Journal of Biomedical Informatics 60 (2016) 187–196

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Analyzing health insurance claims on different timescales to predict days in hospital ☆

Yang Xie [a,*], Günter Schreier [b], Michael Hoy [a], Ying Liu [a], Sandra Neubauer [b], David C.W. Chang [a], Stephen J. Redmond [a], Nigel H. Lovell [a]

[a] *The Graduate School of Biomedical Engineering, UNSW Australia, Sydney, New South Wales 2052, Australia*
[b] *AIT Austrian Institute of Technology GmbH, 8020 Graz, Austria*

ARTICLE INFO

Article history:
Received 16 April 2015
Revised 5 January 2016
Accepted 5 January 2016
Available online 28 January 2016

Keywords:
Health care
Temporal data mining
Predictive modeling
Health insurance claims

ABSTRACT

Health insurers maintain large databases containing information on medical services utilized by claimants, often spanning several healthcare services and providers. Proper use of these databases could facilitate better clinical and administrative decisions. In these data sets, there exists many unequally spaced events, such as hospital visits. However, data mining of temporal data and point processes is still a developing research area and extracting useful information from such data series is a challenging task. In this paper, we developed a time series data mining approach to predict the number of days in hospital in the coming year for individuals from a general insured population based on their insurance claim data. In the proposed method, the data were windowed at four different timescales (bi-monthly, quarterly, half-yearly and yearly) to construct regularly spaced time series features extracted from such events, resulting in four associated prediction models. A comparison of these models indicates models using a half-yearly windowing scheme delivers the best performance on all three populations (the whole population, a senior sub-population and a non-senior sub-population). The superiority of the half-yearly model was found to be particularly pronounced in the senior sub-population. A bagged decision tree approach was able to predict 'no hospitalization' versus 'at least one day in hospital' with a Matthews correlation coefficient (MCC) of 0.426. This was significantly better than the corresponding yearly model, which achieved 0.375 for this group of customers. Further reducing the length of the analysis windows to three or two months did not produce further improvements.

1. Introduction

Hospitalization is usually the largest component of health expenditure [1]. Early identification of those with a higher risk of hospitalization could help in making efficient health resource management decisions. There exists work which attempts to predict hospitalizations for specific disease groups based on laboratory data and medical records. Niewoehner et al. [2] developed hospitalization risk indexes for chronic obstructive pulmonary disease (COPD) patients using their spirometry, demographics, and medical history data. Sugimoto et al. [3] discovered that serum intact parathyroid hormone levels obtained in outpatients with heart failure were shown to be an independent predictor of hospitalization. However, limited work has investigated the possibility of developing a general model which is not disease specific, which can set a benchmark against which disease-specific models may be compared. To our best knowledge, the most relevant recent work includes a data mining competition ('Heritage Health Prize') [4] and research done by the authors [5]. Both studies developed models to predict days in hospital for a general population based on health insurance claims.

In our previous work [5], we utilized bagged decision tree classifier models to predict the total number of days spent in hospital in the subsequent calendar year for individuals from a general population, using large-scale health insurance claims data. The proposed method performs well in the general population as well as different demographic sub-populations.

However, information available to perform the prediction comes from two sources; firstly, a list of customer attributes (e.g., customer demographics and insurance enrollment information), and secondly, the history of customer hospital admission claims. The latter can be considered as a time series of unequally

☆ This research was supported under the Australian Research Council's Linkage Projects funding scheme (Project No. LP0883728) and HCF Foundation.

* Corresponding author.

E-mail addresses: yang.xie@unsw.edu.au (Y. Xie), guenter.schreier@ait.ac.at (G. Schreier), n.lovell@unsw.edu.au (N.H. Lovell).

http://dx.doi.org/10.1016/j.jbi.2016.01.002
1532-0464/© 2016 Elsevier Inc. All rights reserved.

spaced events. Furthermore, each hospitalization event can itself be expressed as a sequence of hospital services utilized by the patient during that particular stay [5]. While it is straightforward to directly present the first type of information as features to a classification algorithm, data mining of time series and point processes, such as the hospitalization events described above, is still a developing research area. It is rather challenging to extract relevant information in a useful manner. The relative timing of recorded hospital visits used for prediction seems likely to contain subtle, yet valuable information which may lead to more accurate predictions if it can be harvested properly.

Reviews of methods applicable to a broader class of *temporal data mining* problems [6,7] highlight the following approaches which are applicable to the analysis of point process data:

1.1. Conversion to time series

When point processes occur relatively frequently, they can be converted to an equally-spaced time series. Prediction of future values from such time series is a well studied area (see [8,9]). However, this is not straightforward with the complex data encountered here, and information would be lost in the conversion process.

1.2. Estimation of the similarity between time series data

If a distance or similarity measure can be developed, then non-parametric classification approaches such as *K-nearest neighbors* (KNN) can be used. This returns a quantity based on the set of the closest matches in the training data. An example is *Edit Distance* (the number of alterations required to convert one time series to another) [10,11]. However, such methods may be too computationally costly for very large data sets like in the problem at hand.

1.3. Estimation of the underlying risk process

Poisson and other similar generalized linear models are often used for insurance purposes, and these are useful for determining the distribution of the possible number of claims [12,13]. Poisson regression makes the assumption that the response variable is drawn from a Poisson distribution. The method also assumes the logarithm of the Poisson distribution's expected value can be modeled by a linear combination of selected model parameters. Another approach taken is to model the system as a Cox process, which is a Poisson process where the expected value of the distribution changes over time in terms of the known factors [14]. However, both Poisson and Cox regression frameworks assume the response variable to be non-linearly and monotonically proportional to the explanatory variables. While useful, in exploratory studies like the one presented here, non-linear interactions and correlations between exploratory variables may be too complex for a log-linear model like this to capture, ultimately leading to poorer performance relative to more flexible pattern recognition models.

1.4. Symbolic point data mining

If all scalar values and perhaps time intervals between observations are discretized into categorical variables, the entire time series can be expressed as a sequence of symbols. The key step in such methods is defining a language that can adequately represent the temporal dimension of the data [15]. Most work relies on the use of temporal abstraction (TA) [16] and temporal logic [17], which allow the description of complex temporal patterns and temporal interactions among multiple time series. TA is the first step, which is the process of segmenting and aggregating time series data into explicit and symbolic representations, making it suitable for human decision making or data mining [18]. Next is mining these temporal patterns derived through TA, which is a relatively young research field. Most work mine temporal association rules (TARs), based on Allen's temporal relations, which represent temporal events using before/after chains (e.g., event A precedes/overlaps/f inishes-by/contains event B) [17]. There have been several reports of the application of these methods to healthcare data sequences, which comprised hybrid events temporally interacting with each other (e.g., medications and physiological measurements) [15,19,20]. However, in the claim data set used for this study, no medication data, laboratory results or physiological measurements were available, and events were temporally sparse, with the average number of procedure claims per customer per year less than two (Table 2). Given the sparse time series and the lack of temporal heterogeneity between feature time series, further confounded by the vast number of potentially predictive features available in the big data set used here, it was decided not to pursue such an approach in this paper.

1.5. Heuristic time series features

Some features can be invented heuristically based on properties of the time series [21,22]. Indeed in one such study, they were found to give better performance compared to other model-based methods [21]. Our previous reported approach [5] falls into this category, in which most of the medical features extracted from properties of the hospital admission records and procedure claims were aggregated for each customer. This was found to give good performance and is meaningful as a first attempt to solve this problem. However, processing features in this way is expected to lead to a loss of temporal information.

The objective of this study is to develop a method which includes time information more explicitly and evaluate its performance on predicting the number of hospitalization days for a general population. Considering that the density of hospital admission events during a year is relatively sparse or zero for most of the claimants, a mixture of heuristic time series features and windowing was utilized. Specifically, features extracted from medical events are sorted into smaller time intervals to bring in more detailed information about the relative timing of events, which would potentially improve prediction performance. In addition, the time intervals are varied through a set of different scales, i.e., a year, half year, quarter, and two months. Our aim is to explore the impact of varying the temporal resolution on the predictive power of such models. At the same time, a model evaluation and comparison routine is proposed to assess whether the differences in the performance of different time scale models is statistically significant.

2. Methods

2.1. Data set

The data set consisted of 100,000 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. The 100,000 subjects were randomly selected from the HCF customer database. Only those customers who had enrolled with HCF before 1/1/2011 were included in the selection process. Three consecutive years of data, from 2011 to 2013, were provided for analysis.

These data contain tables of hospital admission administrative records and hospital procedure claims, as well as basic demographic information of customers [5]. Customer demographics include information related to customers, such as sex, age, the type of HCF product they were subscribed to, the date they joined the fund, and other personal information. Hospital admission records include data fields of primary diagnosis code, admission date, discharge date, length of stay (LoS), and several other information items related to admission. Hospital procedure claims include information related to the procedure delivered during a hospital admission, like date and type of procedures as well as information on the associated costs. Each claim is related to a hospital admission.

Fig. 1 lists the structure of the data set and source data variables from which features were extracted.

Number of customers arranged by age, sex, and year for which they were customers of HCF (2011–2013), were shown in Table 1. Since the cohort is fixed to the same 100,000 customers across three years, we see the number of customers in all the 40+ age categories increase each year, while the 20–39 age groups decrease in size. Table 2 displays average values in key demographic and claim statistics across three years.

It is important to note here that, in Australia, records of all registered deaths are kept by state and territory governments. These data are not shared with health insurers; therefore, we did not exclude those that died during the study period. Customers who

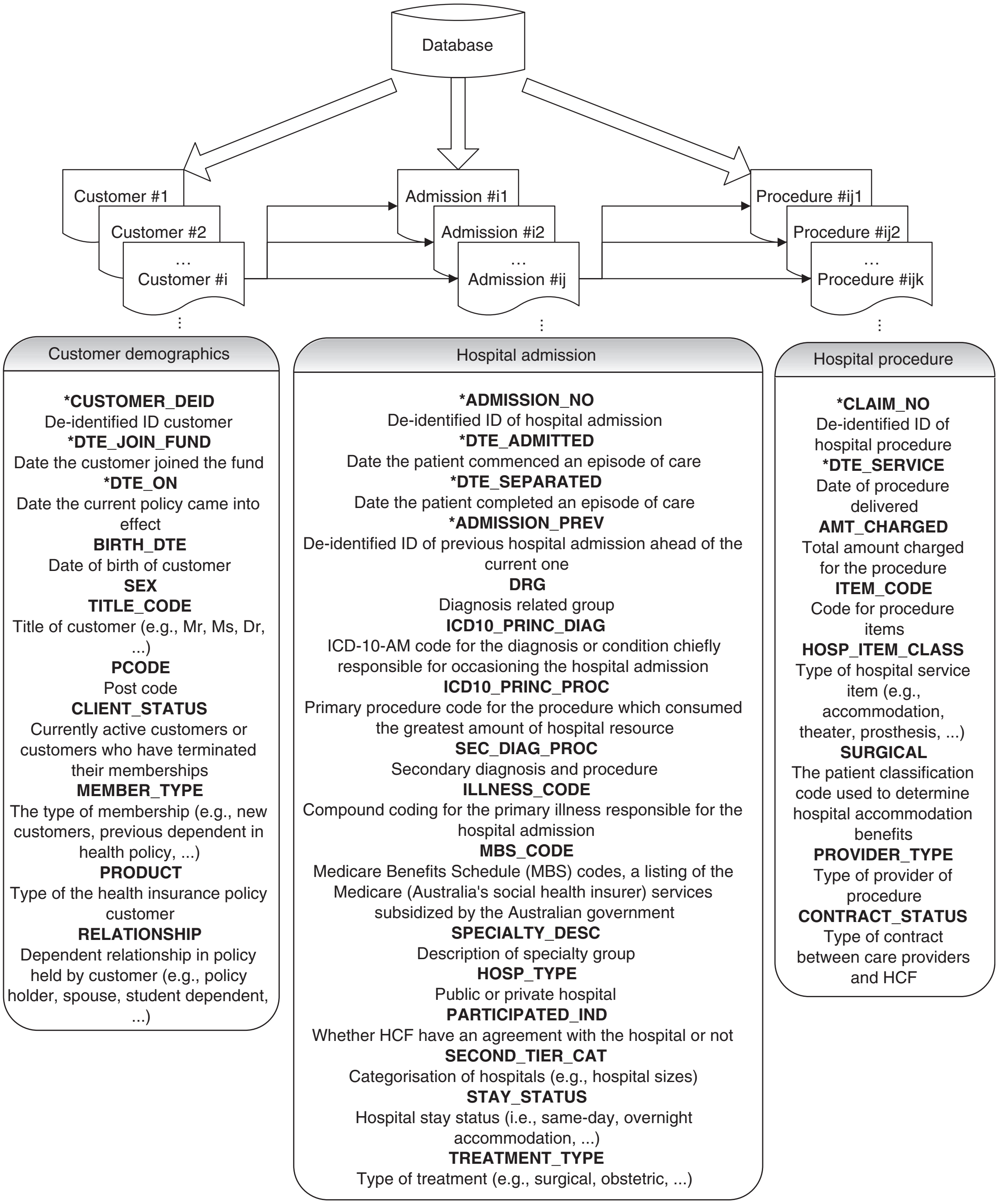
Database
Customer #1
Customer #2
...
Customer #i
Admission #i1
Admission #i2
...
Admission #ij
Procedure #ij1
Procedure #ij2
...
Procedure #ijk
Customer demographics
*CUSTOMER_DEID
De-identified ID customer
*DTE_JOIN_FUND
Date the customer joined the fund
*DTE_ON
Date the current policy came into effect
BIRTH_DTE
Date of birth of customer
SEX
TITLE_CODE
Title of customer (e.g., Mr, Ms, Dr, ...)
PCODE
Post code
CLIENT_STATUS
Currently active customers or customers who have terminated their memberships
MEMBER_TYPE
The type of membership (e.g., new customers, previous dependent in health policy, ...)
PRODUCT
Type of the health insurance policy customer
RELATIONSHIP
Dependent relationship in policy held by customer (e.g., policy holder, spouse, student dependent, ...)
Hospital admission
*ADMISSION_NO
De-identified ID of hospital admission
*DTE_ADMITTED
Date the patient commenced an episode of care
*DTE_SEPARATED
Date the patient completed an episode of care
*ADMISSION_PREV
De-identified ID of previous hospital admission ahead of the current one
DRG
Diagnosis related group
ICD10_PRINC_DIAG
ICD-10-AM code for the diagnosis or condition chiefly responsible for occasioning the hospital admission
ICD10_PRINC_PROC
Primary procedure code for the procedure which consumed the greatest amount of hospital resource
SEC_DIAG_PROC
Secondary diagnosis and procedure
ILLNESS_CODE
Compound coding for the primary illness responsible for the hospital admission
MBS_CODE
Medicare Benefits Schedule (MBS) codes, a listing of the Medicare (Australia's social health insurer) services subsidized by the Australian government
SPECIALTY_DESC
Description of specialty group
HOSP_TYPE
Public or private hospital
PARTICIPATED_IND
Whether HCF have an agreement with the hospital or not
SECOND_TIER_CAT
Categorisation of hospitals (e.g., hospital sizes)
STAY_STATUS
Hospital stay status (i.e., same-day, overnight accommodation, ...)
TREATMENT_TYPE
Type of treatment (e.g., surgical, obstetric, ...)
Hospital procedure
*CLAIM_NO
De-identified ID of hospital procedure
*DTE_SERVICE
Date of procedure delivered
AMT_CHARGED
Total amount charged for the procedure
ITEM_CODE
Code for procedure items
HOSP_ITEM_CLASS
Type of hospital service item (e.g., accommodation, theater, prosthesis, ...)
SURGICAL
The patient classification code used to determine hospital accommodation benefits
PROVIDER_TYPE
Type of provider of procedure
CONTRACT_STATUS
Type of contract between care providers and HCF

Fig. 1. Structure of the data set and the source elements, from which features were extracted. '*' indicates a key attribute. These key variables are either de-identified IDs or timestamp variables, which could be used to compute additional features. For instance, ADMISSION_PREV was used to compute the number of days from the previous admission to the current admission. However, IDs and timestamps themselves were not directly used as features. Note that the difference between 'Client' and 'Customer' is that everybody in the same family may share the same client ID, but everyone has a unique customer ID.

Table 1
Number of customers arranged by age, sex, and year for which they were customers of HCF (2011–2013). Since the cohort is fixed to the same 100,000 customers across three years, we see the number of customers in all the 40+ age categories increase each year, while the 20–39 age groups decrease in size.

Age	Female			Male			Sex unknown			All sexes		
	2011	2012	2013	2011	2012	2013	2011	2012	2013	2011	2012	2013
Unknown	3284	2942	2662	2343	1956	1638	7477	7477	7477	13,104	12,375	11,777
< 10	4932	4781	4531	5330	5216	4957	5	4	4	10,267	10,001	9492
10–19	5087	5064	5154	5346	5331	5390	3	3	3	10,436	10,398	10,547
20–29	5209	5103	4954	4461	4510	4598	218	185	144	9888	9798	9696
30–39	7094	6907	6696	5974	5669	5416	600	575	555	13,668	13,151	12,667
40–49	7861	7927	7957	7053	7143	7177	521	542	564	15,435	15,612	15,698
50–59	6299	6540	6768	5422	5676	5826	295	312	326	12,016	12,528	12,920
60–69	4022	4264	4483	3935	4080	4244	151	162	175	8108	8506	8902
70–79	1927	2056	2238	1934	2073	2267	72	77	86	3933	4206	4591
80+	1692	1823	1964	1346	1490	1631	107	112	115	3145	3425	3710
All ages	47,407	47,407	47,407	43,144	43,144	43,144	9449	9449	9449	100,000	100,000	100,000

Table 2
Average values in key demographic and claim statistics across three years (2011–2013).

Demographics	2011	2012	2013
Average age (years)	33.425	34.294	35.170
Average number of admissions per customer	0.155	0.168	0.182
Average days in hospital per customer (days)	0.374	0.412	0.438
Average days in hospital per admission (days)	2.413	2.452	2.407
Average amount charged per customer (AUD[a])	521.267	594.162	645.821
Average amount charged per admission (AUD[a])	3372.151	3527.649	3558.829
Average number of procedure item codes assigned per customer	1.533	1.706	1.831
Average number of procedure item codes assigned per admission	9.919	10.127	10.089
Average number of DRG codes assigned per customer	0.0986	0.107	0.115
Average number of ICD10-AM primary diagnosis codes assigned per customer	0.0973	0.106	0.113
Total number of admissions	15,458	16,843	18,147
Total number of days in hospital (days)	37,347	41,162	43,779
Total amount charged (AUD[a])	52,126,703	59,416,188	64,582,066
Total number of DRG codes assigned	9862	10,707	11,454
Total number of ICD10-AM primary diagnosis codes assigned	9729	10,588	11,296
Total procedure item codes assigned	153,330	170,569	183,091
Sex[b]	2011–2013		
	F: 47 407	M: 43 144	U: 9 449

[a] AUD is abbreviated for Australian Dollars.
[b] F: female; M: male; U: sex unknown.

died in a given year will be considered as having zero days in hospital in the subsequent years.

2.2. Data pre-processing

Since the source variables contain various data types, such as dates, numeric, text, they needed to be pre-processed before they could be used for modeling. Numeric variables were kept in their numeric format. All categorical variables were enumerated, i.e., replaced by integers for the purpose of making the whole feature matrix numeric and conserving computing memory. For some of the categorical variables, binary features were also extracted by generating a separate column for each of the categories, as a category indicator. Some additional features were also generated using specific calculations. For example, from the secondary diagnoses, co-morbidity scores were computed, utilizing the respective look-up tables, that give weights to certain ICD-10 diagnoses according to the original and the updated Charlson Index [23,24], respectively.

2.3. Bins and feature set

Since we were attempting to predict the number of days in hospital for each customer in a particular year, the prediction was performed at the customer level. In a previously reported method by our research group [5], to further process the feature matrix, all non-customer level information were aggregated into the customer level and sorted into the respective years. This allowed a feature array to be generated for each customer for each year. Since each customer could have multiple hospital admissions and procedures in a year, when sorting these information into a year, descriptive statistical methods were performed to summarize the events that happened in that year [5]. For example, amount charged for multiple procedures in a year would be summarized to an average amount charged for all procedures in this year. Therefore, features used in the reported method were an overall depiction of a year, and variations in respective quarters or months are obviously not represented by this method of aggregation, thereby leading to a loss of temporal resolution.

The present study broadens the concept of bins from fixed yearly bins to flexible time intervals such as half year, quarter, and two months. Features extracted from the hospital admissions and the hospital procedures, were sorted into corresponding bins based on the timestamps associated with them. Similarly, suitable methods were applied during aggregation. For some of the numeric features, such as prior cost information and prior number of days in hospital (DIH), descriptive statistics were computed, including the sum, mean, standard deviation, median, maximum and minimum values, range between maximum and minimum, the most recent element and the most frequent element (if multiple elements occur equally frequently, the element with the smallest value or the element occurring first alphabetically will be considered as the most frequent). Similarly, for categorical variables, when aggregating, descriptive statistics, such as count of unique elements, the most frequent element, and the most recent element were used. If not specified, other features were summed during aggregation.

This aggregated all events into regular temporal bins, and used each bin as a separate feature. Equally spaced time series were then constructed in this way for each feature extracted from temporal events. The final feature table contained a feature array for each customer and each temporal bin. Finally, features that do not vary across bins, such as demographic data like sex, were prevented from being duplicated into multiple bins using a post-processing step. It should be noted that for hospital stays which cross multiple time bins, the entire stay was credited to the bin associated with the date of admission.

Fig. 2 demonstrates the process for converting a series of events with categorical variables and numeric variables into a time series, with an example of a categorical feature (most frequent ICD-10 primary diagnosis) and a numeric feature (number of admissions). The bins used in this example are either one quarter or one year in duration.

2.4. Predictive methods

2.4.1. Temporal models

Four temporal models with four different bin settings were developed: yearly model, half-yearly model, quarterly model, and bi-monthly model.

Fig. 3 illustrates the temporal models. The yearly model and quarterly model are used as examples. As displayed, in the yearly model, three years of data were divided into three yearly bins, while in the quarterly model, each year had four quarter bins and thus twelve bins in total across the three years. For both models, the initial two years (01/01/2011 to 31/12/2012) served as a two-year training period, and the one year prediction period was from 1/1/2013 to 31/12/2013. Since we were forecasting days in hospital for a subsequent year, the feature matrix and target variable, i.e., the entity to be predicted, should come from different years. The feature matrix for training (Training X) was from year 2011, while target for training (Training Y) was from year 2012. The feature matrix (Prediction X) from year 2012 was then input to the trained model to make a prediction. The prediction outcome (Prediction Y) was a forecast of days in hospital for the year 2013. This was later compared to the actual days in hospital in 2013, and a variaty of indicators were calculated to assess performance.

Table 3 lists the settings of the four models. Like the yearly model and quarterly model in Fig. 3, the other two models are a half-yearly model with bin length of six months, and a bi-monthly model with bin length of two months. Table 3 gives the indices to the bins for the training and prediction periods, corresponding to the Training X, Training Y, Prediction X, Prediction Y in Fig. 3.

All of the four models had the same training periods and prediction periods. For example, for bi-monthly model, Training X covered the first six bins (each bin with a length of two months), which were equal to the Training X of yearly model (one bin with a length of a year).

2.4.2. Bagged trees

Bagged regression trees, which are quick to train on large data sets [25,26], were chosen to build the models described in Table 3. Every tree in the ensemble is grown on an independently drawn bootstrap replica of the data. Observations not included in this replica are 'out-of-bag' samples for this tree. To compute a prediction from an ensemble of trees for unseen data, the average of predictions across all individual trees in the ensemble is taken. The out-of-bag observations are used to estimate the prediction error, which helps to avoid severe over-fitting during the training phase. Here a function named 'treebagger' in the statistics toolbox of MATLAB R2013b (MathWorks, Natick, MA, USA) was used to implement the algorithm [27].

2.4.3. Comparison of predictive models

The four model schemes listed in Table 3 predict the same target which is the number of days in hospital in year 2013. Afterward, we compared these four models to investigate the relationship between temporal resolution and their predictive performances.

Firstly, parameters of bagged trees were tuned for each model to optimize performance. The number of trees used in the bagging ensemble was set to 50 since the out-of-bag error did not improve when increasing the number of trees above 50. For each model, the minimum number of observations (customers) per tree leaf was searched among a set of values of 5, 10, 20, 50, 100, 150, and

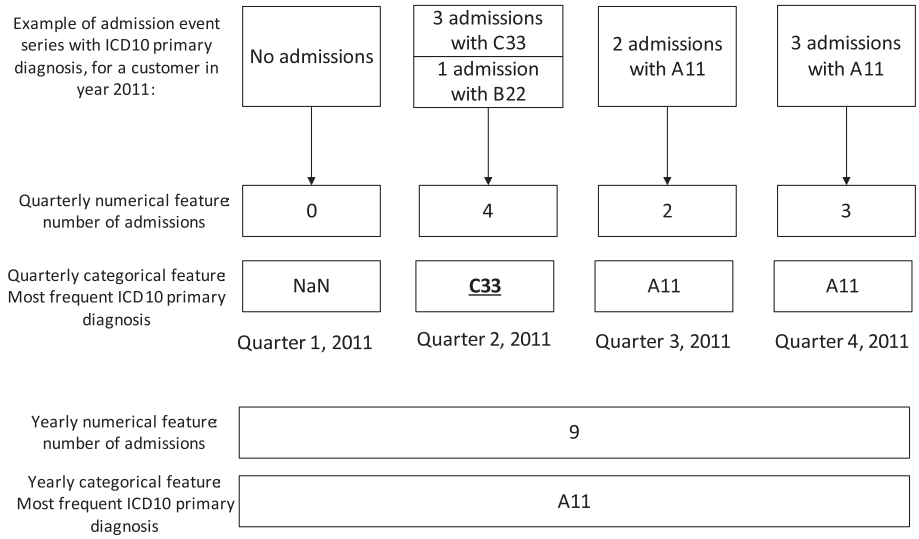
Example of admission event series with ICD10 primary diagnosis, for a customer in year 2011:
No admissions
3 admissions with C33
1 admission with B22
2 admissions with A11
3 admissions with A11
Quarterly numerical feature: number of admissions
0
4
2
3
Quarterly categorical feature Most frequent ICD10 primary diagnosis
NaN
C33
A11
A11
Quarter 1, 2011
Quarter 2, 2011
Quarter 3, 2011
Quarter 4, 2011
Yearly numerical feature: number of admissions
9
Yearly categorical feature Most frequent ICD10 primary diagnosis
A11

Fig. 2. Process for converting a series of events with categorical variables and numeric variable into a time series, with an example of a categorical feature (most frequent ICD-10 primary diagnosis) and a numeric feature (number of admissions). The bins used in this example are quarter and year. A11, B22, C33 represent three ICD-10 primary diagnosis codes.

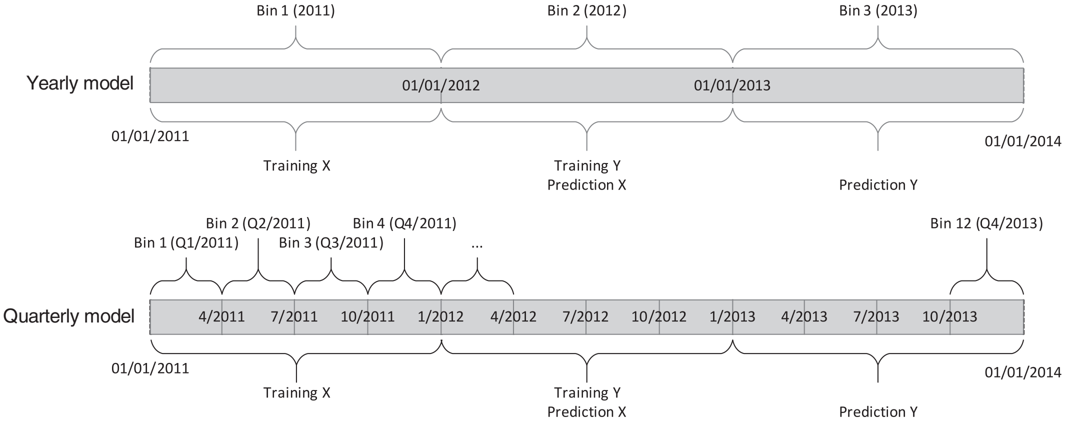

Fig. 3. An illustration of temporal models. Yearly model and quarterly model are shown as examples. For both models, the periods from the initial two years (01/01/2011 to 31/12/2012) serve as a two-year training period with a one-year prediction period from 1/1/2013 to 31/12/2013. Training X is the feature matrix for training (from year 2011); Training Y is target for training (from 2012). The feature matrix from 2012 is input to the trained models to make a prediction for 2013. The prediction outcome (Prediction Y) is a forecast of days in hospital for the year 2013, and is later compared to the actual days in hospital in 2013. The ticks with dates are the borders of each bin. In the sub-plot of the quarterly model, these dates (first day in a month) were displayed as month/year to save space.

Table 3
A list of temporal model schemes. Numbers are the indices to the bins of Training X, Training Y, Prediction X, Prediction Y described in Fig. 3.

Model no.	Bin[a]	Training X	Training Y	Prediction X	Prediction Y
1	Y	1	2	2	3
2	H	1–2	3–4	3–4	5–6
3	Q	1–4	5–8	5–8	9–12
4	B	1–6	7–12	7–12	13–18

[a] Y: yearly; H: half year; Q: quarter; B: two months (bi-monthly model).

200. We chose 100, since all models gave the best performance with this value. Other parameters, if not specified, were default MATLAB settings.

The data set of 100,000 customers were split into 10 non-overlapped subsets, each containing 10,000 customers. Using the selected parameter settings, we trained and tested 10 models on these 10 subsets. A comparison test, which will be described in Section 2.5.2, was later applied to compare the performances of the four models. Finally, the best model from this comparison was applied to three populations, the whole 100,000 customers (Group 1), customers with a birth year in or after 1948 (Group 2) and customers born before the year 1948 (Group 3), respectively. These sub-groupings were chosen since the average number of days in hospital increases significantly between the ages of 63 and 65 years [5]. Customers with a birth year of 1948 would be 63 to 65 years old across the study period (2011–2013). Therefore, a birth year of 1948 was used as a threshold to categorize non-senior (88,587 customers) and senior sub-populations (11,413 customers). We also applied the yearly model to the three populations for comparison, as a yearly model was used in our previous work [5].

Fig. 4 is an illustrative figure of the whole process, including data split, model training and model comparison.

2.5. Performance measures

2.5.1. Statistics for model performance

Table 4 gives an overview of the performance indicators used in this study. One of the regression performance indicators is referred to as the root-mean-square-error (RMSE), which is the root-mean-square of the difference between the logarithm of the estimated DIH and the logarithm of the true number of days [4]. The logarithm ($\ln(\text{DIH} + 1)$) was used to reduce the importance assigned to those with a high number of hospital days. In addition to RMSE, the Pearson correlation coefficient (ρ) was also calculated between the logarithm of the predicted DIH and the logarithm of the actual number of DIH. RMSE and ρ served as an overall measure of the goodness of predictability from a regression perspective.

Additionally, customers can be categorized into two categories, those without hospital days (0 hospital days) and those with hospital days (at least 1 hospital day) per year. By setting a threshold for the predicted number of hospital days, binary classification analysis was applicable to the dichotomized result. The Matthews correlation coefficient (MCC), which is regarded as a balanced measure and can be used even if the classes are unevenly represented, was calculated. An MCC of $+1$ represents a perfect prediction, 0 means no better than chance, and -1 indicates total disagreement between prediction and observation [28,29]. In addition, the area under the receiver operating characteristic curve (AUC) was also computed.

The results reported were obtained with a threshold applied to the continuous output estimate of DIH from the bagged tree regression model. AUC, ρ, and RMSE are not affected by choice of the threshold, and only the MCC changes when the threshold varies. Therefore, the threshold chosen to dichotomize the continuous output estimate from the bagged tree regression model was that which maximized the MCC. Predicted values below the threshold were considered to represent a prediction of zero DIH.

2.5.2. Statistics for comparison of models

The comparison of the performances of the four different temporal models was based on considering the ten runs each as independent samples of the key statistic parameters, AUC, MCC, ρ, and RMSE. Non-parametric descriptive statistics (median, interquartile range, minimum and maximum) are provided and tested for statistical significance utilizing the Kruskal–Wallis test, which is the non-parametric equivalent to the one-way analysis of variance (ANOVA) [30,31]. It is used for comparing two or more samples that are independent, that may have different sample sizes, and does not assume a normal distribution of the residuals. The null hypothesis is that the medians of all groups are equal, and the alternative hypothesis is that at least one population median of one group is different than the population median of at least one other group.

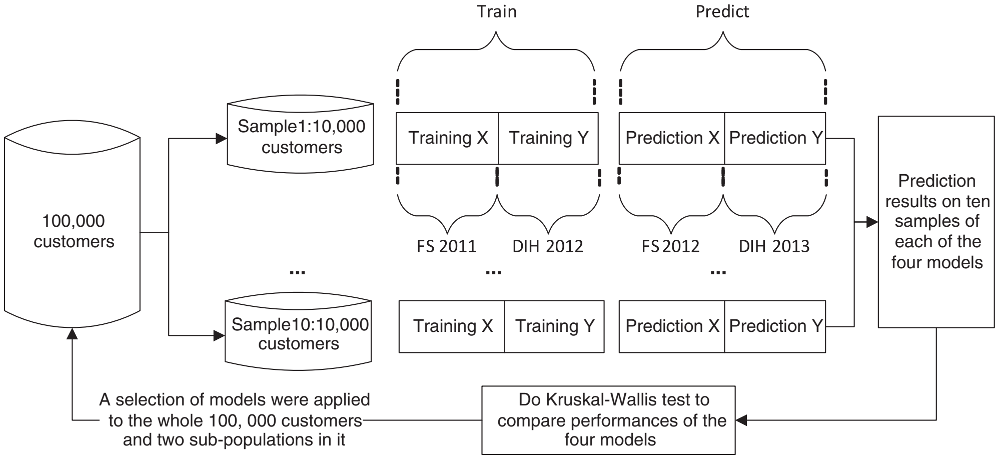
Train
Predict
100,000 customers
Sample1:10,000 customers
Sample10:10,000 customers
Training X
Training Y
Prediction X
Prediction Y
FS 2011
DIH 2012
FS 2012
DIH 2013
...
Training X
Training Y
Prediction X
Prediction Y
Prediction results on ten samples of each of the four models
Do Kruskal-Wallis test to compare performances of the four models
A selection of models were applied to the whole 100, 000 customers and two sub-populations in it

Fig. 4. An illustrative figure of the whole process: data split, model training and model comparison. The data set of 100,000 customers were split into 10 non-overlapped subsets, each containing 10,000 customers. Kruskal–Wallis test was applied to compare the performances of the four models, based on the prediction results of the 10 samples. Finally, the best model from this comparison was applied to three populations of the whole 100,000 customers. FS and DIH are abbreviations of ‘feature set’ and ‘days in hospital’.

Table 4
Performance measures: root-mean-square-error (RMSE), Pearson correlation coefficient (ρ), Matthews correlation coefficient (MCC) and the area under receiver operating characteristic curve (AUC).

Measure	Equation
RMSE	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}[\ln(p_i+1)-\ln(a_i+1)]^2}$
ρ	$\frac{\sum_{i=1}^{N}[\ln(p_i+1)-\overline{\ln(p+1)}][\ln(a_i+1)-\overline{\ln(a+1)}]}{\sqrt{\sum_{i=1}^{n}[\ln(p_i+1)-\overline{\ln(p+1)}]^2}\sqrt{\sum_{i=1}^{n}[\ln(p_i+1)-\overline{\ln(a+1)}]^2}}$ (Pearson)
MCC	$\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
AUC	$\int_{-\infty}^{+\infty} R_{tp}(T)R'_{fp}(T)dT$

N is the total number of persons in the population; p_i is the predicted number of DIH for the ith person; a_i is actual DIH, $i \in [1, N]$; $\overline{\ln(p+1)}$ is the mean value of all the logarithmic predicted number of DIH; $\overline{\ln(a+1)}$ is the mean value of all the logarithmic actual DIH; TP is the number of hospitalized patients who were correctly predicted as having $\geqslant 1$ DIH; TN is the number of subjects who were correctly predicted as having 0 DIH; FP is the number of subjects who were predicted as having $\geqslant 1$ DIH, but actually having 0 DIH; FN is the number of subjects who were predicted as having 0 DIH, but actually having $\geqslant 1$ DIH; $R_{tp}(T)$ and $R_{fp}(T)$ are the true positive rate (sensitivity) and the false positive rate (equals to 1−specificity) for a given threshold T in a binary classification model.

3. Results

3.1. Results of comparison of models

Fig. 5 shows box-plots of the results for the four key performance measures of the four models, from yearly model to the bi-monthly model. Displayed values are median (MED), minimum (MIN), maximum (MAX), inter-quartile range (IQR), the number of elements in each sample (N) and the p-value (p) as obtained by the Kruskal–Wallis test. The diagrams show that the hypothesis of four equal models is rejected, with respect to AUC and MCC.

Rejecting ‘all-equal’ hypothesis means that at least one model is significantly different with the others. From Fig. 5, obvious difference between yearly model and the other three non-yearly models could be observed. Further test was done to test the equality of the three non-yearly models (i.e., half-yearly model, quarterly model and bi-monthly model). Table 5 shows p-values of the Kruskal–Wallis test results for the four key performance measures of three non-yearly models, indicating no significant differences among these three models.

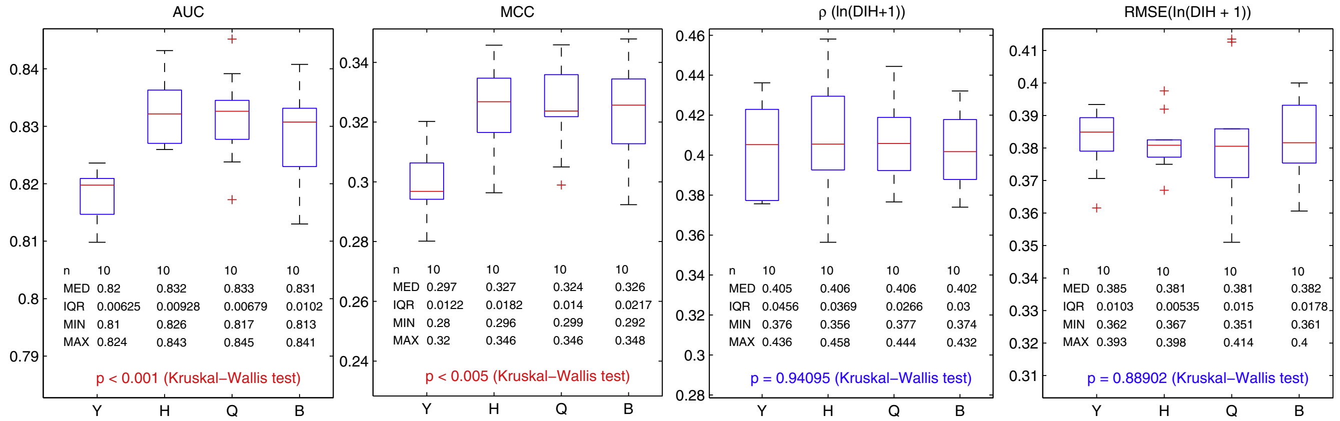

Fig. 5. Box-plot of the results for the four key performance measures of all four models. The central mark is the median, the edges of the box are the 25th (q_1) and 75th percentiles (q_3). Points are drawn as outliers in the symbol of ‘+’, if they are larger than $q_3 + 1.5(q_3 - q_1)$ or smaller than $q_1 - 1.5(q_3 - q_1)$.

Table 5
Kruskal–Wallis test results (p-value) for the four key performance measures of three non-yearly models.

Measure	p-value
AUC	0.67553
MCC	0.97327
ρ	0.83581
RMSE	0.94603

Table 6
Performance metrics of yearly model (Y) and half-yearly model (H) on three sub-populations.

Result	Y[a]	H[a]	Y[b]	H[b]	Y[c]	H[c]
RMSE	0.376	0.373	0.315	0.314	0.688	0.681
ρ	0.437	0.450	0.375	0.380	0.448	0.465
MCC	0.301	0.328	0.241	0.269	0.375	0.426
AUC	0.824	0.837	0.802	0.816	0.804	0.823

[a] Group 1: all 100,000 customers.
[b] Group 2: 88,587 customers in or after year 1948.
[c] Group 3: 11,413 customers born before year 1948.

3.2. Model performance

Table 6 lists the performance when yearly and half-yearly models were applied to the three groups of customers described in Section 2.4.3. Fig. 5 shows that the yearly model performed worst. Considering that no significant differences were found among the other three models, the half-yearly model was used to represent the other three models. It was also applied to the three groups, allowing a comparison with the yearly-model.

4. Discussion and conclusion

A temporal data mining approach for predicting future days in hospital has been proposed. Equally spaced time series features were constructed on time intervals (bins) intended to capture subtle temporal information and improve prediction outcomes. Four models with bins of different lengths, i.e., year, half-year, quarter, and two months, have been built to explore how varying the temporal resolution would impact the predictive power. These bin features become a summary of events occurring within each bin. The smaller the bin, the more subtle time information could be brought into the feature set, however, the frequency of hospitalization and health service utilization will likely put lower bounds on gains achievable by decreasing the bin length.

Fig. 5 shows the four performance measures of the four models, indicating statistically significant differences in the performances of the four models with respect to AUC and MCC, the two measures for binary analysis (classification of patients into those with and without hospitalizations in the prediction year). It demonstrates that at least one population median of one model is different than the population median of at least one other model. The yearly model ranked the lowest for both AUC (median value = 0.82) and MCC (median value = 0.297), while the other three models gave better results for median AUC and median MCC (e.g., for half-yearly model, median AUC = 0.832 and median MCC = 0.327). A visible difference between the yearly model and the other three models can be seen in the AUC and MCC sub-plots in Fig. 5.

Regarding the two measures of goodness of fit for the regression, ρ and RMSE, there was no evidence of a statistically significant advantage of a particular type of model; although yearly models showed the highest (worst) median value of RMSE (0.385). In general, it is notable that a trend seems to exist in median values of AUC, MCC and RMSE. Initially, they increased as temporal resolution becomes better (smaller bins) from yearly to half-yearly, then reached a plateau and did not change observably from the half-yearly model to the bi-monthly model. One possible reason could be that as bin lengths become smaller, more detailed temporal information on events are captured, thereby improving the predictive power. However, given the temporal sparsity of health-related events, decreasing the bin length too far will result in many empty bins, whilst also increasing the dimension of the feature space, which could negatively impact the classifier's performance and offset the benefits introduced by increasing the temporal resolution.

The Kruskal–Wallis test results for the half-yearly model, quarterly model and bi-monthly model(omitting the yearly model) is also shown in Table 5. It reveals that the difference among these three models are not statistically significant. This leads to the conclusion that the significant difference shown in Fig. 5 comes from the yearly model.

Table 6 shows the performance of yearly models and half-yearly models on three populations, namely the complete sample of 100,000 customers (Group 1), customers with a birth year of or after 1948 (Group 2), and customers with a birth year before 1948 (Group 3). The numbers displayed in Table 6 confirm for all three groups of customers that the half-yearly model outperformed the yearly model with respect to all performance measures; MCC, AUC and ρ increased. The improvement in RMSE was not as obvious as compared to the other performance indicators. Especially, in Group 2 (customers born in or after year 1948), the performance of the half-yearly model (RMSE = 0.314) was almost equal to the performance of the yearly model (RMSE = 0.315).

Another phenomenon worth discussing is that the superior performance of the half-yearly model seems to be especially pronounced in Group 3 (customers born before year 1948), when compared to the other two groups. The half-yearly model raised the MCC to 0.426 (from 0.375 for the yearly model). A potential explanation of this rather remarkable improvement could be that senior customers are heavy users of medical resources, and are hospitalized more frequently. Therefore, the medical events data in Group 3 is denser than for the whole population including the non-senior population, thereby boosting the predictive power of the time series features.

Generally speaking, the proper temporal interval is expected to vary between data sets with different characteristics. If the data density in a data set is too sparse, prediction accuracy is expected to decrease when the information is sorted into smaller bins and the dimension of the feature space increases.

Predicting length of stay is a very important area of clinical and informatics research. Some excellent research has been done, such as predicting LoS of stroke patients in rehabilitation [32], patients with acute psychiatric disorders [33], and chemically-dependent individuals [34]. Most of these studies focused on specific diseases in relatively small cohorts. To our best knowledge, there are few comparable studies to what is presented here, developing models that predict the number of hospitalization days for a general population using insurance claims. The model is intended to capture the uncertainty and variability among different disease groups, different age groups and other information. The general model developed here has considerable value as a benchmark against which disease-specific models may be compared [5]. Some related work has been conducted using insurance claims to predict health-care costs in the future [35], or using drug claims to predict pharmacy costs [36]. However, in these studies, the potential predictive power embedded within the timing of claims was not investigated.

Betal et al. proposed a temporal pattern mining approach to classify future patients based on their medication data and laboratory results [15]. One of the key step in their work was defining a language (i.e., TAs [37] and temporal logic [17]) that adequately represent the temporal dimension of the data. An advantage of such methods is allowing a description of interactions among multiple time series so as to reveal the interaction of medication data and laboratory results (e.g., a relative temporal order among particular medication and glucose/cholesterol measurements). Temporal pattern mining approaches have been applied to clinical temporal data and health administrative data. A number of studies have investigated mining temporal patterns for diabetic patients' clinical variables such as blood glucose, cholesterol, and medications [18]. Recent studies have also applied such methods to detect adverse drug reactions from drug prescriptions and clinical data stored in administrative healthcare databases [38,39]. We have not found any applications of these methods to insurance claims. In contrast to the nature of many clinical data sets described in the literature [18,38,39], insurance claim data are temporally sparse and lack a rich temporal interaction across time series. Respecting the sparse nature of our insurance claims data, using the temporal binning approach described in this paper is considered appropriate and also has the advantage of being a readily understandable method of extracting temporal information.

It should be noted that billing codes can be assigned retroactively; for instance, coding errors can happen, which may be retrospectively amended. Physicians may often code a diagnosis in the absence of confirming tests results during emergency treatments, because a diagnosis is required for claim reimbursement. These billing codes could be altered and the claim reimbursement corrected later. Moreover, late claims could be received by insurer after the patient has been discharged from the hospital (e.g., claiming more than a year after discharge). The data set used in this paper (claims between year 2011 and 2013) was provided by HCF in August 2014. Issues concerning with retroactive assignment are present in this data set. HCF estimate about 0.33% of claims will have their billing codes added or amended retroactively in the following year. Also, given that this is a real characteristic of an active insurance business, it is appropriate that such claims are included in the modeling task and considered as part of the noisy big data challenge that proposed models need to work with.

Moreover, since the cohort is fixed at 100,000 customers for the three years, the demographics expectedly change over time; they grow older. This trend is very apparent in Tables 1 and 2. With this aging, the average number of admissions per customer increases from 0.155 in 2011 to 0.182 in 2013, however, the average number of days per admission is approximately fixed at 2.4 for each year (it actually decreases slightly from 2.413 in 2011 to 2.407 in 2013). Expectedly, with more admissions, both the average charge per customer (in a fixed cohort) and total amount charged show an upward trend. The prediction model makes use of these trends during training. The reader should therefore note the following caveat. If this prediction model was then applied in a cohort that was not aging on average over time, its prediction performance would be worse, as it would overestimate the number of days in hospital. This behavior could be dealt with in two ways: (1) when training the model using two years of data from a fixed cohort, it must be then applied to the same fixed cohort when making a prediction for a subsequent year, or; (2) the cohort may be allowed to change (some customers die or leave the fund, new customers are registered) across the two years used to train the model, such that the demographics and claims statistics remain stationary between these two years, hence allowing prediction in a subsequent year for a further altered cohort which also has the same demographic and claims statistics.

As mentioned in Section 2.1, the data set used in this work has no explicit information on customer deaths. From the perspective of modeling, leaving those who died in the data set would likely reduce the model performance, as the model may be predicting hospitalizations for them later that year when they are already deceased and their actual DIHs are assumed to be zero. Therefore, our performance indicators could be considered as estimates of a lower bound on performance which arises when information on death is not available. In the future, if possible, it may be worthwhile to integrate information on registered deaths (e.g., whether/when customers died) into the model, which would likely improve performance.

The time intervals provide an opportunity of building more flexible models by ensembling sub-models; for example, a model could be built based on each single quarter of a given year and forecast days in hospital for each quarter of the subsequent year and then sum the results to estimate the total days in a year. By doing this, we may, for example, have a chance to explore seasonal patterns inside the data, since different models may be needed for winter than for summer. In addition, it would be interesting to see how such models work on data with longer observation periods (i.e., using data from additional years). If pathology/laboratory test data could be acquired in the future, it would also be interesting to build models for specific disease groups (e.g., chronic diseases) which have more frequent medical events.

Conflict of interest

None declared.

References

[1] How much do we spend on health? Australian Institute of Health and Welfare, Australian Government, 2014. <http://www.aihw.gov.au/australias-health/2012/spending-on-health/>.

[2] D.E. Niewoehner, Y. Lokhnygina, K. Rice, W.G. Kuschner, A. Sharafkhaneh, G.A. Sarosi, P. Krumpe, K. Pieper, S. Kesten, Risk indexes for exacerbations and hospitalizations due to copd, Chest 131 (1) (2007) 20–28.

[3] T. Sugimoto, T. Tanigawa, K. Onishi, N. Fujimoto, A. Matsuda, S. Nakamori, K. Matsuoka, T. Nakamura, T. Koji, M. Ito, Serum intact parathyroid hormone levels predict hospitalisation for heart failure, Heart 95 (5) (2009) 395–398.

[4] P. Brierley, D. Vogel, R. Axelrod, Heritage provider network health prize round 1 milestone prize: how we did it – Team 'Market Makers', 2014. <http://www.heritagehealthprize.com/c/hhp/leaderboard/milestone1>.

[5] Y. Xie, G. Schreier, D.C.W. Chang, S. Neubauer, Y. Liu, S.J. Redmond, N.H. Lovell, Predicting days in hospital using health insurance claims, IEEE J. Biomed. Health Inform. 19 (4) (2015) 1224–1233.

[6] T.C. Fu, A review on time series data mining, Eng. Appl. Artif. Intell. 24 (1) (2011) 164–181.

[7] G. Bruno, P. Garza, Temporal pattern mining for medical applications, in: Data Mining: Foundations and Intelligent Paradigms, Springer, 2012, pp. 9–18.

[8] U. Thissen, R. Van Brakel, A.P. De Weijer, W.J. Melssen, L.M.C. Buydens, Using support vector machines for time series prediction, Chemometr. Intell. Lab. Syst. 69 (1) (2003) 35–49.

[9] J.J. Rodríguez, C.J. Alonso, J.A. Maestro, Support vector machines of interval-based features for time series classification, Knowl.-Based Syst. 18 (4) (2005) 171–178.

[10] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, E. Keogh, Querying and mining of time series data: experimental comparison of representations and distance measures, Proc. VLDB Endowment 1 (2) (2008) 1542–1552.

[11] P.F. Marteau, Time warp edit distance with stiffness adjustment for time series matching, IEEE Trans. Pattern Anal. Machine Intell. 31 (2) (2009) 306–318.

[12] J.P. Boucher, M. Denuit, M. Guillen, Models of insurance claim counts with time dependence based on generalisation of poisson and negative binomial distributions, Variance 2 (1) (2008) 135–162.

[13] J.P. Boucher, M. Guillén, A survey on models for panel count data with applications to insurance, Revista Real Acad. Ciencias Exactas, Fisicas Nat. Ser. A. Mat. 103 (2) (2009) 277–294.

[14] J.B. Illian, S.H. Sørbye, H. Rue, D.K. Hendrichsen, Fitting a log gaussian cox process with temporally varying effects – a case study, J. Environ. Stat. 3 (2012) 1–25.

[15] I. Batal, H. Valizadegan, G.F. Cooper, M. Hauskrecht, A temporal pattern mining approach for classifying electronic health record data, ACM Trans. Intell. Syst. Technol. 4 (4) (2013) 63:1–63:22. http://doi.acm.org/10.1145/2508037.2508044.

[16] Y. Shahar, A framework for knowledge-based temporal abstraction, Artif. Intell. 90 (12) (1997) 79–133. <http://www.sciencedirect.com/science/article/pii/S0004370296000252>.

[17] J.F. Allen, Towards a general theory of action and time, Artif. Intell. 23 (2) (1984) 123–154. <http://www.sciencedirect.com/science/article/pii/0004370284900080>.

[18] R. Moskovitch, Y. Shahar, Medical temporal-knowledge discovery via temporal abstraction, in: American Medical Informatics Association (AMIA) Annual Symposium Proceedings, vol. 2009, 2009, pp. 452–456.

[19] R. Henriques, S.M. Pina, C. Antunes, Temporal mining of integrated healthcare data: methods, revealings and implications, SDM IW on Data Mining for Medicine and Healthcare (2013) 52–60.

[20] W. Lin, M.A. Orgun, G.J. Williams, Mining temporal patterns from health care data, in: Data Warehousing and Knowledge Discovery, Springer, 2002, pp. 222–231.

[21] J. Wiens, E. Horvitz, J.V. Guttag, Patient risk stratification for hospital-associated c. diff as a time-series classification task, in: Advances in Neural Information Processing Systems, 2012, pp. 476–484.

[22] R.A. Baxter, G.J. Williams, H. He, Feature selection for temporal health records, in: Advances in Knowledge Discovery and Data Mining, Springer, 2001, pp. 198–209.

[23] V. Sundararajan, T. Henderson, C. Perry, A. Muggivan, H. Quan, W.A. Ghali, New ICD-10 version of the Charlson comorbidity index predicted in-hospital mortality, J. Clin. Epidemiol. 57 (12) (2004) 1288–1294.

[24] H. Quan, B. Li, C.M. Couris, K. Fushimi, P. Graham, P. Hider, J.-M. Januel, V. Sundararajan, Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries, Am. J. Epidemiol. 173 (6) (2011) 676–682.

[25] L. Breiman, Random forests, Machine Learn. 45 (1) (2001) 5–32.

[26] L. Breiman, A. Cutler, Random forests, 2014. <http://www.stat.berkeley.edu/breiman/RandomForests/cc_home.htm>.

[27] Treebagger, The MathWorks, Inc., 2014. <http://www.mathworks.com.au/help/stats/treebagger.html/>.

[28] B.W. Matthews, Comparison of the predicted and observed secondary structure of t4 phage lysozyme, Biochim. Biophys. Acta (BBA) – Protein Struct. 405 (2) (1975) 442–451.

[29] D.M. Powers, Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation, J. Machine Learn. Technol. 2 (1) (2011) 37–63.

[30] W.H. Kruskal, W.A. Wallis, Use of ranks in one-criterion variance analysis, J. Am. Stat. Assoc. 47 (260) (1952) 583–621.

[31] G.W. Corder, D.I. Foreman, Nonparametric Statistics for Non-statisticians: A Step-by-step Approach, John Wiley & Sons, 2009.

[32] T. Galski, R.L. Bruno, R. Zorowitz, J. Walker, Predicting length of stay, functional outcome, and aftercare in the rehabilitation of stroke patients. The dominant role of higher-order cognition, Stroke 24 (12) (1993) 1794–1800.

[33] D.A. Huntley, D.W. Cho, J. Christman, J.G. Csernansky, Predicting length of stay in an acute psychiatric hospital, Psychiatric Services 49 (8) (1998) 1049–1053.

[34] M.T. Williams, C.S. Roberts, Predicting length of stay in long-term treatment for chemically dependent females, Int. J. Addictions 26 (5) (1991) 605–613.

[35] D. Bertsimas, M.V. Bjarnadottir, M.A. Kane, J.C. Kryder, R. Pandey, S. Vempala, G. Wang, Algorithmic prediction of health-care costs, Oper. Res. 56 (2008) 1382–1392.

[36] Y. Zhao, A.S. Ash, R.P. Ellis, J.Z. Ayanian, G.C. Pope, B. Bowen, L. Weyuker, Predicting pharmacy costs and other medical costs using diagnoses and drug claims, Med. Care 43 (2005) 34–43.

[37] Y. Shahar, A framework for knowledge-based temporal abstraction, Artif. Intell. 90 (1) (1997) 79–133.

[38] H. Jin, J. Chen, H. He, G.J. Williams, C. Kelman, C.M. O'Keefe, Mining unexpected temporal associations: applications in detecting adverse drug reactions, IEEE Trans. Inform. Technol. Biomed. 12 (4) (2008) 488–500.

[39] E. Chazard, C. Preda, B. Merlin, G. Ficheur, R. Beuscart, Data-mining-based detection of adverse drug events, in: Studies in Health Technology and Informatics, vol. 150, 2009, pp. 552–556.