

Semi-Supervised Prediction of Comorbid Rare Conditions Using Medical Claims Data

Chirag Nagpal¹, Kyle Miller¹, Tiffany Pellathy², Marilyn Hravnak²,
Gilles Clermont³, Michael Pinsky³, Artur Dubrawski¹

¹School of Computer Science
Carnegie Mellon University
Pittsburgh, PA, USA

²School of Nursing
University of Pittsburgh
Pittsburgh, PA, USA

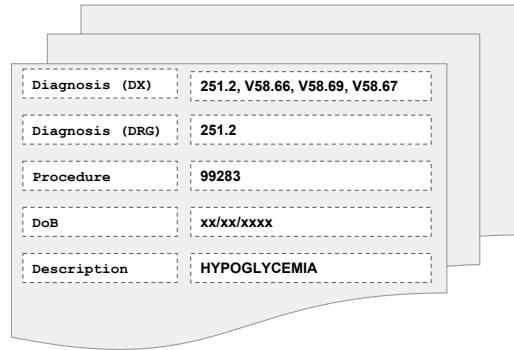
³School of Medicine
University of Pittsburgh
Pittsburgh, PA, USA

Abstract—Medical insurance claims data offer a coarse view of a patient’s medical profile, including information about previous diagnoses and procedures performed. These data have been exploited in the past to predict presence of unmanifested conditions. Rarer conditions however, provide an extremely limited amount of ground truth to train supervised models, but predicting relevant co-morbidities can help reduce failure to rescue from a treatable, yet potentially life threatening condition. In this paper, we aim at a formidable task of improving models built to predict comorbidity of rare conditions that emerge during hospitalization and present PRECORC, a novel approach that leverages hierarchical structures of diagnosis and procedure codes to alleviate the relatively low prevalence of specific types of Failure to Rescue (FTR) incidents. It can be applied *post-hoc* over previously learnt predictive models, and used to discover parts of the underlying hierarchies that contribute to the task. Our experimental results demonstrate that PRECORC carries promise for operational utility in clinical settings, and offer insights into potential leading indicators of life threatening complications.

I. INTRODUCTION

Failure to rescue (FTR) is defined as the death of a hospitalized patient from one of the following treatable complications: pneumonia, shock or cardiac arrest, upper gastrointestinal bleeding, sepsis, or deep venous thrombosis. According to [1], the mean death rates among medical and surgical patients with one of these complications approach 20%, but the risk of death in these cases may be reduced through early identification and clinical intervention. However, the challenge of discovering and prioritizing for close monitoring those patients who are at risk for developing one of the FTR conditions is exacerbated by low prevalence of documented cases in available data. For example, an apparent under-reporting of deep venous thrombosis yields extremely low recorded prevalence of 0.4% among surgical and 0.5% among medical patients [1]. Relative sparseness of case data makes it difficult to use standard machine learning approaches to train predictive models that would yield operationally useful performance.

We present PRECORC (*P*rediction of *C*omorbid *R*are *C*onditions) a novel approach to predict the risk of FTR complications that leverages hierarchical structures of diagnosis



Diagnosis (DX)	251.2, V58.66, V58.69, V58.67
Diagnosis (DRG)	251.2
Procedure	99283
DoB	xx/xx/xxxx
Description	HYPOGLYCEMIA

Fig. 1: Fragment of a medical insurance claim record that includes Diagnostic and Procedure Codes, as well as Basic Demographic Information.

and procedure codes to alleviate the relatively low prevalence of specific types of such cases. It can be applied on top of previously learnt predictive models, and used to discover parts of the underlying hierarchies that support predictions. We model diagnosis and procedure hierarchies as well as simple demographics of patients in the form of a graph and perform soft label propagation to jointly leverage all available knowledge. In order to prime the label propagation algorithm, we apply prior risk score distributions that can be obtained with any standard machine learning classifier (in our experiments, we use Logistic Regression and Random Forest classifiers as examples). The resulting models outperform both the baseline classifiers and the label propagation method used with uninformed priors. We test these models using as examples one of the conditions directly represented among FTR risk factors, Venous Thrombo-Embolism (VTE), and one treatment scenario which often occurs in such context, Intubation and Mechanical Ventilation (IMV). In the following sections of this paper we briefly summarize prior work on incorporating hierarchies to improve predictions in healthcare applications, introduce our method, describe the data used and its featurization, explain the experimental settings, and summarize and discuss empirical results.

¹Corresponding Authors: {chiragn,mille856,awd}@cs.cmu.edu

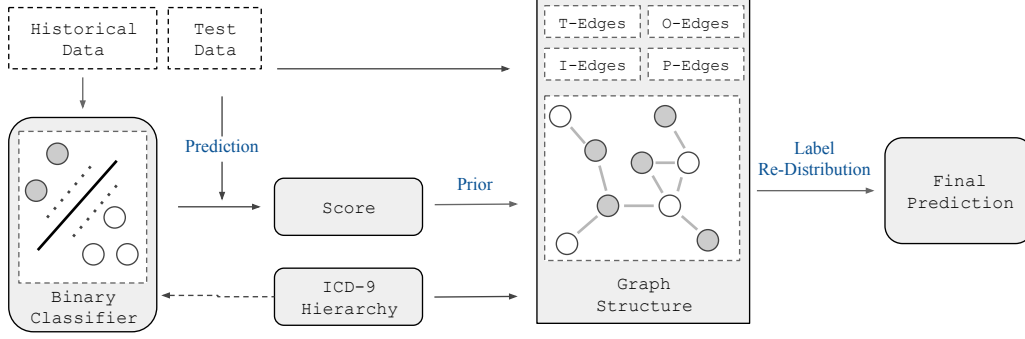


Fig. 2: Pipeline for our proposed approach. Historical Data is utilized to train standard classifiers, the performance of whom is boosted using the underlying medical ontology.

II. RELATED WORK

The ability to build models to predict the onset of rare conditions, or conditions that are hard to discern and that can escalate rapidly, using historical patient data, has been of recent interest to the machine learning in healthcare community [2], [3], [4], [5], [6], [7], [8], [9]. Notably, [10]’s work was the first step in the direction of exploiting the hierarchical structure of ICD-9 Diagnostic codes, where they incorporate the hierarchical structure explicitly by encoding it in their feature representations. On the other hand, [11] explored the hierarchy, while relying on smoothing the sparse feature representations extracted from the hierarchy by using co-occurrence between the various codes. In a more recent work, [12] investigated the use of medical domain corpora to train dense Word Vector representations, and have demonstrated significant improvements to performance of standard classifiers to predict disease onset, especially when ground truth data comes in short supply. These prior works rely on heuristics for informative feature representations that encode a hierarchy, and then proceed to train a common supervised machine learning models. This paper treats the problem in a fundamentally different manner. We propose to leverage hierarchical structures of medical context more directly by redistributing through them the label distributions represented by the training data. To the best of our knowledge, this is the first work that exploits an explicit graphical structure for the task.

III. PROPOSED APPROACH

Each patient’s hospital stay record includes the patients age, the set of diagnoses using the ICD-9 naming convention and the set of procedures performed. It also includes some basic demographic information about the patients race and gender. Each record is first featurized and a binary classifier is trained on this data to predict presence of a condition. The output of this binary classifier, along with the ICD-9 ontology and each patient’s record is then used to generate a graph like structure over which label propagation is performed, the output of which is treated as the final score for the propensity for a particular condition. Figure 2 is a broad representation of our pipelined system.

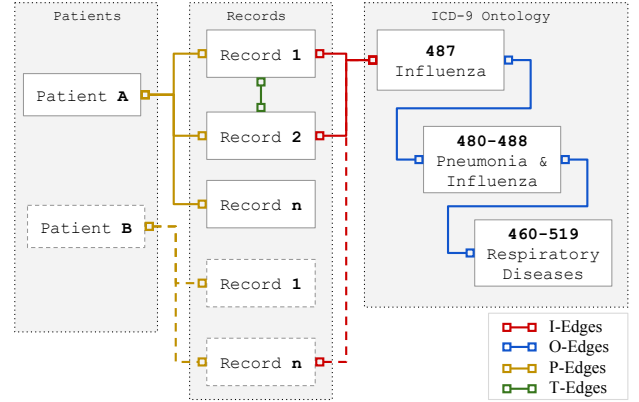


Fig. 3: Graph construction.

IV. GRAPH CONSTRUCTION

PRECORC is highly sensitive to the underlying structure of the graph. This section describes the graph construction using the nodes, along with the edges and their respective weightings. We will demonstrate the effect of different choices of graph structure with empirical results in the later sections. Figure 3 is a representative example of the graph construction methodology, that demonstrates the nodes and edges present between them.

Incidence(I-) Edges: Incidence edges describe relations between the diagnoses that were present in the record of the current hospitalization. Let us say that during the hospitalization the patient was diagnosed with ICD-9: 487.0 (PNEUMONIA WITH INFLUENZA). We add an edge between the patient’s current stay and the node representing the 487.0 code in the ICD-9 hierarchy graph. The edge weight is normalized by the count of any previous diagnoses with the same code, by dividing with the total number of distinct diagnoses in the current admission.

Ontology(O-) Edges: The ontology edges describe the ICD-9 hierarchy by forming a tree structure. Thus, for 487.0, we add nodes corresponding to 487 (INFLUENZA), 480-488 (PNEUMONIA & INFLUENZA) and 460-519

(DISEASES OF RESPIRATORY SYSTEM). Edges are added between each consecutive layer in the ICD-9 hierarchy with unit weight.¹ Ontology edges allow skewing of the label distribution in the branch that has propensity to be comorbid, and this distribution being shared by all leaves of the given branch, rather than just the diagnosis explicitly present in the current record.

Patient(P-) Edges: These edges are between each patient stay and a node representing the individual patient and are reflective of the normalized frequency of the records for each patient. The motivation for these edges is that they would allow label propagation to jointly leverage all historical records of the given patient. Often, a patient who has been diagnosed with a condition sometime in their history, would have a higher propensity for presenting with the similar condition in their future admissions, and the P-edges allow supporting such hypotheses.

Temporal(T-) Edges: T-edges are present between two subsequent patient stays and are weighed by the normalized time difference between each stay. They serve as an alternative to the Patient edges with a similar intended effect. Since T-edges are present only between subsequent records, their effect is more temporally localized as opposed to P-edges. Moreover, T-edges do not add explicit nodes to the graph structure to represent patients, thus using T-edges does not increase computational burdens much.

V. LABEL PROPAGATION

Consider the following energy function for label propagation over the graph labeling, commonly known as Harmonic Energy Minimization [13]. This function has a random walk interpretation, and along with Page Rank [14] it is extensively studied and applied to problems involving semi-supervised learning in structured data [15], [16].

$$E(f) = \sum_{i \in \mathcal{L}} (y_i - f_i)^2 D_{ii} + \lambda \sum_{i,j} (f_i - f_j)^2 A_{ij} \quad (1)$$

Here, \mathcal{L} and \mathcal{U} represent the set of labeled and unlabeled nodes respectively. A is the affinity matrix of the graph, and D is a diagonal matrix with the diagonal elements reflecting the degree of each node. y is the vector of labels for nodes in \mathcal{L} .

Label Propagation using the above Harmonic Energy Minimization formulation allows the model to only exploit information relating to ICD-9 codes, while our medical claims data has other features including Procedure Codes, Age, Gender, Diagnosis-Related Group (DRG) category, etc. Straightforward use of the above formulation does not allow label propagation to leverage this additional knowledge. Thus, as described by [17] we employ the soft label model,

¹While in practice, it can be argued that a weight corresponding to the prevalence of each sibling in the hierarchy would be a better choice, for simplicity we set it to 1.0 here.

which also includes a parameter enforcing a prior on each of the nodes. The modified energy function is thus given as:

$$E(f) = \sum_{i \in \mathcal{L}} (y_i - f_i)^2 D_{ii} + \lambda \left(w_0 \sum_{i \in \mathcal{U}} (f_i - \pi_i)^2 D_{ii} + \sum_{i,j} (f_i - f_j)^2 A_{ij} \right) \quad (2)$$

Which is minimized when, $\mathbf{f} = [\mathbf{D}(\mathbf{I} + \mathbf{S}) - \mathbf{A}]^{-1} \mathbf{D} \mathbf{S} \mathbf{y}$ (3)

$$\mathbf{S} = \begin{bmatrix} \frac{1}{\lambda} \mathbf{I}_{\mathcal{L}} & 0 \\ 0 & w_0 \mathbf{I}_{\mathcal{U}} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_{\mathcal{L}} \\ \boldsymbol{\pi} \end{bmatrix}, \quad \mathbf{D}_i = \sum_j \mathbf{A}_{ij}$$

In literature, such a prior is typically set to the estimated prevalence of the positive class in data. The nature of such prior makes it flexible, and we instead propose using the output of any standard binary classifier to set it. This design choice makes the maximum achievable performance of this approach, for at least some subset of values for the parameters, λ , w , lower-bounded by the performance of the originally trained binary classifier. We expect the proposed approach to perform at least as well as that initial model, given the right set of values for the parameters.

VI. DATA AND ITS FEATURIZATION

Dataset: The dataset for our experiments consists of a total of 222,202 de-identified admission records corresponding to 10,000 patients collected from a large hospital in Pittsburgh, PA, over a period of two and half years from January, 2014 to August, 2016. Each patient may have multiple such records, reflective of subsequent hospital admissions. For each of these records, we remove any information that implies or suggests diagnosis of VTE or IMV and treat any such record as a positive case for the two separate tasks, respectively all other records are treated as negative cases. We use thusly prepared data as our ground truth.

Featurization: For each of these admission records, we create a separate feature vector that encodes the diagnostic codes, the DRG codes, and the procedure codes corresponding to that admission. Each such feature vector is complemented with an equally-sized set of features which cumulatively reflect all prior admissions, by summing the counts of any features that have been previously encountered. Instead of using every code uniquely, which would result in highly sparse feature representations, we truncate each code to their first, most significant, three digits. We also include slow-changing and static information such as age and gender. Motivated by [10], we also add features from each code using “propagated binary” method. Note that while creating feature vectors, we do not include the diagnostic codes for IMV and VTE, since they are used to generate labels for the two targeted tasks, respectively.

VII. EXPERIMENTS

We empirically evaluate the performance of PRECORC on the two following prediction tasks using the data prepared

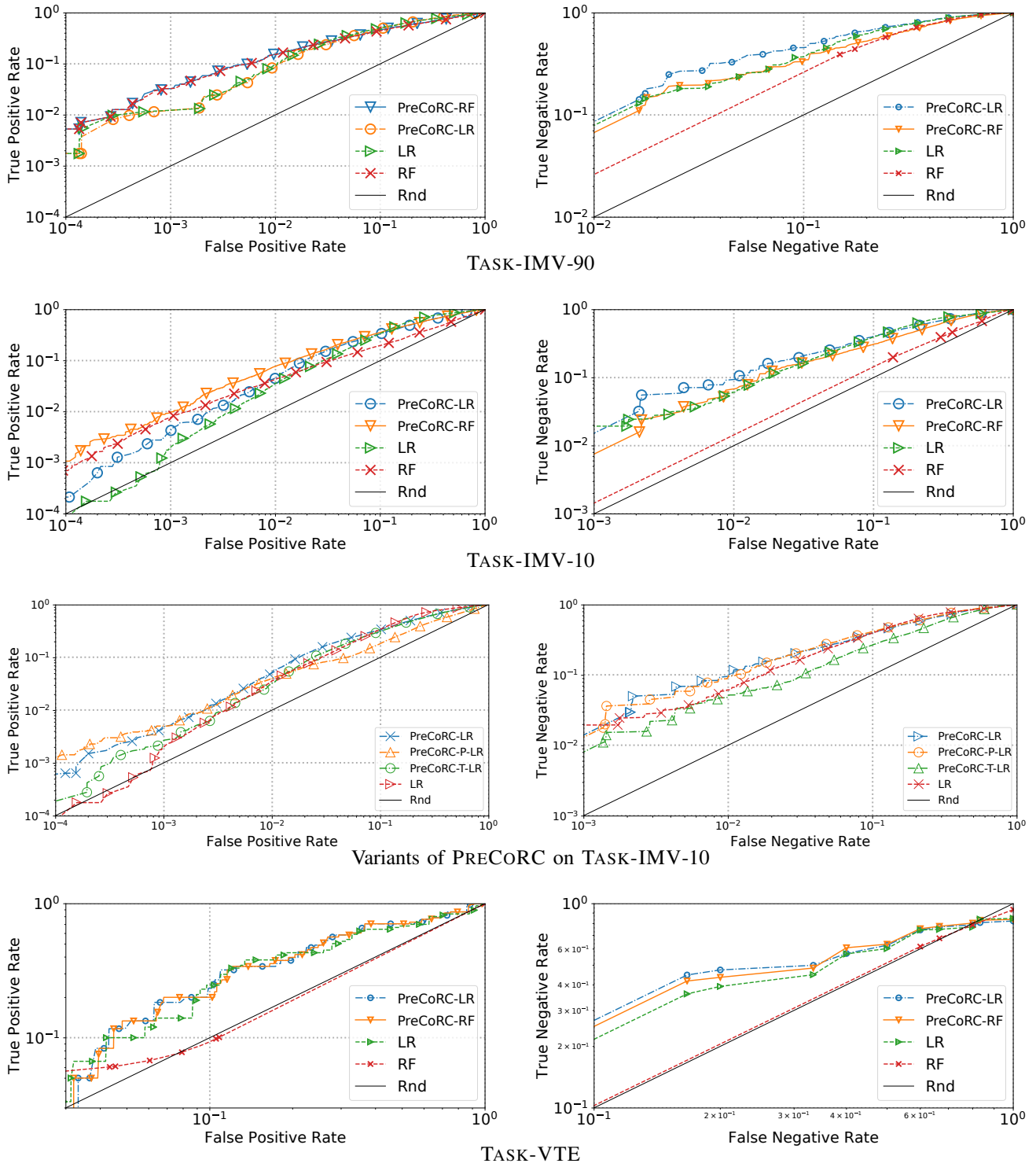


Fig. 4: ROC Plots corresponding to the various tasks described. Note that for PRECORC we report the result of the best performing variant.

	TPR@FPR=10 ⁻³	TPR@FPR=10 ⁻²	TPR@FPR=10 ⁻¹	TNR@FNR=10 ⁻²	TNR@FNR=5%	AUC
RF-BASELINE	0.0328	0.1525	0.4449	0.0230	0.1250	0.7410
RF-PCA	< 10 ⁻⁴	< 10 ⁻⁴	0.0088	0.0258	0.1289	0.6535
RF-NMF	< 10 ⁻⁴	0.0006	0.0239	0.0266	0.1329	0.7197
RF-PRECoRC	0.0369	0.1569	0.4682	0.1246	0.1246	0.7660
LR-BASELINE	0.0121	0.0924	0.5104	0.0705	0.2532	0.7947
LR-PCA	< 10 ⁻⁴	0.0005	0.0319	0.0897	0.2124	0.7942
LR-NMF	< 10 ⁻⁴	0.0005	0.0259	0.1065	0.2373	0.7670
LR-PRECoRC	0.0144	0.1100	0.5167	0.1478	0.3715	0.8211

TASK-IMV-90

	TPR@FPR=10 ⁻³	TPR@FPR=10 ⁻²	TPR@FPR=10 ⁻¹	TNR@FNR=10 ⁻²	TNR@FNR=5%	AUC
RF-BASELINE	0.0085	0.0454	0.2057	0.0118	0.0588	0.5598
RF-PCA	< 10 ⁻⁴	< 10 ⁻⁴	0.0037	0.0095	0.0474	0.4785
RF-NMF	< 10 ⁻⁴	≈ 10 ⁻⁴	0.0078	0.0090	0.0452	0.4816
RF-PRECoRC	0.0153	0.0833	0.3650	0.0630	0.2265	0.7277
LR-BASELINE	0.0020	0.0356	0.3367	0.0568	0.2405	0.7663
LR-PCA	< 10 ⁻⁴	< 10 ⁻⁴	0.0129	0.0634	0.2387	0.7551
LR-NMF	< 10 ⁻⁴	< 10 ⁻⁴	0.0121	0.0932	0.2750	0.7231
LR-PRECoRC	0.0048	0.0515	0.3951	0.1058	0.2637	0.7624

TASK-IMV-10

	TPR@FPR=10 ⁻³	TPR@FPR=10 ⁻²	TPR@FPR=10 ⁻¹	TNR@FNR=10 ⁻²	TNR@FNR=5%	AUC
RF-BASELINE	< 10 ⁻⁴	< 10 ⁻⁴	0.1247	0.0103	0.0514	0.5116
RF-PCA	< 10 ⁻⁴	< 10 ⁻⁴	0.0059	0.0107	0.0534	0.5224
RF-NMF	< 10 ⁻⁴	≈ 10 ⁻⁴	0.0096	0.0111	0.0554	0.5437
RF-PRECoRC	< 10 ⁻⁴	< 10 ⁻⁴	0.2500	0.1133	0.14393	0.6545
LR-BASELINE	< 10 ⁻⁴	0.0167	0.2428	0.0184	0.0809	0.6230
LR-PCA	< 10 ⁻⁴	< 10 ⁻⁴	0.0123	0.0184	0.0135	0.5224
LR-NMF	< 10 ⁻⁴	< 10 ⁻⁴	0.0055	0.0165	0.0202	0.4938
LR-PRECoRC	< 10 ⁻⁴	0.0167	0.2500	0.1553	0.2021	0.6663

TASK-VTE

TABLE I: Summary of experimental results. Note that for PRECoRC we report the results of the best performing variant. RF: Random Forest, LR: Logistic Regression, BASELINE: Raw Features.

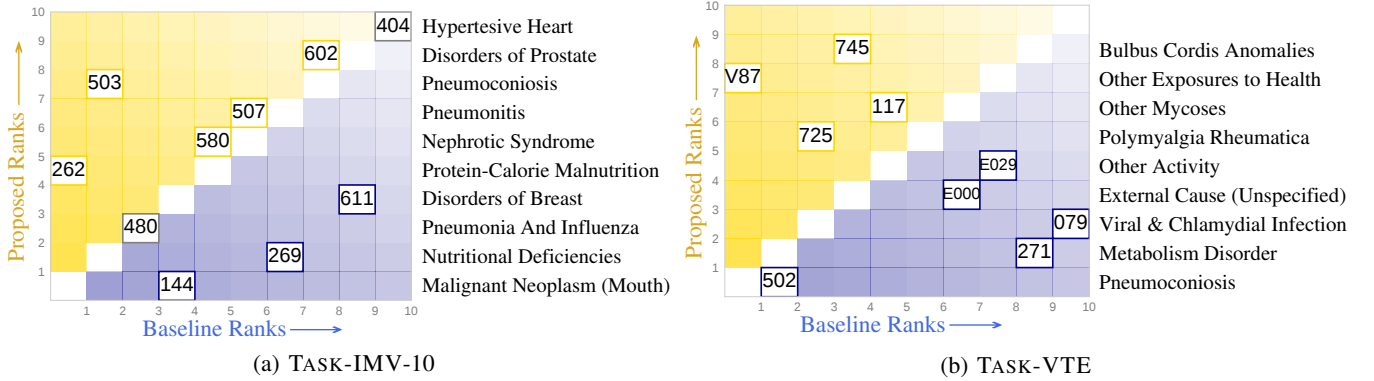


Fig. 5: Features re-weighted by the proposed approach.

CODE	DESCRIPTION
482	Bacterial Pneumonia
038	Streptococcal/Pneumococcal Septicemia
359	Muscular Dystrophy; Myopathy
238	Neoplasms; Myelodysplastic syndrome

TABLE II: Reweighted features for IMV.

CODE	DESCRIPTION
608	Seminal vesiculitis; Spermatocoele; Hematospermia
999	Infection due to central venous catheters
364	Iridocyclitis; Hyphema of iris; Iridoschisis
466	Acute bronchitis

TABLE III: Reweighted features for VTE.

as described above.

TASK-IMV: The first task is the prediction of whether the admitted patient will require Intubation and Mechanical Ventilation (IMV) during the current admission, given their past medical history. IMV has a reasonable prevalence in our dataset, out of 10,000 patients, we have 1,266 who required intubation at least once during their multiple hospital stays. Since we have multiple records corresponding to different hospital stays per patient, this corresponds to a prevalence of 1.173%. For TASK-IMV we perform two experiments, **TASK-IMV-10** that uses 10% of the labeled data as training data, and the rest as test. The other experiment, **TASK-IMV-90** is trained using 90% of the labeled training and tested on the remaining 10%. The rationale behind this is that by training on fewer data, IMV’s prevalence would resemble that of a rarer condition.

TASK-VTE: Our data set includes patients diagnosed with Pulmonary Embolism and/or Deep Vein Thrombosis, related conditions labeled together as VTE. A relative rarity of VTE in our data manifests with a high class imbalance with just 56 patients diagnosed with VTE yielding prevalence of 0.0519%. For TASK-VTE, we train our model with the entire labeled set and perform 10-fold cross validation.

Baselines: In order to demonstrate performance gains of our approach we first train two standard classifiers as baselines, a binary Logistic Regression model, with a ℓ_2 penalty on weights trained using [18], and a Random Forest ensemble with 100 component models using the package Scikit [19]. The choice of our baselines allows the comparison of simpler classifiers with linear decision functions, along with more powerful ensemble methods that often yield more complicated hypotheses spaces. We also train our model using label propagation, but without priming it with any classifier outputs. For label propagation, the algebraic multigrid method is leveraged to reduce the linear system described previously in Eq. 3 using the package PyAMG [20]².

Decomposition: Since our feature representations are sparse, for the Baseline classifiers, we also experimented with Principal Component Analysis [21], and Non-Negative Matrix Factorization [22], in order to extract denser, lower dimensional representations of features, in hopes to obtain more robust supervised models. Medical conditions and procedures have inclinations to be a correlated set of features and given enough data, we hypothesize that projecting it to a lower dimensional subspace by such a decomposition could remove spurious correlations while preserving the informative elements.

²Typically, the graph Laplacian matrix is sparse, and hence leveraging multigrid techniques can help improving the rate of convergence of an iterative linear system solver.

PRECORC: For the proposed approach, we exploit the same types of models as Baselines, i.e. Logistic Regression and Random Forest as the initial ‘seed classifiers’, before constructing the graphical structure to perform soft label propagation. It is worth noting that while training the seed classifiers, the classifier hyper-parameters are kept fixed to the same settings as used in Baselines, since tuning them would violate fairness of the experimental setup. Only the parameters λ and w_o for label propagation are tuned by performing grid search via cross-validation.

Variants: In order to appreciate the effect of various choices of the underlying graph architecture, we collect the results of PRECORC that include the Patient edges (PRECORC-P), the one that includes temporal edges (PRECORC-T), and PRECORC that just includes the Incidence and Ontology edges, and the results of the best performing variant are finally reported.

VIII. RESULTS

This section summarizes the results of our experiments. For both TASK-VTE and TASK-IMV we perform 10-fold cross validation. Unlike for IMV, for VTE, however, we do not create separate tasks for the full and under-sampled dataset since its prevalence is already very low. Tests of statistical significance are performed between the experiments and baselines using paired t-test, treating the AUC obtained from each split as an observation. We compare the performance of the classifiers for different True Positive Rates (TPR) at different thresholds of False Positive Rate (FPR), along with the True Negative Rates (TNR) at fixed False Negative Rates (FNR). Due to the rarity of the positive class instances in our data, we would want a classifier to perform especially well at the lower values of the False Positive Rate. We thus compare the TPR of the models at FPR corresponding to 10^{-3} , 10^{-2} and 10^{-1} , as well as the TNR at FNRs corresponding to 1% and 5%. Table I summarizes these scores for the various described tasks, while Figure 4 represents the ROC plots corresponding to the same.

TASK-IMV-90: PRECORC consistently outperformed the corresponding baselines in terms of all metrics, and the overall AUC for each task was statistically significantly superior with $p\text{-value} < 10^{-3}$. However, although there was some gain in TPR at lower FPRs, this gain was not very pronounced, and also not found to be significant. The gain in the TNR at fixed FNRs however, was greatly pronounced and also found to be statistically significant.

TASK-IMV-10: PRECORC significantly outperformed the corresponding TPR@FPR baselines, with a TPR greater by a factor of as much as 2 than that of the baselines. We further compared the statistical significance of the TPRs at lower thresholds, and found that these were significant with $p < 10^{-3}$ in all the cases. While the overall AUC with PRECORC-LR was lower than the LR-Baseline, this was not significant with a $p\text{-value} > 0.05$. LR-NMF

outperformed PRECORC at the TNR@FNR=5%, although this too was not found significant.

Variants: Figure 4 presents the performance of the variants of PRECORC trained with a Logistic Regression prior on TASK-IMV. PRECORC+P outperforms PRECORC in the high precision range, however the performance deteriorates at increasing FPR. PRECORC-T, although it still outperforms the Baseline Logistic Regression for the positive class, it performs worse with the negative class.

TASK-VTE: For VTE, our method did not show improvement over the TPR at the FPR=0.01 at higher FPRs, PRECORC outperformed baselines although we did not have strong p-values for these cases.³ We attribute this to the extremely low prevalence of the positive class.

In general, the gain in the True Negative Rate (TNR) at fixed False Negative Rates (FNR) was also pronounced in all the tasks, including VTE, which has very few positive instances. The results point to a significant boost in precision for both the positive and negative class, and motivates the application of our technique as a practical screening tool to identify individuals at either high or low risk, with a high confidence. However, for the negative class, models initialized with LR seem to outperform RF, which can be explained as a result of RFs tendency to favour the positive class.

IX. FEATURE ANALYSIS

For both tasks, we analyze which features are considered important by the baseline Logistic Regression and the proposed algorithm. For the logistic regression, we utilize the coefficient vector corresponding to the decision function learnt as an indicator for the feature relative importance, while for the proposed model we use the propagated value corresponding to the particular ICD-9 codes node after performing label propagation. We ignore diagnostic codes that have low per stay-record frequency in our dataset, as for the purposes of comparison, these are overly myopic and might not be generalizable enough for the classification task.⁴

Most of the top features for the proposed approach correspond to those also ranked high by the baseline, but with a few notable exceptions. Figures 5a and 5b present the overlap set of relative rankings of the importance of top 10 features used by the classifiers. Tables II & III report some canonical examples of features that were in the set of 30 most relevant features for the proposed approach but not in the top 100 most relevant features of the baseline. We hypothesize that these features may be held responsible for boosting classification performance of the proposed model when compared to the performance of the baselines.

³This is expected, since we have just 56 positive samples, performing 10 cross validation results in only 6 positive test cases in each fold, far too few to get appreciable p-values

⁴Nodes corresponding to these codes have much lower degree and are influenced by fewer neighbors, during label propagation.

Diagnoses such as Pneumonia & Muscular Dystrophy are ranked higher by PRECORC, supporting the intuition that patients with such histories would be prone to the need of life support in the form of intubation or mechanical ventilation, as corroborated in clinical research [23]. Perhaps more intriguing are the features corresponding to VTE. For instance, Iridocyclitis, a manifestation of Behçet’s Syndrome which has been identified to have a high tendency to cause Deep Vein Thrombosis [24] has been reranked high by our approach. It was also able to highlight ICD-9 ‘999’, (Infections due to Central Venous Catheters), which is a VTE Risk Factor [25], [26].

Post-hoc Interpretability: PRECORC can be thought of as a post-hoc interpretable model as described by [27] for the underlying binary classifier. Most post-hoc interpretable models are aimed at creating a simpler model to imitate the performance of a larger, more complicated, black-box-like model, perhaps at the expense of lower performance. It is worth noting that PRECORC generates an interpretable pipeline, while preserving, if not boosting, performance. In our case, the label propagation model is created over a feature subset that is more naturally represented by a graphical structure, as opposed to the feature vector representation exploited by the underlying ‘seed’ classifier.

X. DISCUSSION & CONCLUSION

We proposed an approach to boost the performance of a medical complication co-morbidity classifier by leveraging the medical diagnostic code hierarchy, along with the patient profiles, in a graph structure. We hypothesize that using this graph structure to perform label propagation, induces a similarity function over the diagnostic codes, that helps augment the knowledge available from standard binary classifiers. The increase in performance, especially in the high precision range for both the positive and negative class, offers potential practical utility. PRECORC can be effectively employed as a screening tool at both the low FPR and low FNR range. For other thresholds, the users can still default back to the pre-existing standard screening protocol. In the absence of any labeled data, the proposed graphical structure can help the user identify potential cases of interest through model-driven search, in an active-learning/active-search like setting.

In its current form, our approach has certain limitations. Firstly, grid search over the label propagation parameters can be expensive. Secondly, the graphical structure for label propagation, as presented, assumes a rather simplistic featurization, especially in the context of the temporal nature of the underlying data. Better graph construction methods would need to be employed in order for the model to reflect such temporal context and underlying knowledge. While our approach makes the model interpretable to an extent, further research needs to be carried out as to how to best present the label distributions over the ontology as an explanation to the end user: in our case, the medical practitioner.

Our results indicate that it is possible to predict the co-occurrence of rare conditions in patients based upon

prior data. Applied operationally in clinical settings, risk for developing FTR conditions can potentially be assessed earlier enabling preventative treatment, decreasing mortality, improving patient outcomes, reducing costs of care, and improving hospital's quality ratings. We expect that further improvements in the classification capability can be attained when adding information from patients electronic health records collected upon admission and throughout their hospital stay. Data sources could include medical histories, in particular pre-existing conditions, demographics, and accruing information from laboratory tests, imaging diagnostic tests, and evolving vital signs whenever they become available. If successful, such forecasting tools could help change the paradigm of current medical care from reactive to proactive. That is the overall goal of our future work.

ACKNOWLEDGMENTS

This work has been partially supported by NIH under award R01NR013912 and by DARPA under awards FA8750-12-2-0324 and FA8750-17-2-0130.

REFERENCES

- [1] J. Needleman, P. Buerhaus, S. Matke, M. Stewart, and K. Zelevinsky, "Nurse-staffing levels and the quality of care in hospitals," *New England Journal of Medicine*, vol. 346, no. 22, pp. 1715–1722, 2002.
- [2] J. Chen and A. Dubrawski, "Learning to extract actionable evidence from medical insurance claims data," in *Actionable Intelligence in Healthcare*, J. Liebowitz and A. Dawson, Eds. Taylor and Francis, 2017, ch. 12.
- [3] J. Sun, J. Hu, D. Luo, M. Markatou, F. Wang, S. Ebadollahi, Z. Daar, and W. F. Stewart, "Combining knowledge and data driven insights for identifying risk factors using electronic health records," in *AMIA*, vol. 2012, 2012, pp. 901–10.
- [4] K. Ng, A. Ghoting, S. R. Steinhubl, W. F. Stewart, B. Malin, and J. Sun, "Paramo: A parallel predictive modeling platform for healthcare analytic research using electronic health records," *Journal of biomedical informatics*, vol. 48, pp. 160–170, 2014.
- [5] J. Sun, C. D. McNaughton, P. Zhang, A. Perer, A. Gkoulalas-Divanis, J. C. Denny, J. Kirby, T. Lasko, A. Saip, and B. A. Malin, "Predicting changes in hypertension control using electronic health records from a chronic disease management program," *Journal of the American Medical Informatics Association*, vol. 21, no. 2, pp. 337–344, 2014.
- [6] J. Wiens, E. Horvitz, and 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. 467–475.
- [7] M. R. Pinsky and A. Dubrawski, "Gleaning knowledge from data in the intensive care unit," *American journal of respiratory and critical care medicine*, vol. 190, no. 6, pp. 606–610, 2014.
- [8] L. Chen, O. Ogundele, G. Clermont, M. Hravnak, M. R. Pinsky, and A. W. Dubrawski, "Dynamic and personalized risk forecast in step-down units. implications for monitoring paradigms," *Annals of the American Thoracic Society*, vol. 14, no. 3, pp. 384–391, 2017.
- [9] M. Guillaume-Bert, A. Dubrawski, D. Wang, M. Hravnak, G. Clermont, and M. R. Pinsky, "Learning temporal rules to forecast instability in continuously monitored patients," *Journal of the American Medical Informatics Association*, p. ocw048, 2016.
- [10] A. Singh, G. Nadkarni, J. Guttag, and E. Bottinger, "Leveraging hierarchy in medical codes for predictive modeling," in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2014, pp. 96–103.
- [11] M. Vukicevic, S. Radovanovic, A. Kovacevic, G. Stiglic, and Z. Obradovic, "Improving hospital readmission prediction using domain knowledge based virtual examples," in *International Conference on Knowledge Management in Organizations*. Springer, 2015, pp. 695–706.
- [12] Y. Liu, C. Stultz, J. Guttag, K.-T. Chuang, F.-W. Liang, and H.-J. Su, "Transferring knowledge from text to predict disease onset," in *Proceedings of the 1st Machine Learning for Healthcare Conference*, 2016, pp. 150–163.
- [13] X. Zhu, Z. Ghahramani, J. Lafferty *et al.*, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, vol. 3, 2003, pp. 912–919.
- [14] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford InfoLab, Tech. Rep., 1999.
- [15] L. Bing, S. Chaudhari, R. C. Wang, and W. W. Cohen, "Improving distant supervision for information extraction using label propagation through lists," in *EMNLP*, 2015, pp. 524–529.
- [16] L. Bing, M. Ling, R. C. Wang, and W. W. Cohen, "Distant ie by bootstrapping using lists and document structure," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
- [17] X. Wang, R. Garnett, and J. Schneider, "Active search on graphs," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 731–738.
- [18] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.
- [19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.
- [20] W. N. Bell, L. N. Olson, and J. B. Schroder, "PyAMG: Algebraic multigrid solvers in Python v3.0," 2015, release 3.2. [Online]. Available: <https://github.com/pyamg/pyamg>
- [21] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.
- [22] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001.
- [23] A. K. Simonds, "Respiratory complications of the muscular dystrophies," in *Seminars in respiratory and critical care medicine*, vol. 23, no. 03. Copyright© 2002 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA., 2002, pp. 231–238.
- [24] M. Houman, I. B. Ghorbel, I. K. B. Salah, M. Lamloum, M. B. Ahmed, and M. Miled, "Deep vein thrombosis in behcet's disease," *Clinical and experimental rheumatology*, vol. 19, no. 5; SUPP/24, pp. S–48, 2001.
- [25] C. Rooden, M. E. Tesselaar, S. Osanto, F. R. Rosendaal, and M. V. Huisman, "Deep vein thrombosis associated with central venous catheters—a review," *Journal of Thrombosis and Haemostasis*, vol. 3, no. 11, pp. 2409–2419, 2005.
- [26] J. L. Baskin, C.-H. Pui, U. Reiss, J. A. Wilimas, M. L. Metzger, R. C. Ribeiro, and S. C. Howard, "Management of occlusion and thrombosis associated with long-term indwelling central venous catheters," *The Lancet*, vol. 374, no. 9684, pp. 159–169, 2009.
- [27] Z. C. Lipton, "The mythos of model interpretability," *arXiv preprint arXiv:1606.03490*, 2016.