A Data and Knowledge Driven Randomization Technique for Privacy-Preserving Data Enrichment in Hospital Readmission Prediction

Milan Vukicevic [*] Sandro Radovanovic [*] Gregor Stiglic [†] Boris Delibasic [*]
Sven Van Poucke [‡] Zoran Obradovic [§]

Abstract

In health care predictive analytics, limited data is often a major obstacle for developing highly accurate predictive models. The lack of data is related to various factors: minimal data available as in rare diseases, the cost of data collection, and privacy regulation related to patient data. In order to enable data enrichment within and between hospitals, while preserving privacy, we propose a system for data enrichment that adds a randomization component on top of existing anonymization techniques. In order to prevent information loss (inclusive loss of predictive accuracy of the algorithm) related to randomization, we propose a technique for data generation that exploits fused domain knowledge and available data-driven techniques as complementary information sources. Such fusion allows the generation of additional examples by controlled randomization and increased protection of privacy of personally sensitive information when data is shared between different sites. The initial evaluation was conducted on Electronic Health Records (EHRs), for a 30-day hospital readmission prediction based on pediatric hospital discharge data from 5 hospitals in California. It was demonstrated that besides ensuring privacy, this approach preserves (and in some cases even improves) predictive accuracy.

Keywords: virtual examples, electronic health records, hospital readmission

1 Introduction

Healthcare predictive analytics have a potential for high-impact applications for many stakeholders. Hospitals can benefit from healthcare predictive analytics by monitoring of quality indicators, planning of healthcare capacities, optimization of supply levels etc. Insurance companies can define adequate charging policies; medical doctors can optimize treatment using decision support in diagnostics while patients can receive better quality of care, assessment of real costs by different hospitals etc. Prediction of 30-day hospital re-admission takes a special place in predictive analytics research [22]. Timely identification of potential unplanned readmissions can have a high impact on the improvement of healthcare services for patients, by reducing the need for unnecessary interventions and hospital visits. In addition, hospital readmission is considered as a major indicator of quality of care for hospitals, with significant economic impact. It is reported that readmission rate was 19.6% within 30 days, 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 the annual cost of about $25 billion. [21]

Many researchers addressed this problem by building predictive models on secondary healthcare data, but they often failed to develop highly accurate models because of the lack of data. Regulations and privacy concerns often hinder the exchange of healthcare data between hospitals or other healthcare providers [22, 24]. This problem can be solved by two major strategies: secure multi-party computation (SMC) [22, 14], and randomization [8]. In the case of SMC, the sites cooperate to build the global prediction model without sharing the data itself, and these techniques have already shown their usefulness in many application areas [22, 14].

On the other hand, these techniques cannot help in situations where the lack of data originates from long and expensive clinical trials [2] or in the case of data from rarely observed diseases [1]. In such situations a randomization based pre-processing could be applied, where some noise is added to the original data prior to predictive modeling. Still, randomization techniques often hamper the utility of the model [14].

One way to address these problems is the inclusion

[*] Faculty of Organizational Sciences, University of Belgrade, Belgrade, Serbia. {milan.vukicevic, sandro.radovanovic, boris.delibasic}@fon.bg.ac.rs

[†] Faculty of Health Sciences, University of Maribor, Maribor, Slovenia. gregor.stiglic@um.si

[‡] Department of Anesthesiology, Critical Care, Emergency Medicine and Pain Therapy, Ziekenhuis Oost-Limburg, Genk, Belgium. svanpoucke@gmail.com

[§] Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, USA. zoran@ist.temple.edu

of additional training examples created from the current set of examples by utilizing specic knowledge about the task at hand (often called virtual examples VE [16]). Compared to simple randomization techniques incorporation of VE as training examples in machine learning not only preserves model accuracy (and data privacy) but often improves it [25]. This is explained because incorporation of prior knowledge may contain information on a domain not present in the available domain dataset [15, 18] and thus exploits advantages in domain knowledge (knowledge driven) and data driven knowledge as complementary information sources. In the area of healthcare predictive modeling, virtual examples are successfully used for sepsis analysis [17]. VEs 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 [2] or medical models [5] as prior knowledge sources. It can be concluded that VE can be useful, and sometimes are the only possible technique to compensate for the lack of data in predictive modeling.

In this paper, we propose a method for VE generation which uses labeled examples and domain knowledge in the form of the ICD-9-CM hierarchy of diseases. The proposed technique is based on perturbation techniques that preserve privacy, but also allow generation of unobserved comorbidities. We consider three perturbation techniques based on apriori probabilities (data driven) and ICD-9-CM hierarchy information (knowledge driven) in order to randomize examples in a controlled manner while preserving privacy and addressing the problem of potentially existing, but non-observed comorbidities in data at hand. Additionally, features that indicate patient identity (patient identification, hospital identification and year of admission) are excluded. The intuition for the inclusion of hierarchy of diseases is based on too specific diagnoses that medical experts can assign. In case of similar symptoms, medical experts can make mistakes and assign false diagnoses. However, such diagnoses often belong to the same group of diagnoses due to their similarity. Therefore, the inclusion of a hierarchy of diseases in order to emphasize diagnoses from the same group or to generate unseen comorbidities can be used as a privacy preserving technique.

2 State of the art

Virtual examples are very popular in areas where data are hard to obtain or where data interchange is impossible due to regulations. This kind of problems is defined as small sample data set problems. This means that sample size is smaller compared to the number of features which leads to a poor generalization of classifiers.

One of the first efforts in investigating the effects of virtual examples is presented in [19]. They compared two ways for incorporation of domain knowledge in learning algorithm. One approach was based on changing the learning algorithm and other by incorporating virtual examples. The first approach may perform better and faster, but it requires significant effort in changing the goal function, or optimization of the learning algorithm. An approach based on virtual examples have two major advantages. First, it improves accuracy of the learning algorithm and second, it can be readily implemented for any learning algorithm. However, the virtual examples based approach increases overall training time and specifically for support vector machines leads to situation where many virtual examples becomes support vectors hence decreasing classification speed.

Virtual examples can be divided into two categories. First, more popular category, is to generate virtual examples by extracting the nontrivial knowledge hidden in the problem being solved. The task of extracting nontrivial knowledge is being formulated as a probability density function estimation [16]. This approach improved performance of pattern recognition by, given a 3D view of an object, generating new images from other angles through mathematical transformations. Extracting prior knowledge is a highly challenging task which requires a lot of efforts. This approach assures rationality, however adaptability to other problems is very low. Our approach can be categorized as the extraction of nontrivial knowledge. However, we added formally specified domain knowledge in the form of ICD-9-CM ontology (hierarchy). This way information about similarity of features is included in the generation of virtual examples.

Another approach, called perturbation, is to generate virtual examples by adding the noise to the original examples. This approach often adds noise using uniform or normal distribution. An interesting perturbation examples are presented in [27] where training samples of the rare class are divided into p groups using the k nearest-neighbor algorithm, then generated virtual examples by averaging every two samples of each group, and leaving the labels unchanged. Compared to the first approach, adaptability of these methods is more evident, but rationality cannot be assured.

There are four problems the virtual generator needs to address: inputs, the strategy of the virtual examples generator, outputs and the number of virtual examples. Most of the virtual example approaches differ in strategy. To the best of our knowledge, there are several strategies of virtual example generation.

The first strategy randomly picks a sample inside of the hypersphere of a real samples input point, where the hypersphere is defined by uniform or Gaussian distribution. Since the point is selected near an original data point, it is similar, but not the same as an original data point. Further, the output is selected by a weighted average of original data points in a hypersphere or using evolutionary approach. This approach emphasizes that utilizing virtual examples does improve the performance of the classifier. [3]

Another approach of functional virtual population generation [11] is developed for specific types of systems such as manufacturing systems. The process of virtual examples generation starts from one system attribute and generate a specified number of virtual examples in the neighborhood of selected attribute. To test a virtual population, neural network is used, where the real performance of manufacturing system is used as an output. Once accuracy reaches its peak, a different system attribute is selected and the process is repeated. When all system attributes are processed, an integrated virtual population is used for artificial testing of the manufacturing process. Experimentally this approach dramatically improved learning accuracy and scheduling in a manufacturing system. This system inspired our strategy of virtual example generation. Similarly as selecting one system attribute, we select one diagnosis from ICD-9-CM codes (the proposed system is explained in more details in the following section).

The need for virtual examples in a small sample studies is explained in [7]. They elaborate that a classical network cannot recognize a non-linear function with a small sample. Therefore, they utilized the information diffusion principle. This principle asserts that, when an incomplete data set is used to estimate a relationship between features, there must exist reasonable diffusion means to change observations into fuzzy sets to partly fill the gap. They proposed a random generator controlled by probability density function as a diffusion function. Derived patterns are controlled to match using BP networks.

Another paper demonstrationg the importance of virtual examples in small data sets [9]. Small sample size learning cannot achieve high performance with respect to the overall population independent of the learning algorithm employed. However, small sample has a certain distribution, and virtual examples can be derived from density function obtained from intervalized kernel density estimation. This approach is shown to improve the performance of the learning algorithm. However, this approach seemed less suitable with nominal attributes (which is the case in our problem).

Virtual examples are highly utilized in medical domain, especially for the problem of rare disease and microarray analysis, which are formulated as small sample problems. One paper which utilizes virtual examples in order to improve the performance of learning algorithm for cancer identification is [12]. Their procedure for the binary classification problems with small sample data sets consists of three steps. First, a gene selection algorithm, which selects genes based on t-statistic value is employed to reduce dimensionality and improve learning ability (which is expected in gene expression problems due to high noise in attributes). Then, by utilizing group discovery technique, they profile related characteristics of each discriminative gene within a dataset. This step primarily searches for sample grouping (clusters) based on the spatial relationship between each other. As such, outliers are presented as a separate group. It is expected that clusters have the same label. Further, random noise is added to real examples using mean and standard deviation for each cluster. Simulation on both synthetic and real world data sets have shown that performance improved dramatically compared to the original data set. This paper motivated us to use groupings of features, not samples. Since we can utilize domain knowledge in the form of ICD-9-CM ontology (hierarchy) we grouped features which are similar in terms of effects and charging.

Virtual examples are also utilized in scheduling. Because of limited information in early dynamic flexible manufacturing systems, scheduling knowledge is difficult or impossible to obtain. Therefore, virtual examples must be generated and utilized for simulation. In [13] mega-trend-diffusion technique is performed to develop virtual examples. The mega-diffusion method diffuses a set of data using a common diffusion function with the objective to determine possible dataset coverage on a group consideration basis. When the group is found (domain range) samples are randomly selected from group and value of the diffusion function is added. The idea of grouping of data is also implemented in our research, where we grouped data based on hierarchy (domain knowledge). Namely, in the process of selection of samples, we used samples which have a same or similar diagnosis. Our approach is explained in more details in the following section.

Smoothing aiming for better estimation of pdf is used in [26]. Their virtual examples generator estimates parameters of Normal distribution from data. Further, using a random number generator they produce a virtual example. Although, this seems obvious it is mathematically and empirically shown that virtual examples improve the performance of learning algorithms for small size data sets and imbalanced data sets. These results motivated us to develop domain knowledge-based virtual examples generator.

Metaheuristics have also been used for virtual examples generators. In [10] a genetic algorithm is used for virtual example generation specially designed for small data prediction problems. Their mathematical model optimized mean absolute percentage error of linear regression function with constraints. The acceptable value of each attribute was determined with lower and upper bound. Virtual examples were defined as units in genetic algorithms, which were optimized in each iteration. The output of virtual example is defined based on real world samples with defined upper and lower bound. If the output does not satisfy these conditions the process is repeated. This way virtual examples are generated with an optimization procedure which reduces the error of learning algorithm. However, since virtual examples use class information adding the noise is essential in order to prevent overfitting.

Medical records including rare diseases are one of the most challenging prediction tasks where virtual examples generator is needed in order to obtain acceptable performance of learning algorithm.

In [5] a population of virtual patients is generated by random initialization of some parameters and by random initialization of the states initial conditions. Further, a patient is tracked over time using ordinary differential equations and based on results it can be either in survival group or in non-survival group. This random initialization and random selection of states, both using pdf from real data, have shown promising results. Another paper [2] produced in silico or virtual patients for sepsis prediction. Virtual examples were created using dynamical equation, but each of the patients has a unique set of parameters and therefore unique response to the CLP induction of sepsis. This is especially important since sepsis is highly progressive disease and early prediction is a must. As in majority of virtual example generator papers parameters are randomly sampled from predefined intervals and and if the likelihood for sepsis is high enough over time then the virtual patient is accepted as valid. It has been shown that this approach in combination with domain knowledge improves performance of prediction compared to a data driven approach. Therefore, we find this motivating to include formally written domain knowledge in order to improve performance of learning algorithm. In same domain (sepsis prediction) there is another approach for virtual patients generation which shown promising results

In [23] a feasibility based programming method is used as a virtual example generator. Model optimize mean absolute prediction error. Inputs are chosen randomly while outputs are defined using a genetic algorithm and backup propagation neural networks based feasibility-based programming model, with a constraint on output (must be inside lower and upper bound). When a virtual example is created latent spectral features are extracted which simplify model (thus reducing model training time). It has been shown that this approach improves the performance of learning algorithm for shell vibration and acoustic spectral data of a laboratory-scale ball mill.

This paper extends the first method for generating virtual examples which utilizes structured domain knowledge in the form of ontology (hierarchy) [25]. The hypothesis in this work is that using higher level concepts for probability smoothing and selection of diagnoses (as a step of virtual example construction) would positively influence readmission prediction and that this approach would enable data sharing between hospitals.

3 Proposed System

In order to address problems discussed above we propose a system for data enrichment and sharing of information about EHRs between hospitals that adds an additional layer of privacy protection into existing predictive modeling systems (Figure 1).The process of privacy protection starts with traditional anonymization techniques, which map personal and hospital identity into encrypted form. Additionally, time and duration of hospital visits are presented in relative form (number of days from initial admission), while exact dates are removed. Even though these techniques can substantially reduce the risk of patient identification, the state of the art predictive techniques theoretically can still identify the person based on procedures, diagnoses, and other data that cannot be encrypted if they serve as a basis for collaborative building and evaluation of predictive models. In order to increase privacy protection and allow data sharing and building of the more accurate predictive models, we propose a data enrichment mechanism that is based on randomization. However, data enrichment based on simple probabilistic randomization most often reduces the predictive performance, because of additional noise that is added to data. In order to prevent data quality loss by randomization we introduce a mechanism for fusion of data randomization techniques with the domain knowledge sources (ontologies or rules), and thus, randomization of the original data in a controlled manner.

We consider three types of EHR randomization: *a priori*, *knowledge-based* and *hybrid*. For the purpose of clarity, this will be more thoroughly explained later in the text. After anonymization and randomization, this additional example can be used for data enrichment within or between hospitals. Further, each hospital can build predictive models on enriched data (generated on its own or by other hospitals) and these models can be used for assessment of the risk of readmission for new patients. Finally, predictive models (classification, regression, etc.) can be built on enriched data sources and applied for many different problems in healthcare e.g. prediction of re-admission risk, a number of admissions in hospitals, cost-to-charge ratios, etc. In this research, we built and evaluated predictive models for readmission risk prediction. These models should serve as decision support for medical doctors when making a decision about diagnoses and/or therapy. High readmission risk can indicate that diagnosis or therapy are not adequate for the given patient and that doctor should re-examine the patient, or send him to additional testing in order to prevent potential readmission. The proposed system is depicted on Figure 1.

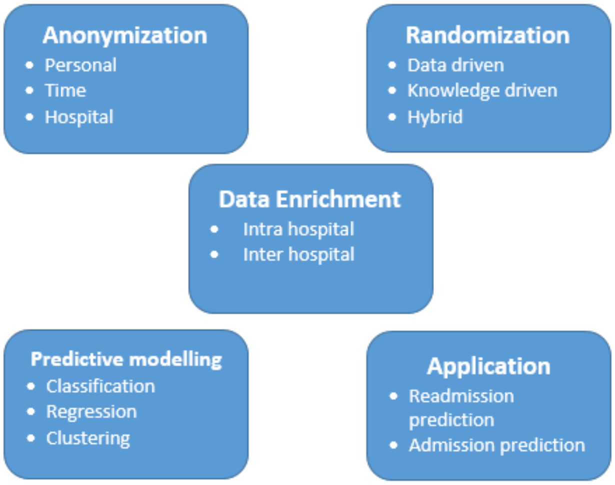

Figure 1: DSS for privacy-preserving sharing of data

In the further text, we explain in more detail the procedures for enabling data sharing through a priory (probability) based and knowledge guided randomization. These techniques are similar to one recently proposed [25], which was previously used for the generation of rare diseases and improved generalization of predictive algorithms. Here it will be used as a general knowledge-based randomization mechanism that allows more secure data sharing.

The additive (a priori) randomization approach uses a technique for smoothing the probabilities of every diagnosis, in a similar way as Laplace smoothing in the Naive Bayes algorithm. For each new VE, we start the generation process by selecting a diagnosis based on a priori probabilities of all diagnoses that are smoothed (increased) with parameter . The initial disease may be selected based on the highest probability of appearance (if most common disease from the hospitals should be shared), or inverted probabilities (if rare diseases should be shared). When the first feature (disease) is selected, the next disease (comorbidity) is selected in the following way: First, the comorbidity subset (CS) is formed with all diagnoses that have comorbidities with the previously selected diagnosis. Next, features are chosen based on λ-updated probabilities from CS. This procedure is iteratively repeated by forming CS based on conditional probabilities of comorbidities for already selected features. It is intuitively clear that this procedure will result in feature distribution that is similar to the original data. Namely, all new features will have the same or reduced set of features compared to the original dataset, where privacy will be preserved, but there is no chance of generating unseen comorbidities.

Knowledge-based randomization - enables generation of features (i.e. comorbidities) that are not observed in the original dataset. This generation can preserve privacy, but also, could be useful in situations when hospitals did not have patients with a specific set of diseases (and it is known that such a set can appear in the future). Of course, by using simple randomization such VE cannot be generated, and thus the process of randomization has to be guided by some form of domain knowledge.

In this study, we use hierarchical ICD-9 (excerpt of hierarchy is given in Figure 2) classification of diseases as a knowledge source. The ICD-9 codes are organized in a hierarchy where an edge represents an is-a relationship between a parent and its children. Hence, the codes become more specific as we go down the hierarchy [20]. When leveraging the ICD-9 hierarchy for generating virtual examples, we can assume that the child nodes have a correlated relationship with the feature of interest (selected feature). There are about 15,000 diagnostic codes in the ICD-9-CM hierarchy. Each three-digit diagnostic code is associated with a hierarchy tree. In this paper, we refer to it as a top-level diagnostic code. Figure 2 shows a part of hierarchy within the top-level (most general) diagnostic code that represents infectious and parasitic diseases. Top-level can be represented as a set of lower level concept group of diagnoses, which present more specific diagnoses. Further, that set of diagnoses can be specified to more specific concepts (five digit codes). Hierarchy used in this paper is Clinical Classification Software (CCS) which clusters patient diagnoses and procedures into clinically meaningful categories. [4]

When leveraging the ICD-9 hierarchy for generating virtual examples, we can assume that the child nodes

Algorithm 1 Pseudo-code for VE generator

```
Inputs: dataset D, # examples n, smoothing $\lambda$,
continue parameter cp, number of examples k
Output: list of virtual examples VE

VE = $\emptyset$ //initialize list of virtual examples
while k virtual examples are created do
  set CS = D //create comorbidity subset CS
  V = $\emptyset$ //initialize virtual example
  while cp is true do
    //calculate probabilities of diagnoses in CS
    //smooth probabilities of every ICD-9 code
    //smooth probabilities of similar diagnoses
        $\mathbf{P} = \frac{|X| + \lambda \times |X| + \lambda \times |X_{cs}|}{n_{cs}}$
    if first step then
      //invert probabilities
        $\mathbf{P} = \frac{1}{\mathbf{P}}$
    end if
    //add disease i to V
    //using roulette wheel selection
    Add(V, i)
    //select CS with examples having at least
    //one diagnosis from three level group of
    //selected diagnosis
    CS = $D_{cs}$
    //calculate ratio of examples in CS with
    //higher number of diagnoses and number of
    //examples with lower number of diagnosis
        $ratio = \frac{|CS_{>}| + \lambda \times |CS_{>}|}{|CS| + 2\lambda \times |CS_{>}|}$
    if random number $\geq$ ratio then
      cp = false
    end if
  end while
  //roulette wheel selection for other features
  //excluding hospital and date of admission

  //add virtual example to list
  Add(VE, V)
end while
```

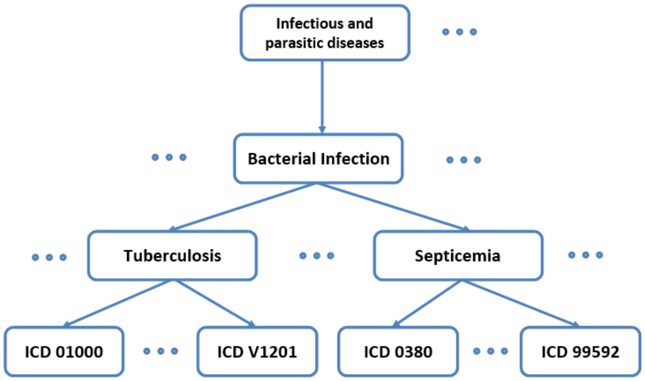

Figure 2: ICD-9 hierarchy of diseases

and a parent node are both correlated with the feature of interest (in this case, the risk of 30-day hospital readmission). So the main idea is to generate VEs with similar readmission outcome for diseases or comorbidities (combination of diseases) from the same hierarchy group.

The first step is the same as in Additive smoothing: the first diagnosis is chosen from rare diagnoses that are favored for selection. The main contribution is the iterative step, where CS is formed not only from comorbidities with previously selected diseases but all comorbidities of 3-digit hierarchy level that selected diagnoses to originate from. This extends the space of possible diagnosis (now not only comorbidities with one diagnosis are considered, but comorbidities with the hierarchy group) 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 and that on the low level of hierarchy diagnosis could be too specific, since various diagnoses from the same group at symptom level seem to share similar behavioral symptoms and diagnostic criteria [28], meaning that real diagnosis could be overlooked.

This way it is possible to adapt models for the unseen cases, but also to randomize them in a controlled manner and thus preserve privacy when sharing data.

Integrated randomization (Additive and ICD9 based) smoothing combines previously described approaches by executing Additive and ICD9 smoothing, respectively. After execution, feature probabilities are updated by the sum of aforementioned smoothing updates. Further CS is formed the same way as in the ICD9 smoothing. The level of randomization and ICD9 influence is controlled by smoothing parameters that control smoothing levels for each type of smoothing. Users also provide the number of examples to be gen-

Table 1: Accuracy of logistic regression (AUC) when using enriched data of a single hospital versus using an individual hospital data alone or shared data from all hospitals.

# Examples	# Readmitted	% Readmitted	Individual	Shared	Enriched
7884	1,336	16.95	0.695	**0.820**	0.815
6394	1,450	22.68	0.693	**0.793**	0.771
6317	1,064	16.84	0.644	**0.782**	0.762
5103	705	13.82	0.621	0.780	**0.794**
4405	813	18.46	0.636	0.728	**0.761**
7884	1,336	16.95	0.695	**0.825**	0.817
6394	1,450	22.68	0.693	0.802	**0.810**
6317	1,064	16.84	0.644	**0.791**	0.741

erated and a parameter for smoothing variables other than diagnoses. Pseudo-code is given in Algorithm 1.

4 Experimental Evaluation

In this research, we addressed the problem of hospital readmission prediction in situations where EHRs are not shared between hospitals. Our main hypothesis was that the controlled (knowledge guided) randomization of data can provide additional examples that can be shared in a more secure way and increase the performance of predictive models built by each hospital.

4.1 Data In Hospital discharge data from California, State Inpatient Databases (SID), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality was used [6]. This data tracks all hospital admissions at the individual level, having a maximum of 15 diagnoses for each admission. Since there are over 14,000 ICD-9-CM codes, and using diagnoses as 15 polynomial attributes would be unfeasible for any learning algorithm to handle, we transformed the feature space by presenting each code as a feature. Therefore, we have about 14,000 binary features, where positive value marks the presence of the diagnosis. The final data set was preprocessed as in [22], with 850 input features (diagnoses) and as predictors for single binary output (patient was re-admitted within 30 days or not).

4.2 Experimental setup Since Data from 2009 and 2010 (about 2/3 of the entire data set) were used for training, while data from 2011 was used for testing. As a learning algorithm, we used logistic regression (LR), since it often showed good performance in medical applications, also performing well on this type of data and most importantly, providing interpretable models. Interpretability of models is especially important in Healthcare predictive analytics because of high costs of wrong decisions. We used all pediatric patient data from 8 hospitals with the highest numbers of patients and highest numbers of different diseases and highest number of patients.

Hybrid strategy (both additive and knowledge based randomization) was used in order to generate additional examples in a controlled (knowledge guided) manner. For each hospital, the same number of randomized examples is created, leading to a repository of 30,103 examples that were used for data enrichment of each hospital. In order to show usefulness of enriching data from specific hospitals with virtual examples, we made the following sets of data (on which logistic regression is applied and evaluated):

- **Individual** LR was trained on data from a single hospital.to predict readmission at that hospital.
- **Shared** LR model is developed on integrated data from all hospitals.
- **Enriched** LR was trained on data from an individual hospital enriched with data from VE repository.

Since the data has a high class imbalance (about 20% of all patients were readmitted), we evaluated all models with Area Under Curve (AUC) instead of classification accuracy.

4.3 Results In contrast to medical applications where data sharing is not applied or not allowed, the proposed method can generate additional examples, which can allow developing more accurate and with better generalization power. Since there are a lot of hospitals with a relatively small number of admissions, at these hospitals this method can supplement missing examples. Table 1 shows brief data description and AUC values for each experiment on each hospital (larger values are better and the best performance is presented in bold letters).

It can be seen at Table 1 that sharing the data drastically improves model performance. All models that are built on Original data have AUC less than 0.696, while models on Shared data had AUC performance from 0.728, up to 0.825. Still, such sharing of data is often not possible due to strict data privacy regulations. On the other hand, models built using data from VE repository allow sharing the data without compromising privacy. It can be seen that models that are built on data from VE repository (and original data from each hospital) achieved results comparable to using shared data. Performance on all hospitals was very similar and for hospitals 4, 5 and 7 results were even slightly better.

5 Conclusion and Future Research

In this paper we proposed a method that allows privacy while preserving data sharing between hospitals. The system is based on domain knowledge guided randomization techniques, where domain knowledge is presented in the form of a hierarchy of diagnoses. It is shown that sharing the data through generated virtual examples as such improves model performance for hospital readmission prediction. We conclude that hospitals could reduce costs for readmitted patients by using data sharing and virtual examples.

In future work, we plan to extend the system to other types of domain knowledge sources, such as other hierarchies and ontologies, where additional information about relations between diseases is present.

6 Acknowledgment

This research was partially supported by the U.S. Office of Naval Research grant N00014-15-1-2729, and by SNSF Joint Research project (SCOPES), ID: IZ73Z0_152415.

References

[1] Bavisetty, S., Grody, W. W., & Yazdani, S. (2013). Emergence of pediatric rare diseases: Review of present policies and opportunities for improvement. *Rare Diseases*, *1*(1), e23579.

[2] Cao, X. H., Stojkovic, I., & Obradovic, Z. (2014, January). Predicting Sepsis Severity from Limited Temporal Observations. In *Discovery Science* (pp. 37-48). Springer International Publishing.

[3] Cho, S., Jang, M., & Chang, S. (1997). Virtual sample generation using a population of networks. *Neural Processing Letters*, *5*(2), 21-27.

[4] Elixhauser, A., Steiner, C., & Palmer, L. (2008). Clinical classifications software (CCS). *Book Clinical Classifications Software (CCS)*(Editor ed eds).

[5] Ghalwash, M., & Obradovic, Z. A Data-Driven Model for Optimizing Therapy Duration for Septic Patients. In *Proc. 14th SIAM Intl. Conf. Data Mining, 3rd Workshop on Data Mining for Medicine and Healthcare*, Philadelphia, PA, USA (April 2014).

[6] Healthcare Cost and Utilization Project. (2008). Introduction to the HCUP Kids inpatient database (KID) 2006. Rockville (MD): Agency for Healthcare Research and Quality.

[7] Huang, C., & Moraga, C. (2004). A diffusion-neural-network for learning from small samples. *International Journal of Approximate Reasoning*, *35*(2), 137-161.

[8] Kantarcioglu, M., Nix, R., & Vaidya, J. (2009). An efficient approximate protocol for privacy-preserving association rule mining. In *Advances in Knowledge Discovery and Data Mining* (pp. 515-524). Springer Berlin Heidelberg.

[9] Li, D. C., & Lin, Y. S. (2006). Using virtual sample generation to build up management knowledge in the early manufacturing stages. *European Journal of Operational Research*, *175*(1), 413-434.

[10] Li, D. C., & Wen, I. H. (2014). A genetic algorithm-based virtual sample generation technique to improve small data set learning. *Neurocomputing*,*143*, 222-230.

[11] Li, D. C., Chen, L. S., & Lin, Y. S. (2003). Using functional virtual population as assistance to learn scheduling knowledge in dynamic manufacturing environments. *International Journal of Production Research*, *41*(17), 4011-4024.

[12] Li, D. C., Fang, Y. H., Lai, Y. Y., & Hu, S. C. (2009). Utilization of virtual samples to facilitate cancer identification for DNA microarray data in the early stages of an investigation. *Information Sciences*, *179*(16), 2740-2753.

[13] Li, D. C., Wu, C. S., Tsai, T. I., & Lina, Y. S. (2007). Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge. *Computers & Operations Research*, *34*(4), 966-982.

[14] Mathew, G., & Obradovic, Z. (2013). Distributed Privacy-Preserving Decision Support System for Highly Imbalanced Clinical Data. *ACM Transactions on Management Information Systems (TMIS)*, *4*(3), 12.

[15] Mirchevska, V., Lutrek, M., & Gams, M. (2014). Combining domain knowledge and machine learning for robust fall detection. *Expert Systems*, *31*(2), 163-175.

[16] Niyogi, P., Girosi, F., & Poggio, T. (1998). Incorporating prior information in machine learning by creating virtual examples. *Proceedings of the IEEE*, *86*(11), 2196-2209.

[17] Radosavljevic, V., Ristovski, K., & Obradovic, Z. (2013). A data-driven acute inflammation therapy. *BMC Medical Genomics*, *6*(Suppl 3), S7.

[18] Radovanovic, S., Vukicevic, M., Kovacevic, A., Stiglic, G., & Obradovic, Z. (2015). Domain knowledge based hierarchical feature selection for 30-day hospital readmission prediction. In *Artificial Intelligence in Medicine* (pp. 96-100). Springer International Publishing.

[19] Schlkopf, S. P., Simard, P., Vapnik, V., & Smola, A. J. (1997). Improving the accuracy and speed of support vector machines. *Advances in neural information processing systems*, *9*, 375-381.

[20] Singh, A., Nadkarni, G., Guttag, J., & Bottinger, E. (2014, September). 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.

[21] Srivastava, R., & Keren, R. (2013). Pediatric readmissions as a hospital quality measure. *JAMA*, *309*(4), 396-398.

[22] Stiglic, G., Wang, F., Davey, A., & Obradovic, Z. (2014). Pediatric Readmission Classification Using Stacked Regularized Logistic Regression Models. In *AMIA Annual Symposium Proceedings* (Vol. 2014, p. 1072). American Medical Informatics Association.

[23] Tang, J., Jia, M., Liu, Z., Chai, T., & Yu, W. (2015, August). Modeling high dimensional frequency spectral data based on virtual sample generation technique. In *Information and Automation, 2015 IEEE International Conference* on (pp. 1090-1095). IEEE.

[24] Vest, J. R., & Gamm, L. D. (2010). Health information exchange: persistent challenges and new strategies. *Journal of the American Medical Informatics Association*, *17*(3), 288-294.

[25] Vukicevic, M., Radovanovic, S., Kovacevic, A., Stiglic, G., & Obradovic, Z. (2015). Improving Hospital Readmission Prediction Using Domain Knowledge Based Virtual Examples. In *Knowledge Management in Organizations* (pp. 695-706). Springer International Publishing.

[26] Yang, J., Yu, X., Xie, Z. Q., & Zhang, J. P. (2011). A novel virtual sample generation method based on Gaussian distribution. *Knowledge-Based Systems*, *24*(6), 740-748.

[27] Zhang L. & Chen G.H. (2006), Method for constructing training data set in intrusion detection system, *Computer Engineering and Applications*, *42*(28). 145 146.

[28] Zhou, J., Lu, Z., Sun, J., Yuan, L., Wang, F., & Ye, J. (2013, August). FeaFiner: biomarker identification from medical data through feature generalization and selection. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1034-1042). ACM.