Impact of Hierarchies of Clinical Codes on Predicting Future Days in Hospital

Yang Xie, Student Member, IEEE, Sandra Neubauer, Günter Schreier, Senior Member, IEEE,
Stephen J. Redmond, Senior Member, IEEE, Nigel H. Lovell, Fellow, IEEE

Abstract—Health insurance claims contain valuable information for predicting the future health of a population. Nowadays, with many mature machine learning algorithms, models can be implemented to predict future medical costs and hospitalizations. However, it is well-known that the way in which the data are represented significantly affects the performance of machine learning algorithms. In health insurance claims, key clinical information mainly comes from the associated clinical codes, such as diagnosis codes and procedure codes, which are hierarchically structured. In this study, it is investigated whether the hierarchies of such clinical codes can be utilized to improve predictive performance in the context of predicting future days in hospital. Empirical investigations were done on data sets of different sizes, considering that the frequency of the appearance of lower-level (more specific) clinical codes could vary significantly in populations of different sizes. The use of bagged trees with feature sets that include only basic demographic features, low-level, medium-level, high-level clinical codes, and a full feature set were compared. The main finding from this study is that different hierarchies of clinical codes do not have a significant impact on the predictive power. Some other findings include: 1) Sample size greatly affects the predictive outcome (more observations result in more stable and more accurate outcomes); 2) Combined use of enriched demographic features and clinical features give better performance as compared to using them separately.

I. INTRODUCTION

With rapid advances in the area of big data analytics, healthcare providers are seeking to refine their care delivery models, and health insurers are aiming to reduce their financial costs, by analyzing large-scale healthcare datasets. One valuable dataset is that of medical claim data, which has broad coverage of the general population. It has been widely used by health insurance payers to build medical cost predictive models [1], [2]. Nowadays, the value of medical claims has been broadened to cover multiple areas. Bjarnehtöttir et al. built a mathematical framework for a real-time drug surveillance system to discover side effects using claim data [1]. Studies have investigated the power of medical claims in predicting future days in hospital (DIH) [3], [4], which would help reducing unnecessary hospitalizations and providing preventive care.

It is well-known that the predictive performance of a model is greatly affected by the way in which the data are represented (i.e., as features extracted from raw data). Medical claims contain various types of data, ranging from demographics to clinical data (e.g., diagnoses and procedures). The latter are usually translated into codes by means of consistent coding systems. For instance, diagnoses are most commonly encoded using the International Classification of Diseases and Related Health Problems (ICD), which is published and maintained by the World Health Organization (WHO). It structures codes hierarchically according to body systems and etiology [5].

However, the vast number of diagnoses and procedures is one of the many potential challenges for preparing claim data for modelling. For example, there are typically thousands of distinct diagnosis codes in the claim data of a large population. Building predictive models for such a population using all of these codes could result in very high dimensionality and sparsity of data within this high-dimensional feature space. Therefore, a hierarchical organisation of such codes is needed. The most straightforward way of doing this is to utilize the existing well-structured hierarchies of the ICD codes. But how these hierarchies would affect predictive performance has not been comprehensively investigated.

We found only a single study that directly addressed this topic and in which Zhao et al. investigated the effect of concept hierarchies of clinical codes in detecting adverse drug events [6]. Their main finding was that predictive performance can be kept at a high level even without employing the more specific levels in the concept hierarchies. However, one limitation of this study was the relatively small size of the datasets used, ranging from 48 to 3,586 patients, thereby having many more features than instances. As expected, many diagnosis codes do not appear in such small populations. Hence, low-level clinical codes may remain under-exploited in predictive modeling when using small sample sizes. Whether use of large datasets would lead to a different conclusion, and whether this conclusion holds in a different context (e.g., predicting hospitalizations), provides motivation for our study. A hypothesis is that for small sample sizes, high-level (more general) codes may be sufficient to achieve reasonable performance, low-level (more specific) codes may further improve the performance when a larger sample size is available.

In our previous work [3], we built a model 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 claim data. The present investigation of the predictive power of different hierarchical clinical codes was done with the same goal of predicting future days in hospital (DIH).

II. METHODS

A. Data Pool

The data pool consisted of the hospital claim data for 100,000 individuals from a private health insurer called Hospitals Contribution Fund of Australia (HCF), one of Australia’s largest combined registered private health fund and life insurance organizations. Three-year claim data (from year 2010 to 2012), including demographics, were available.

B. Clinical Codes and Hierarchies

The present study investigated the hierarchies of two elements in the claims, which contained medical information and involved coding schemes.

Primary diagnosis code: An ICD-10-AM (International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Australian Modification) code is assigned to each acute inpatient admission. It represents the diagnosis or condition which necessitated the use of health care services. ICD-10-AM is a modified version of the WHO’s ICD-10 base classification system. It is structured mono-hierarchically within each chapter and uses codes ranging from three to five characters for categorization into five different levels of detail. Since the WHO’s core classification at the three-character code level remains the same within ICD-10-AM, the generally valid grouping scheme was used for extracting features from levels 1, 2 and 3 [5], [7].

Diagnosis-related group: The Australian Refined Diagnosis-Related Group (AR-DRG) is an Australian hospital inpatient classification system. Each discharge is assigned a four character code. The first character can either be a letter, designating one of 23 Major Diagnostic Categories (MDC) or a 9, indicating an error DRG. The MDCs are subdivided into basic-DRGs by partitioning the codes into surgical, medical, or other, which is denoted by the following two numbers. The last digit provides information about the patients’ resource consumption [8].

Note that both ICD-10-AM and AR-DRG codes can be further expanded to more specific levels. However, considering the scale of our dataset, only the first three levels were investigated. An illustrative example of the first three levels of ICD-10-AM and AR-DRG codes are shown in Fig. 1. ICD-10-AM is abbreviated as ICD10 and AR-DRG is abbreviated as DRG in the remainder of the article for convenience.

C. Experimental Design

The whole investigation was done in the context of predicting future DIH. The predictive goal was to use claims from the observation period, the initial two years (01/01/2010 to 31/12/2011), to predict DIH in the third year (year 2012) for each customer. Since the effect of hierarchies of clinical codes was expected to be affected by sample size, in our study, the impact of hierarchies of clinical codes on predictive performance was investigated on data sets of different sample sizes.

1) Data set preparation: Our data pool included a sample of 100,000 randomly selected customers. Sizes of experimental data sets ranged from 1,000, 2,000, 4,000, 8,000 to 16,000. Under sample size, 10 data replicas (of the same size), randomly sampled from the pool of 100,000, were prepared.

2) Predictive model: Bagged regression trees, with a reputation as a classifier of choice for high-dimensional and sparse data, and being quick to train on large data sets [9], were chosen to build the predictive models in this study. The algorithm constructs an ensemble of decision trees. Each tree is independently built on a bootstrap replica of the data. The average of predictions across all individual trees in the ensemble is assigned to unseen data.

3) Performance measure: The performance metrics were: the area under the receiver operating characteristic curve (AUC), and the Pearson correlation coefficient (ρ). AUC depicted the performance from the perspective of classification (classifying customers into two categories of ‘no hospitalization’ and ‘at least one day in hospital’), and ρ served as a measure of the goodness of predictability from the perspective of regression to estimate exactly how many DIH for the coming year.

When two competing models were compared (pairwise comparison), the Wilcoxon signed-rank test was used for statistical hypothesis testing. When multiple models were involved in comparison, the Kruskal-Wallis test was instead employed, where the null hypothesis was that all models performed equally well.

4) Experiment setup: To study the impact of hierarchies in the coding system, a level-wise comparison was conducted. Additionally, a baseline feature scheme, and a full feature scheme were used, against which the level-wise feature schemes could be compared. All feature schemes used are listed below:

1) baseline: Basic feature related to age, gender, days in hospital (DIH) and hospitalization cost of previous year were included. Two additional features were, average days to previous admission (DAYS2PREV), and Charlson index of previous year.

2) drg1: In addition to baseline, all level 1 DRG features were included.

3) drg2: In addition to baseline, all level 2 DRG features were included.

4) drg3 In addition to baseline, all level 3 DRG features
were included.
5) icd1: In addition to baseline, all level 1 ICD10 features were included.
6) icd2: In addition to baseline, all level 2 ICD10 features were included.
7) icd3: In addition to baseline, all level 3 ICD10 features were included.
8) full: Full feature scheme, includes enriched demographic features and clinical features. Comparing to the basic demographic features, additional demographic information such as customers’ HCF membership type and health insurance products purchased, were also included. In addition to DRG and ICD10 features, information of procedure items used during hospitalization were also used to enrich the feature set. More details of this full feature set can be found in researchers’ previous work [3].

Predictive models, with the above feature schemes, were built and evaluated on the 10 data replicas of every sample size, as described in Section II-C.1. For every sample size, the 10 data replicas were used to ensure that the variability of populations of this size could be captured. Therefore, outcomes obtained from the 10 replicas allowed the variations of predictive performance to be assessed under each combination of sample size and feature scheme.

III. RESULTS

A. Impact of Sample Size on Predictive Outcome

Fig. 2 shows box-plots with descriptive statistics of AUC and ρ for increasing sample sizes. Also displayed is p-values derived from Kruskal-Wallis test. As can be seen from the red line in each box, the median of the predictive performance increases when using larger sample sizes, for both classification and regression. Purely including 16,000 observations in the learning data instead of 1,000, for instance, leads to a performance improvement in mean AUC from 0.69 to 0.74. Furthermore, results become more stable within the area of larger sample sizes, which is indicated by smaller boxes (smaller inter- replica variance). The red colored p-value represents results are significantly different to the median result for all experiments.

B. Impact of Feature Scheme on Predictive Outcome

Fig. 3 demonstrates box-plots with descriptive statistics of AUC and ρ for changing feature schemes. p-values obtained from a Kruskal-Wallis test are also shown. According to the result, ‘all-equal’ hypothesis was rejected, which means at least one feature scheme is significantly different from the others.

C. Impact of Combination of Feature Scheme and Sample Size on Predictive Outcome

Fig. 4 displays a 3D-plot of the average AUC performance on 10 data replicas under every combination of feature scheme and sample size. The vertical axis represents the mean AUC value. Since the 3D-plot of ρ exhibits similar pattern as AUC, it is not displayed because of redundancy.

Fig. 5 displays the box-plots of the pairwise rank differences between eight feature schemes under five different sample sizes. Only the rank box-plot for ρ is provided, as AUC exhibits similar patterns. Yellow colored cases indicate significant differences. The tick label on the x-axis shows which two features schemes are compared. A positive rank difference means that the first method performs worse than the second. Index of 1 to 8 corresponds to the eight feature schemes described in Section II-C.4. From the pairwise comparison, it is apparent that the majority of significances comes from the full feature scheme. No statistically significant differences were found among three levels of clinical codes, neither with ICD10, nor with DRG codes. Moreover, the advantage of full feature scheme was not significant on small samples (N=1,000) and became more obvious increasing sample sizes.

6854
IV. DISCUSSION AND CONCLUSIONS

Some main findings of the present study are:

First, our study demonstrates that increasing the sample size could greatly improve the predictive performance when predicting future days in hospital. This discovery is reasonable as more observations make learning algorithms more stable and accurate.

Secondly, it is unexpected to see that all the features which were derived either from ICD10 or DRG codes by using their standardized hierarchical structure, did not increase the predictive value significantly when compared to the baseline model. In addition, all hierarchies did not differ with each other significantly, even when sample sizes increased. This result does not support our original hypothesis that low-level (more specific) codes could improve the performance for larger datasets. Moreover, our results are not in line with the findings from Zhao et al. [6]. This may be caused by addressing different scientific questions or by the fact that our approach for extracting features from ICD10 codes is similar, but not identical.

Thirdly, combined use of enriched demographic features and clinical features - feature scheme ‘full’ - gave better performance as compared to using them separately. The reason could be that in a general population, the majority has no hospitalization. Therefore, in the full feature scheme, the enriched demographic information, such as the type of health insurance product purchased by customers, is expected to help discriminating individuals of no hospitalization from those who will have hospitalization in the future. Hence, AUC and $\rho$ were improved when using full feature scheme.

In conclusion, our results indicate that it is not straightforward to increase model performance by grouping clinical codes using their hierarchical structure. ICD10 codes are hierarchically structured according to body system and etiology [5]. Such hierarchies do not sufficiently reflect the consumption of hospital resources and therefore may not contribute much to the predictive performance. Hence, further analyses are needed to evaluate other ways of clustering for their ability to better utilize the clinical information with these codes for predicting future health outcomes. One possible solution could be using a similar concept to the Charlson Comorbidity index. The Charlson index is an indicator of general state of health and can easily be obtained from the secondary diagnoses, which are represented by ICD10 codes [10]. However, only a small proportion of ICD10 codes can be used of calculating the Charlson index. Hence, a more sophisticated approach for grouping clinical codes taking account of patients’ use of hospital resources or health status needs to be invented, which will be a very interesting research direction and worth future investigation.

REFERENCES