ, and AUC (area under receiver operating characteristics curve). Bertsimas *et al.* also concluded that their method achieved more significant improvements for more costly members [7]. However, over-fitting at the training stage is also noticed in Fig. 6, which can also be observed in Table II by comparing the statistics between the training and testing sets.

To better improve the result, the following aspects are worthy of exploration in the future. Hospitalization days can be further categorized in various graded categories, as this is a good predictor for targeted interventions from an insurance perspective. Health domain knowledge can be incorporated, which would contribute to generating more meaningful features. Analysis of data with longer observation periods could also help in making more accurate predictions. In addition, if clinical measurements records (i.e., physiological monitoring from telehealth) could be obtained from patients, it is likely that this additional information would greatly benefit the accuracy of predictions for older customers, especially those with chronic disease.

REFERENCES

[1] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J.-F. Chang, and L. Hua, "Data mining in healthcare and biomedicine: a survey of the literature," *Journal of Medical Systems*, vol. 36, no. 4, pp. 2431–2448, 2012.

[2] P. R. Harper, "A review and comparison of classification algorithms for medical decision making," *Health Policy*, vol. 71, no. 3, pp. 315–331, 2005.

[3] P. R. Hachesu, M. Ahmadi, S. Alizadeh, and F. Sadoughi, "Use of data mining techniques to determine and predict length of stay of cardiac patients," *Healthcare Informatics Research*, vol. 19, no. 2, pp. 121–129, 2013.

[4] O. Hasan, D. O. Meltzer, S. A. Shaykevich, C. M. Bell, P. J. Kaboli, A. D. Auerbach, T. B. Wetterneck, V. M. Arora, J. Zhang, and J. L. Schnipper, "Hospital readmission in general medicine patients: a prediction model," *Journal of General Internal Medicine*, vol. 25, no. 3, pp. 211–219, 2010.

[5] M. Silberbach, D. Shurnaker, V. Menashe, A. Cobanoglu, and C. Morris, "Predicting hospital charge and length of stay for congenital heart disease surgery," *The American Journal of Cardiology*, vol. 72, no. 12, pp. 958–963, 1993.

[6] I. Suter-Widmer, M. Christ-Crain, W. Zimmerli, W. Albrich, B. Mueller, P. Schuetz *et al.*, "Predictors for length of hospital stay in patients with community-acquired pneumonia: Results from a swiss multicenter study," *BMC Pulmonary Medicine*, vol. 12, no. 1, p. 21, 2012.

[7] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang, "Algorithmic prediction of health-care costs," *Operations Research*, vol. 56, no. 6, pp. 1382–1392, 2008.