Improving Hospital Readmission Prediction Using Domain Knowledge Based Virtual Examples

Milan Vukicevic[1(✉)], Sandro Radovanovic[1], Ana Kovacevic[1,2], Gregor Stiglic[3,4], and Zoran Obradovic[2]

[1] Faculty of Organizational Sciences, University of Belgrade, Jove Ilića 154, Belgrade, Serbia
{milan.vukicevic, sandro.radovanovic}@fon.bg.ac.rs, ana.kovacevic@temple.edu
[2] Center for Data Analytics and Biomedical Informathics, Temple University, 1925 North 12th Street, Philadelphia, PA, USA
{zoran.obradovic}@temple.edu
[3] Faculty of Health Sciences, University of Maribor, Zitna Ulica 15, 2000 Maribor, Slovenia
[4] Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia
gregor.stiglic@um.si

Abstract. In recent years, prediction of 30-day hospital readmission risk received increased interest in the area of Healthcare Predictive Analytics because of high human and financial impact. However, lack of data, high class and feature imbalance, and sparsity of the data make this task so challenging that most of the efforts to produce accurate data-driven readmission predictive models failed. We address these problems by proposing a novel method for generation of virtual examples that exploits synergetic effect of data driven models and domain knowledge by integrating qualitative knowledge and available data as complementary information sources. Domain knowledge, presented in the form of ICD-9 hierarchy of diagnoses, is used to characterize rare or unseen co-morbidities, which presumably have similar outcome according to ICD-9 hierarchy. We evaluate the proposed method on 66,994 pediatric hospital discharge records from California, State Inpatient Databases (SID), Healthcare Cost and Utilization Project (HCUP) in the period from 2009 to 2011, and show improved prediction of 30-day hospital readmission accuracy compared to state-of-the-art alternative methods. We attribute the improvement obtained by the proposed method to the fact that rare diseases have high percentage of readmission, and models based entirely on data usually fail to detect this qualitative information.

Keywords: Virtual examples · Electronic health records · Hospital readmission · Domain knowledge

L. Uden et al. (Eds.): KMO 2015, LNBIP 224, pp. 695–706, 2015.
DOI: 10.1007/978-3-319-21009-4_51

1 Introduction

Development of accurate predictive analytics models in healthcare has large benefits for many stakeholders. Hospitals can benefit from healthcare predictive analytics by monitoring of quality indicators, planning of accommodation capacities, optimizing level of supplies, etc. Insurance companies can define adequate charging policies; medical doctors can use decision support in diagnostics, while patients can receive better quality of care, assessment of real costs by different hospitals, etc.

Potential of Electronic Health Records (EHR) predictive analytics applications is recognized, and we are witnessing rapid increase in the number of researchers that are trying to create accurate models for prediction of admission and readmission rates [21], readmission risk [24], cost-to-charge ratio [18], mortality rates, disease prediction [4, 23] etc.

Prediction of 30-day hospital re-admission takes a special place in predictive analytics research. Timely identification of potential un-planned readmissions can have high impact on improvement of healthcare services for patients, by reducing the need for unnecessary interventions and hospital visits. In addition, hospital readmission is considered as one of the most important indicators of quality of care for hospitals, with great economic impact. It is reported that readmission rate within 30 days was 19.6 %, 34.0 % within 90 days, and 56.1 % within one year following discharge. According to the Institute for Healthcare Improvement, of the 5 million U.S. hospital readmissions, approximately 76 % can be prevented, generating annual cost of about $25 billion [2].

Because of its complexity, importance, and involved financial resources (additional costs of several billion dollars annually), this problem is highly regarded in the medical community, and it is frequently addressed in recent research [6]. Unfortunately, researchers often fail to evolve highly accurate predictive models because of high dimensionality, class and feature imbalance, and input space sparsity. Therefore, this problem is highly challenging for the data mining community. Additionally, in this area researchers are often confronted with the lack of data that can show up because of different reasons, e.g. due to long and expensive clinical trials [3] or lack of data due to rare appearance of diseases [1, 11]. In other words, for each prevented re-hospitalization we could save 6,579 USD.

One possible way to address these problems is utilization of virtual examples (VE) [17], used as additional training samples and created from the current set of examples by utilizing specific knowledge about the task at hand. Compared to simple randomization techniques, incorporation of prior knowledge may contain information on a domain not present in the available domain dataset [15] and thus exploits advantages in domain knowledge (knowledge driven) and data driven complementary information sources. Virtual examples are successfully used in computer vision [10, 25] using the knowledge about rotated representation of 3D models and smoothness. In the medical area, for situations where clinical trials are long and expensive, VE that use differential equations [3] or medical models [5] are successfully applied. However most of these models are designed for generation of continuous data and to the best of our knowledge, there is no VE generator that can produce binary data from Electronic Health Records by utilization of domain medical knowledge.

Our contribution: We propose a method for virtual example generation that exploits domain knowledge given in the ICD-9-CM hierarchy of diseases. The proposed technique allows characterization of unobserved comorbidities (disease co-occurrences diagnosed in the moment of patient discharge from hospital) and thus removing bias of algorithms towards frequently observed diseases and comorbidities. We evaluate the proposed method on 66,994 pediatric hospital discharge records from California, State Inpatient Databases (SID), Healthcare Cost and Utilization Project (HCUP) in the period from 2009 to 2011 to illustrate the proposed idea.

2 Related Work

Machine learning algorithms are common tools for prediction of hospital readmission [8, 24]. Steele et al. [22] used a Bayesian Belief Network to predict perioperative risk of clostridium difficile infection following colon surgery. Data used in this research are gathered from the Nationwide Inpatient Sample (NIS) registry from 2005 to 2007. Naive Bayes method is used to select features for the Bayesian Belief Network, which is constructed using data from years 2005 and 2006. Evaluation of the generated model was conducted on 2007 data, having area under curve 0.746. Additionally, authors stated that a lot of data preprocessing is needed in order to get usable results.

Shams et al. [19] created a hybrid classification model which is used to distinguish readmission for heart failure, acute myocardial infarction, pneumonia and chronic obstructive pulmonary disease on 2011-12 Veterans Health Administration data in the State of Michigan. Authors concluded that a lot of efforts were made in order to build risk prediction models, but most of them fail to reach satisfactory accuracy level. A potential reason that authors noticed for this is that developed models often do not make distinctions between planned and unplanned readmissions. Therefore, they developed a new readmission metric that has the ability to identify potentially avoidable readmission from all other types of readmission. Results obtained on the proposed readmission metric were very promising, with area under curve just over 0.8 for all above-mentioned diseases.

Even though many efforts are invested in EHR analysis and predictive modeling, and secondary use of EHR has large potential, researchers are not satisfied with the results for diagnosis or readmission prediction. Therefore, Ooi et al. [16] advise that machine learning alone cannot be used for this purpose. They propose a hybrid human-machine database engine where the machine interacts with the subject matter experts as part of a feedback loop to gather, infer, ascertain, and enhance the database knowledge and processing and discuss the challenges towards building such a system through examples in healthcare predictive analysis. McCoy et al. [14] also stress that machine learning on only EHR records is not sufficient and that clinical text is a major component of EHR data and often contains rich information describing patients.

However, in situations with the lack of data, high class or feature imbalance machine learning algorithms have limited potential. In previous studies, for addressing such problems, virtual examples have been mainly used for dealing with limited observations or low quality data [85]. Several studies proposed powerful tools for effective generation of virtual examples based on specific prior knowledge in domain

specific applications. For example, in computer vision, rotated representation of 3D models and smoothness is successfully used to consider the integrated effects and constraints of data attributes by using mega-trend diffusion functions, and feasibility-based programming models [10, 25]. They showed that oversampling with VE improves performance and stability of learning algorithms.

In the area of healthcare predictive modeling, virtual examples are successfully used for sepsis analysis. VE are of crucial significance for early sepsis prediction, since patients infected by this disease often die in the early stage and thus temporal data cannot be gathered. Recently proposed predictive models for addressing this problem are based on VE that use differential equations [3], or medical models [5] as prior knowledge sources.

It can be concluded that VE are a useful, and sometimes the only possible, way to address the problem of lack (or imbalance) of data in predictive modeling. To the best of our knowledge there is no VE generator that uses domain knowledge in order to enrich EHR. Therefore, we developed a virtual examples generator that uses domain knowledge information extracted from ICD-9 hierarchy of diseases and complement HER records about 30-day hospital re-admission prediction.

3 ICD-9 International Classification of Diseases

The ICD-9 EHR codes are organized in a hierarchy where an edge represents 'is-a' relationship between a parent and its children. Hence, the codes become more specific as we go down the hierarchy [20]. Each three digit diagnostic code (group of diagnoses) is associated with a hierarchy tree of ICD codes on lower level. In this paper, we refer to it as a 'top-level' diagnostic code. Figure 1 shows a part of the hierarchy within the top-level (most general) diagnostic code 530 that represents diseases of the esophagus. Top-level can be represented as a set of four level diagnoses, which present more specific diagnoses. Further, one four digit code can be specified to more specific concepts (five digit codes).

4 Exploratory Analysis of Pediatric 30-Day Hospital Readmission Data

In this research, hospital discharge data from California, State Inpatient Databases (SID), Healthcare Cost and Utilization Project (HCUP) [7], Agency for Healthcare Research and Quality was used. This data tracks all hospital admissions at the individual level, having a maximum of 25 diagnoses for each admission. Every diagnosis is presented as an ICD-9-CM code. Beside diagnoses, every admission is explained with an additional 21 features (e.g., sex, age, month of admission, length of stay, total charges in USD, etc.). We used all pediatric patient data in California from January 2009 through December 2011 in the pre-processing phase.

Since there are over 14,000 ICD-9-CM codes, usage of codes as categorical features would be infeasible. Therefore, we transformed diagnoses to binary features, where if a patient had that diagnosis, value would be positive, otherwise negative.

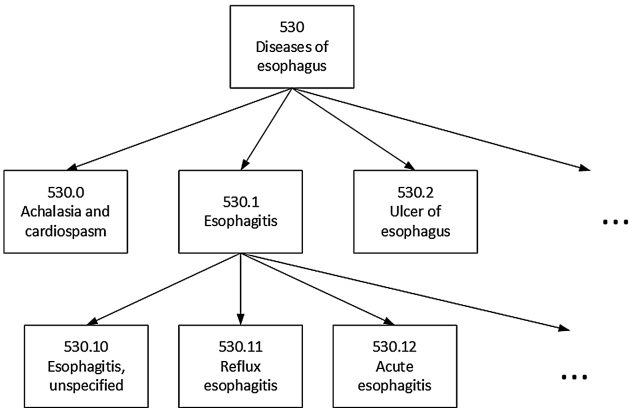

Fig. 1. Excerpt of three digit ICD-9 hierarchy

With this step, we transformed 15 features with very high cardinality (over 14,000 possible values) to over 14,000 binary features. Further, we excluded features with zero positive values. After this transformation, 851 ICD-9-CM codes remained, leading to a total of 872 binary features that we used to predict whether a patient will be readmitted within 30 days.

In order to characterize examples that could potentially be useful for improvement of the accuracy of predictive algorithms, we conducted exploratory analyses of the data with regard to the knowledge provided by ICD-9 hierarchy.

First, we investigated the balance between appearance of different diseases (on the lowest level of hierarchy) and their contribution to cumulative re-admission. Figure 2 (left), shows that most of the diseases appear rarely. In contrast, in cumulative, rare diseases constitute a large portion of re-admissions. This can be clearly seen on Fig. 2 (right), where on the X-axis diseases are represented in ascending order by the frequency of appearance, while on the Y-axis, cumulative share of each disease in total number of readmissions is showed. In particular, point (2000, 0.8) in this figure means that 80 % of readmissions are related to 1000/14000 (or, 14 %) of the least common primary admission diagnoses.

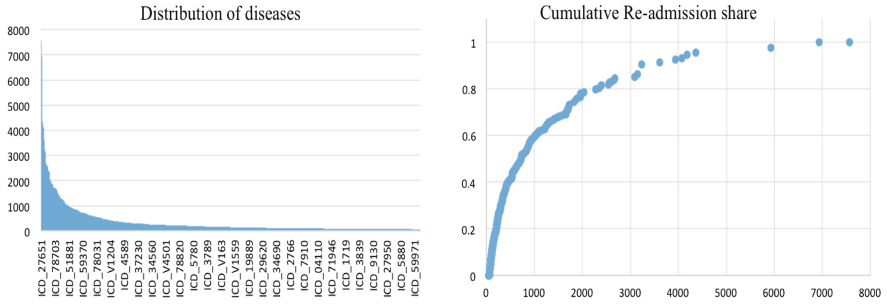

Fig. 2. Distribution of diseases (left) and cumulative readmission rate (right)

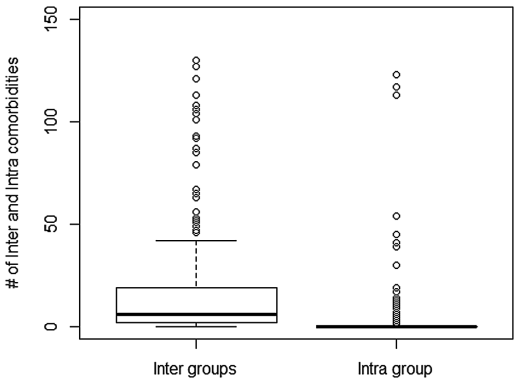

Fig. 3. Comparison of inter and intra 3digit ICD groups comorbidity frequencies

The large share of total re-admissions that is caused by rare diseases induces a problem for evolving of accurate predictive algorithms, since most of them are biased towards frequent features. Further, we expanded our research on investigation of comorbidity (co-occurrences of different diseases in each patient record) frequencies and found that comorbidities between ICD9-3 digit groups are much more frequent than comorbidities within groups. This is clearly shown on Box Plots in Fig. 3.

Based on previous observation, we focused on the inter-group comorbidities and further investigated their relation with re-admission. We analyzed standard deviation of re-admission rate between comorbidities from one group with diseases in other groups. The main hypothesis here was that if diseases from one ICD9-3digit group have small standard deviation of re-admission rate when they appear as comorbidities with diseases from other groups, then we can generate virtual examples that will connect rare (or even un-observed) diseases with high confidence about their re-admission. Analyses of standard deviation of readmission risk by all 3-digit groups (Fig. 4) showed that each 3-digit group has average standard deviation of re-admission less than 0.2. Most of them are below 0.11 or even smaller.

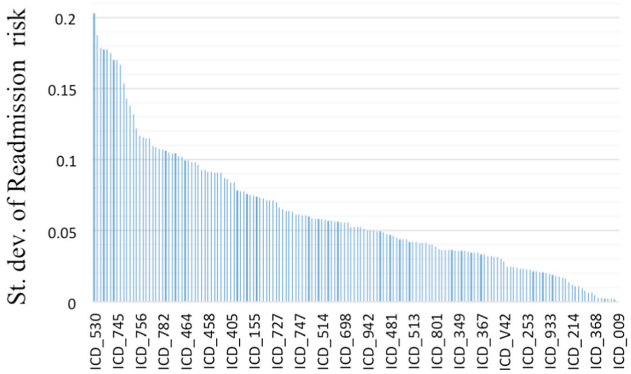

Fig. 4. Average standard deviation of re-admission between diseases of one group with other groups

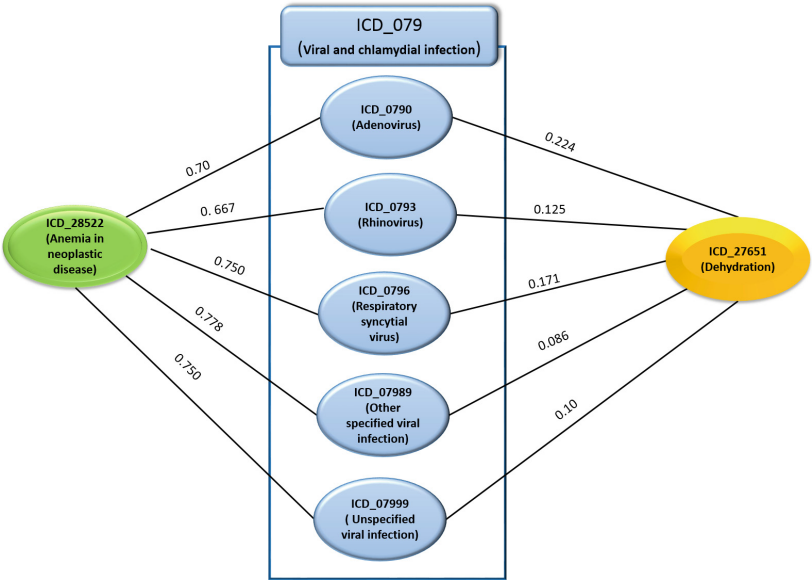

Fig. 5. Similar readmission risk of comorbidities from 3-digit group

This finding allowed generation of virtual examples that could potentially improve accuracy of predictive algorithms. In the following section we propose a data generator that implements findings from exploratory analyses of 30-day Hospital re-admission data. Figure 5, shows the situation where all diseases from one 3-digit group (central part) have similar readmission risks (RR) when connected with other diseases (all diseases from ICD_079 have high RR when co-occur with ICD_28522 and low RR when co-occur with ICD_27651).

5 ICD-9-VEG: Knowledge Based VE Generator

Based on considerations from the previous section we propose an ICD-9 based virtual example generator (ICD9-VEG). In order to address the problem of imbalanced features (since it is observed that rare diseases make large portion of re-admissions) generation of new examples (patient discharge record) goes as follows: First *a priory* probability of disease appearance is calculated for each disease. Initial disease is selected based on inverted probabilities (largest probabilities for selection have rare diseases).

When the first diagnosis is selected, comorbidity subset (CS) is formed from all diagnoses that have comorbidities with the selected diagnosis. The next feature is chosen based on λ-updated probabilities from CS with probability update not only from comorbidities with previously selected diseases, but all comorbidities of 3-digit hierarchy level that selected diagnoses originate from. This extends the space of possible diagnosis and allows knowledge-guided selection of unseen cases. Intuition behind this approach is that diagnoses from the same hierarchical group are often treated the same way, having

similar symptoms and diagnostic criteria (i.e. [9]), meaning that real diagnosis could be overlooked. This intuition is also confirmed in Sect. 4. Further, after selection of each disease, the decision about adding more comorbidities is made based on conditional probability that selected subset of diseases appears with other diseases. Thus, if examples in CS have more diagnoses than new virtual examples, then probability of continuing is greater. In order to enable generation of unseen comorbidities we set up a parameter of continuation that allows adding new comorbidities even if they are not observed in the initial set. Finally, when all diseases are created for the record, readmission outcome (true or false) is assigned based on readmission risk of the CS selected.

When new virtual examples are created they are stored in the VE variable, but they are also added to the initial dataset. The reason for this step is explained in Sect. 3, where we observed that "rarely" observed diseases and comorbidities make large portion of re-admissions. With this step we perform probability averaging of diagnoses, which leads to balancing of disease appearance in the initial data. Level of randomization and ICD9 influence is controlled by a smoothing parameter that controls smoothing level (by increasing **λ,** probability of generating unseen comorbidities grows). Users also provide the number of examples to be generated a parameter for smoothing variables other than diagnosis. The flow of the ICD9-VEG algorithm is depicted on Fig. 6.

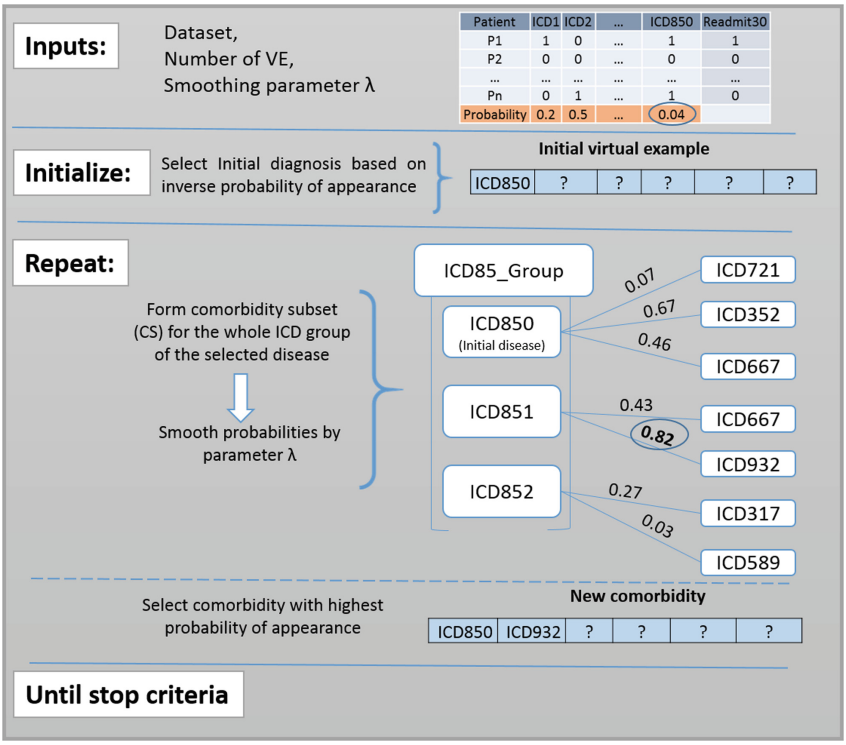

Fig. 6. Description of ICD9-VEG generator

6 Experimental Evaluation

In order to evaluate an influence of virtual examples on the predictive performance of algorithms we used Logistic regression and Naïve Bayes, because of their interpretability and success in healthcare applications [24]. Evaluation is performed on all pediatric patient data (described in previous section) where patient records that were admitted in period 2009-2010 were used for training (46,682 cases), and 2011 for testing (20,312 cases).

Influence of adding virtual examples generated by ICD9-VEG is compared with commonly used oversampling. We created two oversampling sets as a benchmark. First, "Class balance" was prepared by oversampling minority class to completely balance classes (37872 examples with positive re-admission rate are replicated from the initial 7611 samples, and added to the training set). In order to inspect the influence of using domain knowledge in the data generation process we prepared another benchmark set by using similar strategy as ICD9-VEG. "Feature balance" set is prepared by oversampling records that have rare diagnosis (we used the same technique of randomization as in ICD-9VEG). This way the influence of incorporating domain knowledge can be clearly explained and fair comparison of the approaches is made possible.

Additionally, we compared our approach to common ensemble algorithms such as AdaBoost and Bagging. AdaBoost and Bagging are performed in 10 iterations, where Bagging selected 70 % of the dataset with replacement. Experiments are performed 10 times and average is presented. Under-sampling is not considered in this research, since a huge amount of information would be lost.

Since hospital re-admission is highly class imbalanced (in our case there are 11,884 positive and 55,810 negative cases in all three years), in order to inspect in more detail the influence of new information on predictive performance we used four accuracy measures: Area Under Curve (AUC), Precision, Recall and F1 score. Table 1 shows performance of Logistic regression with and without oversampling or boosting techniques.

It can be seen from Table 1 that ICD9-VEG based oversampling gave the best results measured by AUC, precision, and F1 score. It is interesting to note that the "Class balance" technique led to the best recall. This means that Logistic regression managed to predict accurately the largest percentage of readmitted patients. On the

Table 1. Logistic regression performance

Logistic Regression	AUC	Precision	Recall	F1 score
ICD9-VEG	**0.802**	**0.715**	0.357	**0.476**
Logistic regression	0.787	0.693	0.344	0.460
Feature balance	0.775	0.670	0.313	0.427
Class balance	0.784	0.086	**0.366**	0.139
Bagging	0.776	0.669	0.350	0.460
Boosting	0.740	0.693	0.344	0.460

Table 2. Naïve Bayes performance

Naïve Bayes	AUC	Precision	Recall	F1 score
ICD9-VEG	**0.745**	**0.412**	0.683	**0.472**
Naïve Bayes	0.668	0.341	0.664	0.451
Feature balance	0.672	0.339	0.677	0.455
Class balance	0.659	0.080	0.308	0.127
Bagging	0.718	0.326	**0.706**	0.446
Boosting	0.701	0.371	0.664	0.451

other hand, Precision by Class balance is drastically lower compared to other techniques. This means that this type of oversampling biased the classifier to the minority class, and thus it almost always predicts that class. This is reflected in AUC and F1 as general measures of quality of classifiers. In Table 2, the results for Naïve Bayes are showed. It can be seen that performance of Naïve Bayes was worse than Logistic regression and showed higher tendency of predicting rare class (low precision and high recall). The pattern of performance by different oversampling and ensemble techniques stayed the same as for Logistic regression: ICD9-VEG was again the best performing technique by all measures except recall, where Bagging takes the lead.

To take the results into perspective we can estimate the financial impact of our results according to a study by Behara et al. [2] where an estimated cost of prevented re-admission accounts to 6,579 USD. Observing precision, out of 1000 positively classified selected patients, our approach will correctly identify 412 of them as positive. In comparison to Naïve Bayes this is 71 patients more which amounts to almost 0.5 million USD in savings for only 1000 patients selected by both methods (assuming that we can treat each patient in a way that will prevent re-admission).

Discussion: We showed that incorporating domain knowledge from ICD-9 hierarchy into oversampling techniques can improve classifier performance. We attribute that to the fact that rare diseases are the ones that have high percentage of readmission, and models usually fail to detect this information. By generating VEs which use this information, we obtain better performance. It is worth noticing that AdaBoost and Bagging didn't improved performance for logistic regression, which indicates over-fitting. On the other hand, the ensemble approach did improve performance in the Naïve Bayes algorithm. For both Logistic Regression and Naïve Bayes, class balancing with oversampling showed biased results towards minority class (positive re-admission) and feature balancing without incorporation of prior knowledge slightly reduced the performance.

7 Conclusion and Future Research

In this study we addressed the problem of insufficient training data and feature imbalance in healthcare predictive analytics by proposing a method for virtual example generation that exploits domain knowledge and available data as complementary sources. Experimental evaluation showed that this strategy improves predictive

performance of popular classification algorithms and shows competitive performance with traditional oversampling and ensemble strategies. Additionally, the proposed technique allows generation of unobserved comorbidities and thus removes bias of algorithms towards frequently observed diseases and comorbidities. The proposed approach is therefore useful in cases of rare conditions with low prevalence or even in cases where we would like to observe the predictive performance of classification models in unseen scenarios (e.g. disease outbreaks, epidemics or pandemics). As a part of our future work, we will extend the proposed framework by different sources of medical prior knowledge e.g. patient, disease, drug ontology. Additionally, we plan to extend the virtual example generator to generation of continuous data. This will allow generation of more input/output features that are available from original data (i.e. patients age or length of stay in the hospital). The extended generator will be used for enriching EHR repositories for improving predicting performance of the algorithms on readmission prediction as well as other healthcare related tasks, such as length of stay, admission, and mortality rate predictions.

Acknowledgments. This research was supported by DARPA Grant FA9550-12-1-0406 negotiated by AFOSR, National Science Foundation through major research instrumentation, grant number CNS-09-58854, and by SNSF Joint Research project (SCOPES), ID: IZ73Z0_152415.

References

1. Bavisetty, S., Grody, W.W., Yazdani, S.: Emergence of pediatric rare diseases: review of present policies and opportunities for improvement. Rare Dis. **1**(1) (2013)
2. Behara, R., Agarwal, A., Fatteh, F., Furht, B.: Predicting Hospital Readmission Risk for COPD Using EHR Information. Handbook of Medical and Healthcare Technologies, pp. 297–308. Springer, New York (2013)
3. Cao, X.H., Stojkovic, I., Obradovic, Z.: Predicting sepsis severity from limited temporal observations. In: Džeroski, S., Panov, P., Kocev, D., Todorovski, L. (eds.) DS 2014. LNCS, vol. 8777, pp. 37–48. Springer, Heidelberg (2014)
4. Davis, D.A., Chawla, N.V., Christakis, N.A., Barabási, A.L.: Time to CARE: a collaborative engine for practical disease prediction. Data Min. Knowl. Disc. **20**(3), 388–415 (2010)
5. Ghalwash, M., Obradovic, Z.: A data-driven model for optimizing therapy duration for septic patients. In: Proceedings 14th SIAM Int'l Conference Data Mining, 3rd Workshop on Data Mining for Medicine and Healthcare, Philadelphia, April 2014
6. Hasan, O., Meltzer, D.O., Shaykevich, S.A., Bell, C.M., et al.: Hospital readmission in general medicine patients: a prediction model. J. Gen. Intern. Med. **25**(3), 211–219 (2010)
7. HCUP State Inpatient Databases (SID), Healthcare Cost and Utilization Project (HCUP). 2009–2011. Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/sidoverview.jsp. 27 Feb 2015
8. Krumholz, H.M., Lin, Z., Drye, E.E., Desai, M.M., Han, L.F., Rapp, M.T., Normand, S.L.T.: An administrative claims measure suitable for profiling hospital performance based on 30-day all-cause readmission rates among patients with acute myocardial infarction. Cir. Cardiovas. Qual. Outcomes **4**(2), 243–252 (2011)

9. Legname, G.: Novel approaches to diagnosis and therapy in neurodegenerative diseases. In: Toi, V.V. (ed.). IFMBE Proceedings, vol. 46, pp. 155–158Springer, Heidelberg (2015)
10. Li, D.C., Wen, I.H.: A genetic algorithm-based virtual sample generation technique to improve small data set learning. Neurocomputing **143**, 222–230 (2014)
11. Mathew, G., Obradovic, Z.: Distributed privacy preserving decision system for predicting hospitalization risks in hospitals with insufficient data. In: Machine Learning in Health Informatics Workshop: International Conference on Machine Learning Applications–ICMLA, pp. 178–183. Boca Raton, FL, USA (2012)
12. Mathew, G., Obradovic, Z.: Distributed privacy preserving decision support system for highly imbalanced clinical data. ACM Transactions on Management Information Systems. **4** (3) Article No. 12, October 2013 (2013)
13. Mathew, G., Obradovic, Z.: A distributed decision support algorithm that preserves personal privacy. J. Intell. Inf. Syst. **44**, 107–132 (2014)
14. McCoy, A.B., Wright, A., Eysenbach, G., Malin, B.A., Patterson, E.S., Xu, H., Sittig, D.F.: State of the art in clinical informatics: evidence and examples. Yearb. Med. Inform. **8**, 13–19 (2013)
15. Mirchevska, V., Luštrek, M., Gams, M.: Combining domain knowledge and machine learning for robust fall detection. Expert Syst. **31**(2), 163–175 (2014)
16. Ooi, B.C., Tan, K.L., Tran, Q.T., Yip, J.W., Chen, G., Ling, Z.J., Zhang, M.: Contextual crowd intelligence. ACM SIGKDD Explor. Newsl. **16**(1), 39–46 (2014)
17. Poggio, T., Vetter, T.: Recognition and structure from one 2D model view: Ob- servations on prototypes, object classes and symmetries. Artificial Intell. Lab., MIT, Cambridge, MA, A.I. Memo no. 1347, April 1992 (1992)
18. Polychronopoulou, A., Obradovic, Z.: Hospital pricing estimation by gaussian conditional random fields based regression on graphs. In: Proceedings of the 2014 IEEE International Conference on Bioinformatics and Biomedicine, Belfast, UK, Nov 2014
19. Shams, I., Ajorlou, S., Yang, K.: A predictive analytics approach to reducing 30-day avoidable readmissions among patients with heart failure, acute myocardial infarction, pneumonia, or COPD. Health care management science, 1–16 (2014)
20. Singh, A., Nadkarni, G., Guttag, J., Bottinger, E.: Leveraging hierarchy in medical codes for predictive modeling. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 96–103. ACM, September 2014
21. Srivastava, R., Keren, R.: Pediatric readmissions as a hospital quality measure. JAMA **309** (4), 396–398 (2013)
22. Steele, S., Bilchik, A., Eberhardt, J., Kalina, P., Nissan, A., Johnson, E., Stojadinovic, A.: Using machine-learned bayesian belief networks to predict perioperative risk of clostridium difficile infection following colon surgery. Interactive J. Med. Res. **1**(2) (2012)
23. Stiglic, G., Pernek, I., Kokol, P., Obradovic, Z.: Disease prediction based on prior knowledge. In: ACM SIGKDD Workshop on Health Informatics, in Conjunction with the 18th SIGKDD Conference Knowledge Discovery and Data Mining, Beijing, China, Aug 2012
24. Stiglic, G., Wang, F., Davey, A., Obradovic, Z.: Readmission classification using stacked regularized logistic regression models. In: Proceedings of the AMIA 2014 Annual Symposium, Washington, DC, Nov 2014
25. Yang, J., Yu, X., Xie, Z.Q., Zhang, J.P.: A novel virtual sample generation method based on Gaussian distribution. Knowl. Based Syst. **24**(6), 740–748 (2011)