How Good Are Provider Annotations?: A Machine Learning Approach

M. Said Malas[1, 3], MD, Suranga Kasthurirathne[2, 3], BEng, Sharon Moe[1, 4], MD, Jon Duke[3], MD MS

[1]Indiana University School of Medicine, Indianapolis, IN; [2]Indiana University-Purdue University School of Informatics and Computing, Indianapolis, IN; [3]Regenstrief Institute, Indianapolis, IN; [4]Roudebush Veterans Administration Medical Center, Indianapolis, IN

Abstract

Providers' annotations are often used as classifiers for supervised machine learning. Occasionally, annotations of patient status are 'naturally occurring' in clinical documents, such as the morbidities assessment of patients starting dialysis. We aimed to examine the predictability of provider annotations for 8 clinical conditions. We retrieved the reported status (positive/negative) for these conditions from existing clinical documents for a cohort of dialysis patients at Indiana University. We used all available procedure, billing, laboratory, and prescription data to generate predictive models of physician annotations. The best performing algorithms yielded precision and recall metrics ranging from a low of 0.44 and 0.37 for heart failure and a high of 0.86 and 0.71 for cancer. We concluded that the relatively poor prediction of provider annotations points towards heterogeneous and inconsistent annotation behavior. A thorough assessment of provider accuracy should be done prior to using annotations generated during routine clinical care as gold-standard outcomes.

Introduction

A significant challenge in the domain of predictive modeling is obtaining annotated datasets. Such annotations are necessary to classify, and subsequently predict, outcomes of interest using supervised machine learning approaches. Commonly, annotations are generated by manual expert review when the outcome is not readily classifiable from available clinical documentation. Such expert review can be both time-consuming and resource-intensive, and often comprises the bulk of effort in a predictive modeling study. While automated algorithms can be used to perform classification tasks, review and interpretation by an expert is generally considered the gold standard.

Such expert annotations are sometimes collected as part of routine clinical care. For example, sources for such annotations include disability forms, registration forms, and death certificates which contain physician documentation and interpretation of patient conditions and other outcomes. The use of such 'naturally occurring' annotations in the clinical record is a convenient source of classifiers for predictive modeling. However, the quality of these annotations may be influenced by numerous factors, including but are not limited to: 1) complexity of logic required for the annotation 2) cognitive and behavioral biases and heuristics 3) availability and access to necessary data 4) time constraints. Thus, provider annotation performance is itself variable. The objective of this study is to utilize machine learning techniques to assess the predictability of providers' annotation performance for a set of common clinical conditions. Specifically, we looked at nephrologists' documentation of co-morbidities in patients with end-stage renal disease.

Materials and Methods

Patient Cohort

Our cohort included all patients who started chronic dialysis therapy between 2005 through 2014 at Indiana University outpatient dialysis centers. This initial cohort consisted of 296 patients. Approvals were obtained from the Indiana University Institutional Review Board and the relevant institutions participating in the Indiana Network

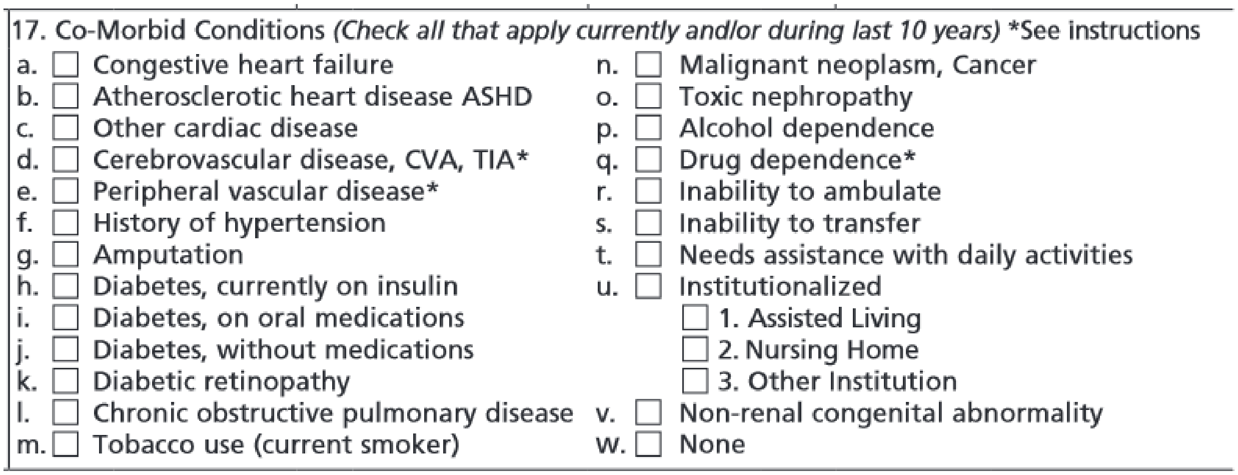
This is the author's manuscript of the article published in final edited form as:
Malas, M. S., Wish, J., Moorthi, R., Grannis, S., Dexter, P., Duke, J., & Moe, S. (2017). A comparison between physicians and computer algorithms for form CMS-2728 data reporting. Hemodialysis International. https://doi.org/10.1111/hdi.12445

for Patient Care (INPC) [1], a health information exchange that aggregates clinical data from all major health systems in Indianapolis, Indiana.

Data extraction

We extracted electronic medical record data (billing and procedure codes, medications, and laboratories) for all patients in the cohort for up to 10 years prior to the onset of End Stage Renal Disease (ESRD). We extracted these data using the INPC.

For the provider annotations, we obtained the ESRD registration forms required by the Center for Medicare and Medicaid Services (CMS) for all new-onset dialysis patients (Figure 1). The forms capture information on disease status (present/absent) for 23 conditions. Completion of these forms is part of the routine clinical care for all new ESRD patients [2]. Each disease state has a corresponding checkbox to indicate the presence of disease, i.e. positive or negative outcome.



17. Co-Morbid Conditions *(Check all that apply currently and/or during last 10 years)* *See instructions

a. □ Congestive heart failure
b. □ Atherosclerotic heart disease ASHD
c. □ Other cardiac disease
d. □ Cerebrovascular disease, CVA, TIA*
e. □ Peripheral vascular disease*
f. □ History of hypertension
g. □ Amputation
h. □ Diabetes, currently on insulin
i. □ Diabetes, on oral medications
j. □ Diabetes, without medications
k. □ Diabetic retinopathy
l. □ Chronic obstructive pulmonary disease
m. □ Tobacco use (current smoker)
n. □ Malignant neoplasm, Cancer
o. □ Toxic nephropathy
p. □ Alcohol dependence
q. □ Drug dependence*
r. □ Inability to ambulate
s. □ Inability to transfer
t. □ Needs assistance with daily activities
u. □ Institutionalized
□ 1. Assisted Living
□ 2. Nursing Home
□ 3. Other Institution
v. □ Non-renal congenital abnormality
w. □ None

Figure 1: The comorbid conditions section of the end stage renal disease patient registration form.

Data preprocessing

We integrated datasets extracted from multiple sources into a single master dataset for analysis. Data sets included each patient's procedure and billing codes, medications, and laboratory data. To these data, we added as outcome measures 8 clinical conditions documented by the nephrologist on each patient's registration form: coronary artery disease, cancer, congestive heart failure, cerebrovascular accident, diabetes mellitus, peripheral vascular disease, chronic obstructive pulmonary disease, and retinopathy. All data analysis, pre-processing and decision model building were performed using version 3.7.12 of the Waikato Environment for Knowledge Analysis (Weka) software [3].

A preliminary analysis of the dataset indicated that many of the outcomes were unbalanced, with the number of negative outcomes for each condition greatly outweighing the number of positive outcomes. We also noted that a number of patients in the dataset had no reported outcomes (not checked yes or no). To ensure better data quality, we excluded those patients from the dataset. This removed 20 patients from the initial cohort leaving 276 patients for our final analysis. We pre-processed the procedure and billing codes, prescription data into binary variables. Unlike the other datasets, the laboratory data were in numeric format. An initial data analysis indicated that these numeric values were highly distributed, and therefore, may reflect negatively on the decision model building. To address this challenge, the laboratory data was discretized into automatically-generated categorical ranges using Weka's built-in discretization support.

To address the unbalanced nature of the master dataset, we decided to perform Synthetic Minority Oversampling Technique (SMOTE) to balance the dataset under evaluation. SMOTE boosted the negative outcomes in the dataset using synthetic data, further improving the quality of the decision model generation.

Machine learning approach

We sought to determine the ability to predict each outcome using the master dataset. We hypothesized that varying (a) feature subset sizes, (b) classification algorithms and (c) boosting percentages would yield varying performance metrics. To test our hypothesis, we built multiple decision models using varying combinations of the abovementioned criteria. The classification algorithms selected for our study were simple logistic regression (SLR), naïve Bayes (NB), random forest (RF), and J48 decision tree (J48). These algorithms were selected based on their widespread use in various matching learning studies, and track record of yielding optimal results [4, 5].

The master dataset had a total of 7655 features. Existing literature indicates that using irrelevant features for decision model building may lead to over-fitting [6]. To prevent this, we limited our feature space by ranking each feature in order of significance using Weka's information gain approach, based on the Kullback-Leibler divergence [7], a widely used method for selecting optimal features for machine learning. From the ranked feature set, we selected feature subsets of 50, 75, 100 and 125 for study. These feature subset sizes were selected based on preliminary analysis of what subset sizes provided the best results for the dataset, coupled with previous literature that recommended smaller feature subset sizes for use [8-10]. Due to the limited dataset, we adopted ten-fold cross validation, aka rotation estimation, a widely used train/test method prescribed for use with relatively small datasets [11].

We built and tested decision models using combinations of the aforementioned feature subset sizes, classification algorithms and boosting percentages (Figure 2). Given 4 feature subset sizes, 4 algorithms, and 3 boosting percentages, this resulted in a total of 48 (4 x 4 x 3) decision models for each outcome being tested.

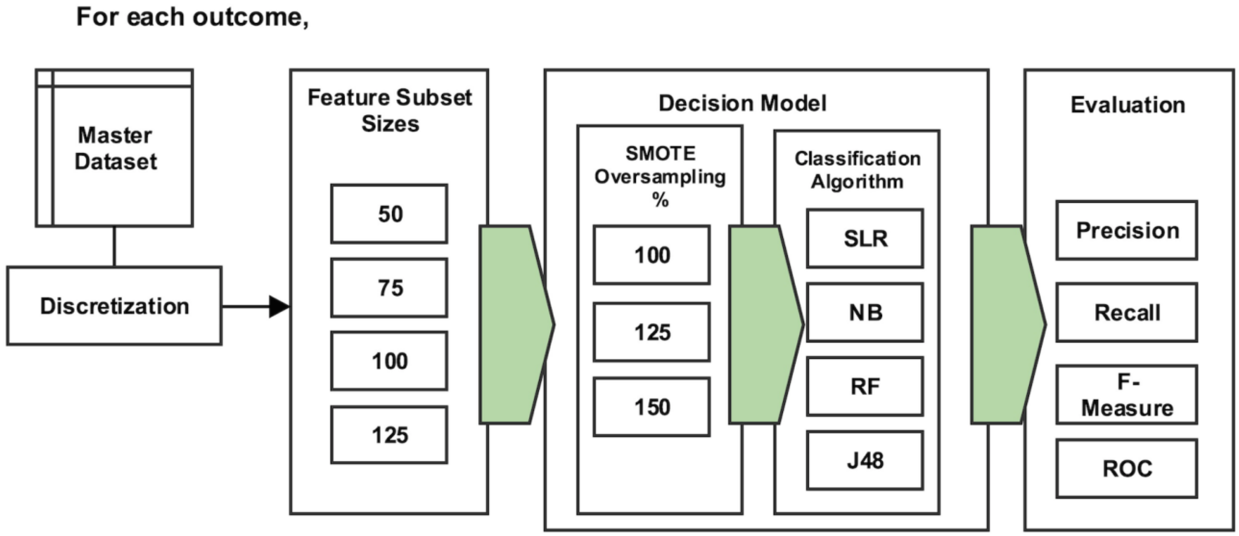


Figure 2. The study approach from the selection of alternative feature subsets to decision model building and the evaluation of results. Synthetic Minority Oversampling Technique (SMOTE), simple logistic regression (SLR), naïve Bayes (NB), random forest (RF), and J48 decision tree (J48), receiver operating characteristic curve (ROC).

Results

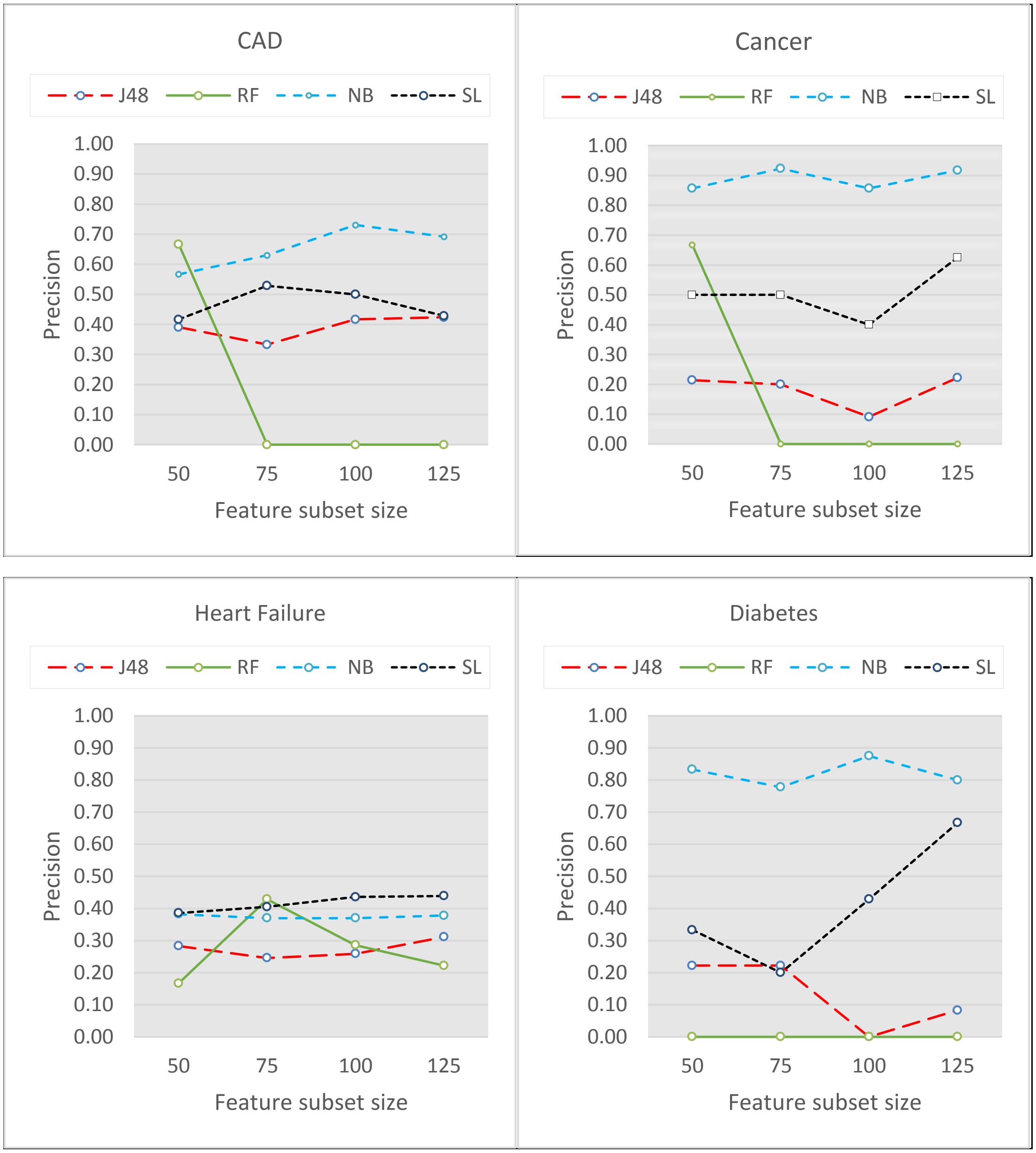
The overall predictive accuracy of providers’ positive annotations was poor to fair, with an average recall of 0.6 and no F-measure exceeding 0.8. Results of our supervised machine learning are summarized in Table 1, which shows the best performing learning algorithm for each condition’s annotations, along with the corresponding accuracy metrics of precision, recall, F –measure, and area under the receiver operating characteristic curve (ROC).

Outcome	Algorithm	Class	Precision	Recall	F-measure	ROC
CAD	NB	Positive	0.731	0.679	0.704	0.923
		Negative	0.964	0.972	0.968	
Cancer	NB	Positive	0.857	0.706	0.774	0.93
		Negative	0.981	0.992	0.987	
CHF	SL	Positive	0.436	0.37	0.4	0.755
		Negative	0.878	0.904	0.891	
CVA	NB	Positive	0.619	0.722	0.667	0.952
		Negative	0.98	0.969	0.975	
DM	NB	Positive	0.875	0.368	0.519	0.927
		Negative	0.955	0.996	0.975	
PVD	NB	Positive	0.769	0.526	0.625	0.945
		Negative	0.966	0.988	0.977	
COPD	NB	Positive	0.667	0.667	0.667	0.928
		Negative	0.981	0.981	0.981	
Retinopathy	NB	Positive	0.354	0.81	0.493	0.907
		Negative	0.982	0.878	0.928	

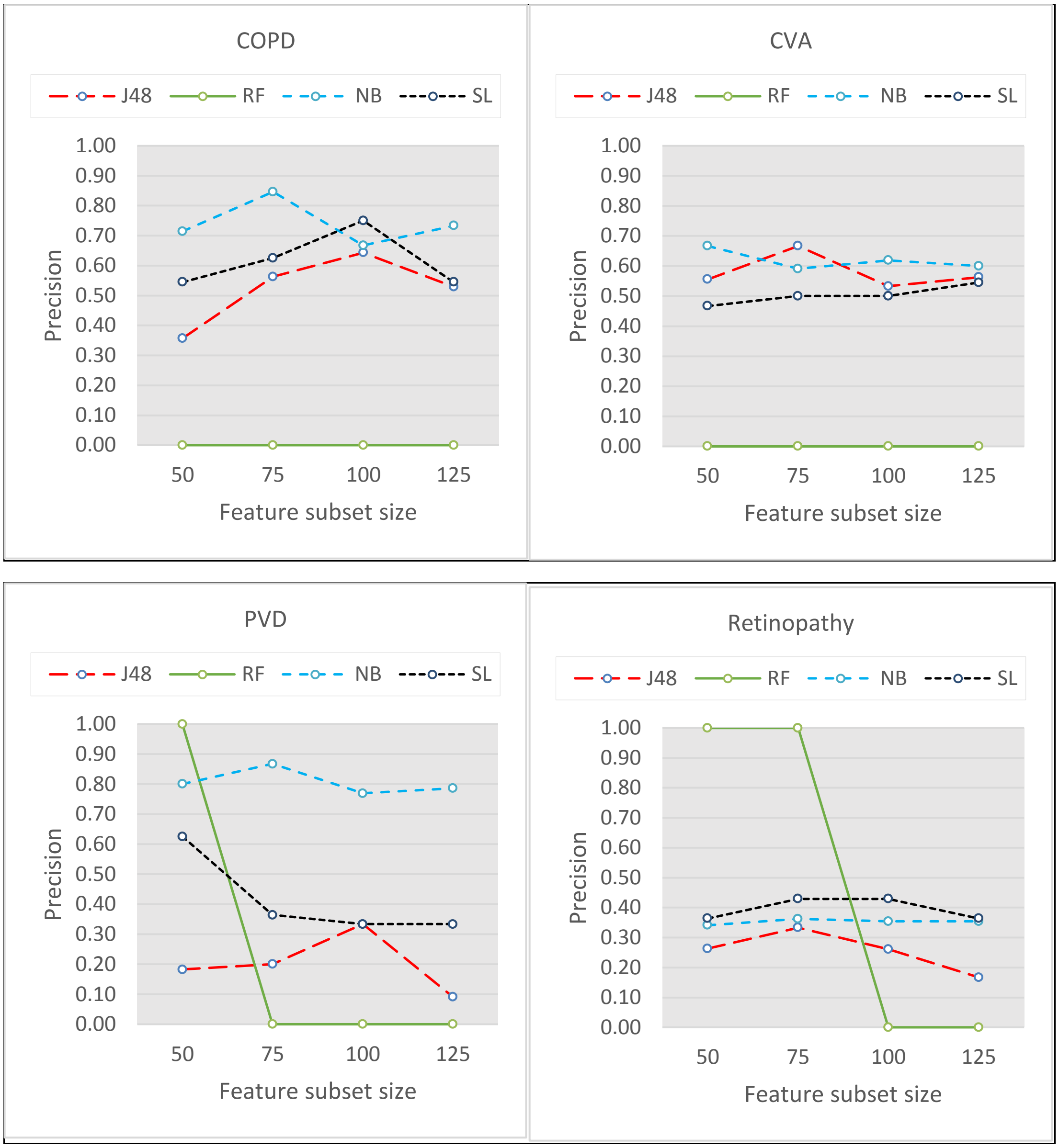
Coronary artery disease (CAD), congestive heart failure (CHF), cerebrovascular accident (CVA), diabetes mellitus (DM), hypertension (HTN), peripheral vascular disease (PVD), chronic obstructive pulmonary disease (COPD), simple logistic regression (SL), naïve Bayes (NB), receiver operating characteristic curve (ROC).

Table 1: Performance metrics of the best performing algorithm for providers' annotations of medical conditions.

An exceedingly high overall predictive accuracy (above 0.9) was achieved for the negative annotations and subsequently the ROC, which can be explained by the high prevalence of negative annotations in the data. In Figure 3, we show the estimated precision for each positively annotated outcome using different decision models and different numbers of selected features. Similarly, the estimated recall for each positively annotated outcome using different numbers of selected features is shown in Figure 4. In some cases, the predictive accuracy was variable when comparing between different decision models.

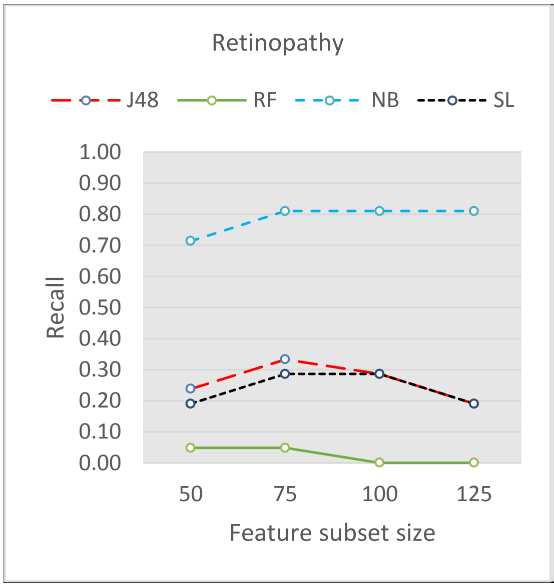

CAD
Cancer
Heart Failure
Diabetes
J48
RF
NB
SL
Precision
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
50
75
100
125
Feature subset size

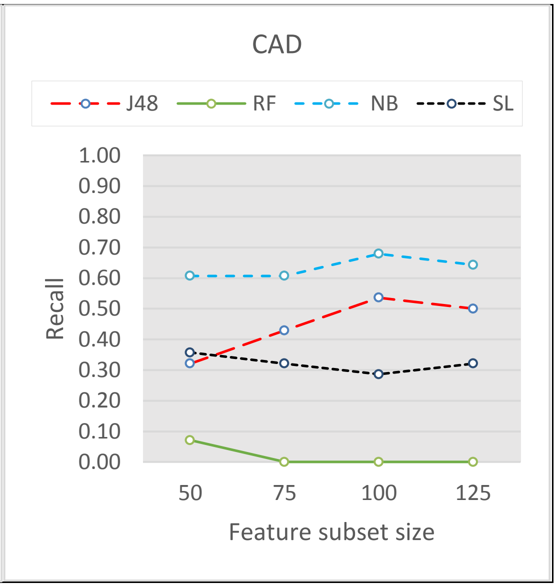
“Continued”

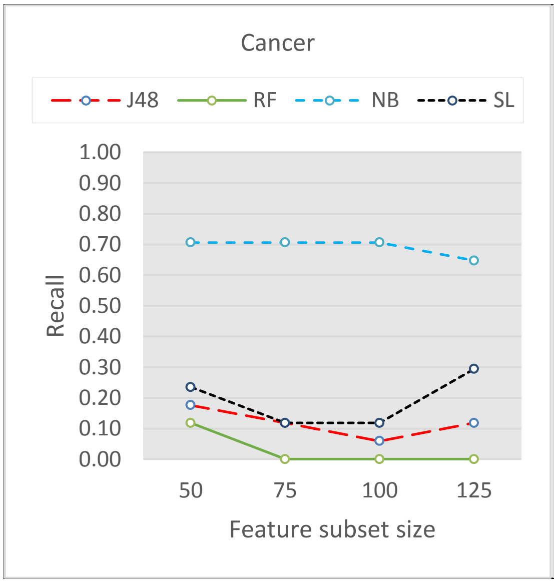


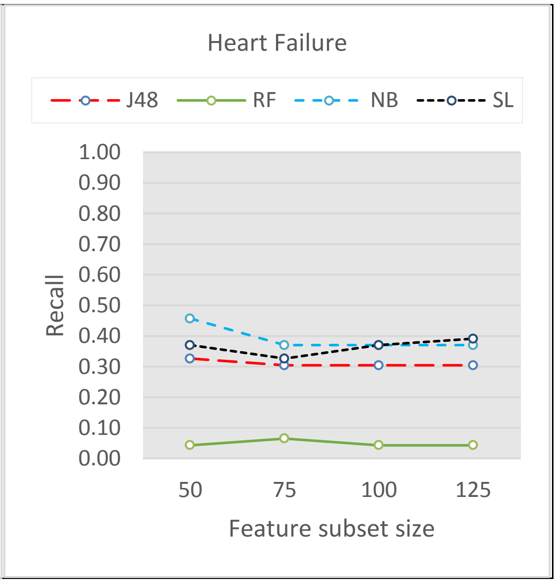
Coronary artery disease (CAD), congestive heart failure (CHF), cerebrovascular accident (CVA), diabetes mellitus (DM), hypertension (HTN), peripheral vascular disease (PVD), chronic obstructive pulmonary disease (COPD), simple logistic regression (SL), naïve Bayes (NB), random forest (RF), J48 trees (J48).

Figure 3: Estimated precision across each positively annotated outcome graphed by varying feature subset size for each decision model.

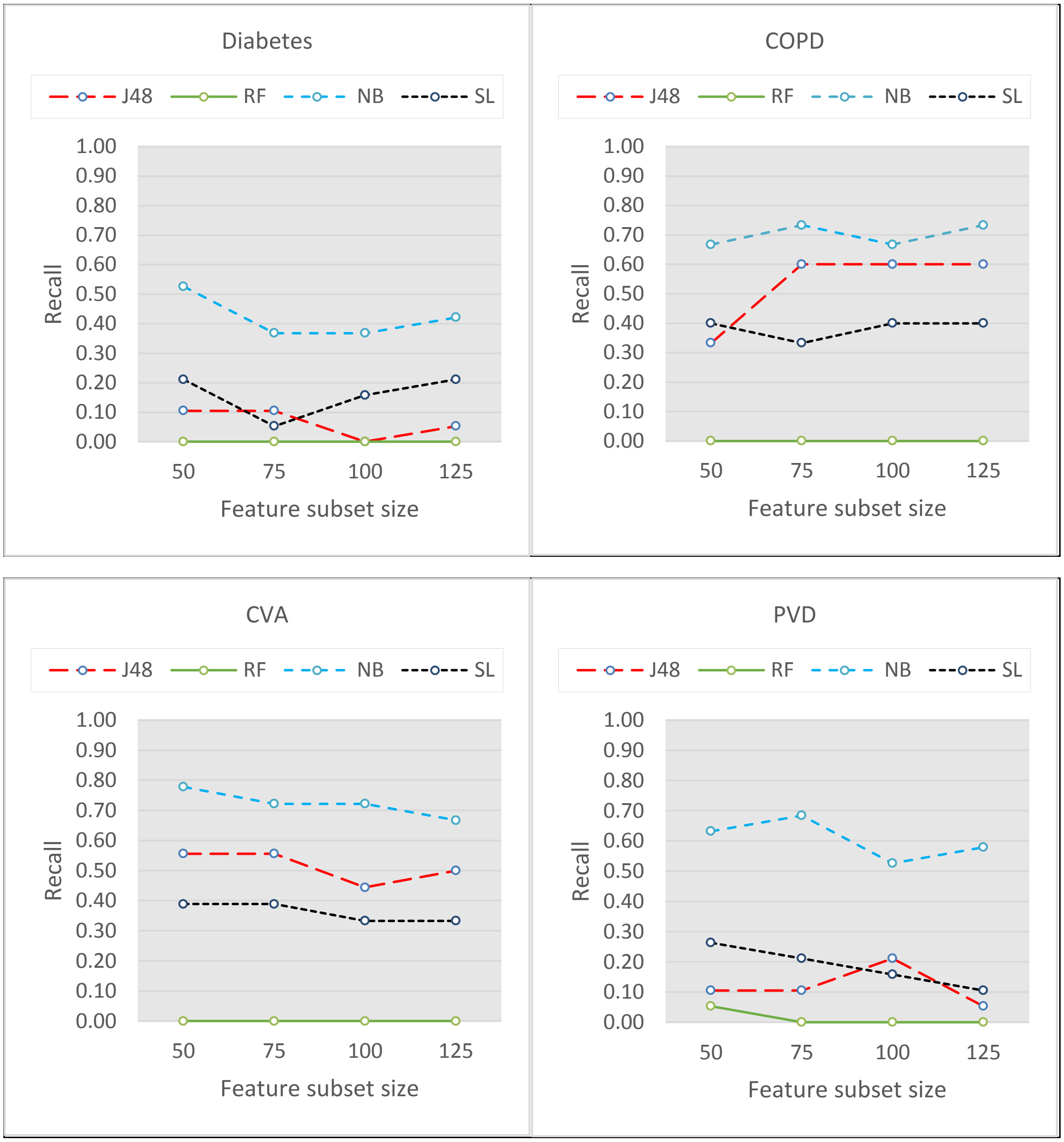

Retinopathy
J48
RF
NB
SL
Recall
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
50
75
100
125
Feature subset size


CAD
J48
RF
NB
SL
Recall
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
50
75
100
125
Feature subset size


Cancer
J48
RF
NB
SL
Recall
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
50
75
100
125
Feature subset size


Heart Failure
J48
RF
NB
SL
Recall
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
50
75
100
125
Feature subset size

"Continued"



Coronary artery disease (CAD), congestive heart failure (CHF), cerebrovascular accident (CVA), diabetes mellitus (DM), hypertension (HTN), peripheral vascular disease (PVD), chronic obstructive pulmonary disease (COPD), simple logistic regression (SL), naïve Bayes (NB), random forest (RF), J48 trees (J48).

Figure 4: Estimated recall across each positively annotated outcome graphed by varying feature subset size for each decision model.

Discussion

Extensive literature exists on the utility of machine learning in predicting clinically relevant outcomes. Literature supports that many conditions such as heart failure [12, 13], diabetes [14], and cancer [8] can be predicted with high accuracy. However, in our study results show fair ability, at best, to predict these provider-reported clinical conditions in our cohort. This indicates wide heterogeneity in how providers annotate these clinical outcomes which is suggestive of a mismatch between annotations and the true state of these clinical conditions. This is further supported by our previous findings where expert-designed phenotypes outperformed providers' annotations to determine the true state of clinical conditions for this cohort [15].

The precision and recall measures for the best performing positive annotation prediction model were as low as 0.44 and 0.37 respectively for heart failure and as high as 0.86 and 0.71 for cancer. This indicates an overall fair prediction for positive annotations. The notable performance differences across the different decision models in many instances such as the precision for cancer and the recall for retinopathy, makes even this fair prediction likely to be an overestimate.

The main strength of our study is the presence of pre-existing annotations that allowed us to approach the behavioral aspect of providers' annotation as the main target. This is in contrast to the usual prediction of the true states of diseases that are more common in the literature and where annotations are created by the research team. We believe that the degree of prediction for un-intervened annotations such as those extracted from routine clinical care have important value for studying the behavioral aspects of providers' decision making process. For example, highly predictable annotations suggest a systematic annotation behavior by providers, regardless whether correct or not for the true disease state, whereas low predictability indicates a more heterogeneous behavior. Such information can be used for further research and in the development of Clinical Decision Support (CDS) systems that target providers' decision making process. The use of machine learning to understand physician annotation behavior has not been previously studied, and, in our opinion, deserves further exploration by the informatics research community as it impacts the rigor of data used for other purposes. For the meantime, providers' annotations as a proxy for the true state of clinical conditions should be utilized with caution.

There are several limitations to this study. First, this is a defined cohort with a sample of only 276 patients. Although this represents a 10 year longitudinal cohort, the size of the cohort is a limiting factor and therefore our findings may not be generalizable. Second, data mining is subject to the known nuances of missing, corrupted, inconsistent, or non-standardized data [16]. However, this was reduced with the use of broad access and standardized retrieval via the Indiana Network for Patient Care network. Further, we adopted a series of best practices advocated in machine learning literature as solutions to pitfalls caused by missing, corrupted, inconsistent, unbalanced or non-standardized data. This means that despite of these limitations, we are following the best possible solutions to build decision models using the data at hand.

Conclusions

We cautiously conclude that providers' annotations of key clinical outcomes are limited based on the lack of their predictability. We suspect that this limitation is likely present in other clinical settings, though this needs further exploration. We suggest that the use of providers' annotations extracted from clinical documents be carefully examined and appropriately challenged for clinical data analytics research. Finally, the use of machine learning techniques can be a valuable tool in understanding annotations patterns in order to guide development of decision support systems.

Disclosures

Dr. Malas is funded by a grant under the Merck-Regenstrief Program in Personalized Health Care Research and Innovation, a collaboration between Merck, Sharp, & Dhome and the Regenstrief Institute, and the Regenstrief-Merck Scholar's Award in Pharmacoepidemiology and Informatics. The authors declare that there are no conflicts of interests.

References

1. Indiana network for patient care. Available from: http://www.ihie.org/indiana-network-for-patient-care.
2. End stage renal disease medical evidence report - medicare entitlement and/or patient registration. Available from: https://www.cms.gov/Medicare/CMS-Forms/CMS-Forms/downloads/cms2728.pdf.
3. Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using weka. Bioinformatics. 2004;20(15):2479-81.

4. Selker HP, Griffith JL, Patil S, Long WJ, D'Agostino RB. A comparison of performance of mathematical predictive methods for medical diagnosis: Identifying acute cardiac ischemia among emergency department patients. J Investig Med. 1995;43(5):468-76.
5. Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W. Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. Ann Behav Med.26(3):172-81.
6. Hawkins DM. The problem of overfitting. J Chem Inf Comput Sci. 2004;44(1):1-12.
7. Polani D. Kullback-leibler divergence. Encyclopedia of systems biology: Springer; 2013. p. 1087-8.
8. Kasthurirathne SN, Dixon BE, Gichoya J, Xu H, Xia Y, Mamlin B, et al. Toward better public health reporting using existing off the shelf approaches: A comparison of alternative cancer detection approaches using plaintext medical data and non-dictionary based feature selection. Journal of biomedical informatics. 2016;60:145-52.
9. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157-82.
10. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507-17.
11. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics surveys. 2010;4:40-79.
12. Wu J, Roy J, Stewart WF. Prediction modeling using ehr data: Challenges, strategies, and a comparison of machine learning approaches. Med Care. 2010;48(6):S106-S13.
13. Rosenman M, He J, Martin J, Nutakki K, Eckert G, Lane K, et al. Database queries for hospitalizations for acute congestive heart failure: Flexible methods and validation based on set theory. J Am Med Inform Assoc. 2014;21(2):345-52.
14. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 2010;10(1):16.
15. Malas MS, Wish J, Moorthi R, Grannis S, Dexter P, Duke J, et al. A comparison between physicians and computer algorithms for form cms-2728 data reporting. Forthcoming in 2016.
16. Koh HC, Tan G. Data mining applications in healthcare. J Healthc Inf Manag. 2011;19(2):65.