Predicting Health Care Utilization After Behavioral Health Referral Using Natural Language Processing and Machine Learning

Nathaniel Roysden[1], Adam Wright, Ph.D[1,2]
[1]Harvard Medical School, Boston, MA; [2]Brigham and Women's Hospital, Boston, MA

Abstract

Mental health problems are an independent predictor of increased healthcare utilization. We created random forest classifiers for predicting two outcomes following a patient's first behavioral health encounter: decreased utilization by any amount (AUROC 0.74) and ultra-high absolute utilization (AUROC 0.88). These models may be used for clinical decision support by referring providers, to automatically detect patients who may benefit from referral, for cost management, or for risk/protection factor analysis.

Introduction

Optimization of cost while preserving or improving quality of care is a central focus of national health reform. One major predictor of high cost of treatment is mental health comorbidity—one study found that, on average, patients with a mental health diagnosis were 2.2 times more expensive than patients without a mental health diagnosis.[1] In addition to mental health diagnoses, there are other behavioral health issues that may not result in a diagnosis, but still predict increased rates of health care utilization.[2] This increase in utilization is theorized to be due to a decreased ability to care for self and high rates of unnecessary care.[3] In addition to cost considerations, the additional health care utilization associated with behavioral health conditions may also pose the health risk of unintended negative treatment outcomes from unnecessary care.[4]

Given the potential for significant improvement of health status and reduction in health costs, patients who may benefit from behavioral health referral are an important cohort for identification. Similar to prediction of readmission, prediction of high care utilization post-referral may signal reconsideration of treatment planning, while prediction of decreased utilization may strengthen the decision to refer. A very high specificity model may even be used to automatically prompt referral for certain patients. Accordingly, the goal of this project is to be able to create a model which can accurately predict the health care cost outcomes of behavioral health referrals.

Machine Learning (sometimes also called "Data Mining") and Natural Language Processing ("NLP") methods have previously been used on Electronic Medical Record ("EMR") and billing data for both predictive and text-classification purposes. While machine learning models excel at determining the content of medical notes,[5,6] their application to prediction of clinical outcomes has varied. One study achieved good results predicting future morbidity of admitted cases of suspected sepsis using a neural network.[7] Another study used a regression tree model to predict post-hospitalization suicides of Veterans Administration patients, again with good results.[8] With regard to more specific clinical outcomes, yet another study employed machine learning models to predict cardiac arrests after an emergency department electrocardiogram.[9] Other common models used in descriptive and predictive machine learning include support vector machines and random forests. He et al. used data mining methods to predict early hospital readmissions.[10] The He et. al. paper is most comparable to this paper, due to its broad cohort and highly variable outcome. Their study differed in its focus on "feature selection and exclusive use of administrative data for ease of portability and understanding," with use of a model that was suited to that particular approach.

We aim to use similar machine learning approaches to predict the outcomes of a patient's first encounter with behavioral health providers ("BH patients"). The studied outcomes are cost-related in both a relative and an absolute sense. The first outcome is decreased health care utilization by any amount following a BH encounter. The second outcome is extremely high post-BH utilization, defined as utilization in the 95th or 99th percentile of non-BH patients. These predictions will be made using random forest classifiers trained on a robust clinical data set, including structured administrative data, free text from provider notes, and lab data.

Methods

Data sources

The Partners Research Patient Data Registry ("RPDR")[11] is an integrated data warehouse with the ability to query an array of discrete databases storing Partners electronic health data from its various member institutions. For this study, the RPDR was queried for patients who were seen at least once in a Brigham and Women's Hospital ("BWH") primary

care practice. This query returned records for approximately 221,000 patients. Those records contained EMR free text notes, procedure codes (billing), encounters (scheduling), problem lists, diagnoses, lab results for A1c and Cholesterol, medication lists, and demographics. The Partners IRB approved this study.

A second set of data was downloaded from the online Medicare Physician Fee Schedule Value Files on an annual basis from 1999 to present and used to estimate standard payments for billed services.[12]

Tools

The Anaconda[13] distribution of Python 3.4 was used with notable packages including the Python Natural Language Toolkit ("NLTK"), Matplotlib, Numpy, Scipy, and Scikit-Learn ("sklearn"), including the Scikit-Learn[14] Random Forest Classifier ("RFC")[15] implementation. Other models from sklearn that were tested but not further described in this paper include AdaBoost Classifier and Random Forest Regressor.

Intervention

Clinic location was selected as the study intervention over psychiatric diagnoses because it represents the treatment benefit provided by any behavioral health services, even for patients who do not have a defined diagnosis.[16] This intervention was recorded as the date of the earliest psychiatry, psychology, or behavioral health clinic encounter from the encounters dataset, based on the listed clinic location for each date of service.

Sample / Cohort Selection

Only patients with at least one behavioral health clinic visit after 2005 were included in the study cohort, reducing from 221,000 patients with at least one visit to a BWH primary care provider to 37,000 patients. Furthermore, only patients with at least one procedure code between 120 and 365 days before and after the first BH visit were included, resulting in a final cohort of 12,759 patients.

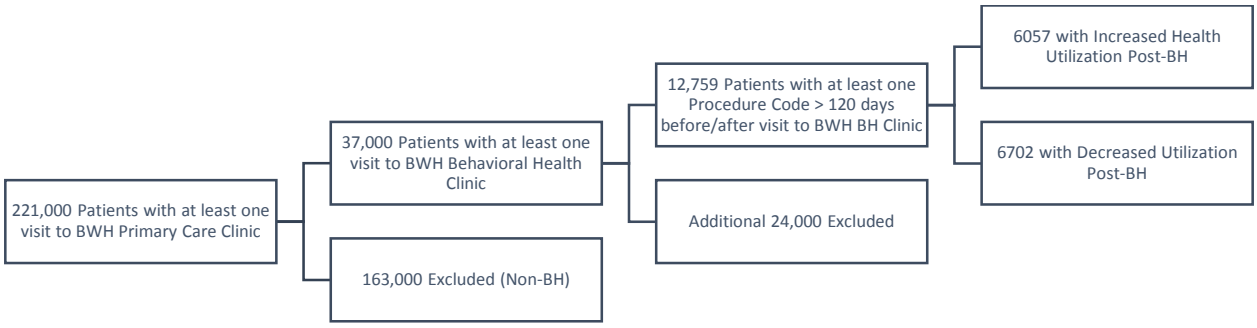

Figure 1: Cohort Selection Strategy

Health Care Utilization

The procedures (billing) dataset was converted from procedure codes to their most recent available Relative Value Unit ("RVU") value via the Physician Fee Schedule files. These RVU values were paired with their corresponding dates for each patient. All entries prior to 2005 were removed from the data set, because there was a significant change in the billing data from 1995 to 2005—an exponential increase in apparent billing that likely represents changes in hospital electronic record keeping. This significantly biased early model attempts, by creating a set of patients who almost all showed increases in health care costs between 1995 and 2005. Billing data from 2005-2014 remained stable.

Health care utilization was calculated for procedure codes in the year prior and the year following the first behavioral health visit ("$date_{BH}$") for each patient in the BH group, as shown in (Equation 1). Billing codes from more than one year prior or more than one year after the first BH visit were ignored. Patients with no billing codes more than 120 days before or without codes more than 120 days after their first BH visit were also removed. This 120 day requirement was intended to ensure that calculated values represented enduring trends in terms of usage before and after the intervention, not singular health interactions causing apparent health utilization to be inflated by a small denominator. Monthly utilization and three-month windowed utilization were each calculated in similar fashion.

Equation 1: $$Usage = \frac{\sum_{i=0}^{n} RVU_i}{|date_n - date_{BH}|} \quad where\ date\_n\ is\ the\ date\ that\ maximizes\ the\ denominator$$

To ensure that utilization calculations were not affected by significantly different denominators, the summary statistics for the "delta days" term were calculated, as shown in (Table 1) below.

Table 1: Number of days in RVU per time calculation, "BH" behavioral health, "SD" standard deviation.

Group	Mean	SD	Median
Before BH	302.5	± 66.0	330.0
After BH	304.1	± 65.7	331.0
Comparison	289.5	± 68.4	311.0

Health utilization was calculated for patients without any visits to a behavioral health provider (heretofore "Comparison Group") using the procedure codes within one year prior to the most recent recorded procedure code, also excluding patients without at least one procedure code occurring more than 120 days before the most recent procedure code. These values were used solely for reasonable cutoff values for classification of BH patients and therefore the slight difference in recorded days between BH and comparison patients does not affect the study outcome. (Table 2) below shows health care utilization in terms of RVU per day over a range of percentiles. These comparison group percentile cutoffs were then used to calculate static cutoffs for absolute utilization by the BH group. In other words, future categorization of BH patient utilization as "above the 99th percentile" means above the 99th percentile of the comparison group and does not reflect the proportion of BH patients falling into that group.

Table 2: Comparison Group Utilization Range

Percentile	RVU per day
50	0.125
80	0.379
90	0.733
95	1.251
99	2.666

Feature Extraction

The frequency of health contacts was calculated as the number of unique dates of service divided by the number of days between the most remote and most recent visit in the year preceding the first behavioral health visit.

The diagnoses, labs, medications, and procedures data sets each include a date with every entry; entries following the first BH visit were discarded. The demographics data set returns values accurate to its date of access; entries were recorded for patients as of the date that particular subset of the data was returned i.e. Oct-Nov, 2014, with each possible value comprising a separate binary feature in the model. Two labs, A1c and Cholesterol, were included as features according to their most recent lab value prior to behavioral health visit. Text results like "cancelled" were entered as 0. Results satisfying the regular expression for ">14.0" were entered as 14.0. Medications were included as discrete features according to their "medication code", a structured identifier of the medication without dosage information; multiple occurrences were counted. Diagnosis and procedure codes were recorded with each unique item entered as a separate feature; their values were entered as a count of the number of individual times they were encountered in each patient's data set.

Finally, each patient's notes—prior to their first behavioral health visit—were concatenated and then extracted into frequency distributions (counts) of tokens, bigrams, and trigrams after removal of the NLTK English "stopwords"[17] set, with the exception of negating terms ("no, nor").

Cohort Evaluation

(Figure 2) below shows a 40x40 bin heat map of health resource utilization in RVU/day before and after the first BH clinic encounter. The likelihood of a patient having decreased health resource utilization following their first BH encounter, by any amount, was 52.8%. The plot is log scaled by number of patients in each category as specified in the color bar on the side of the image, white areas represent no data.

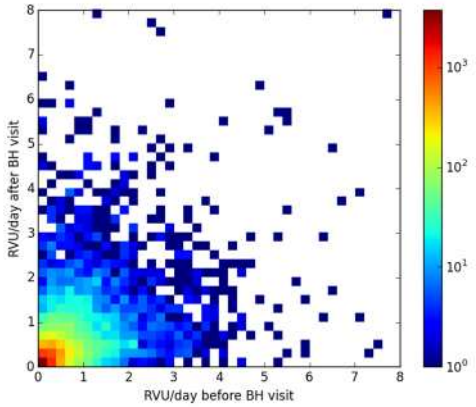

Figure 2: Heat map of patient utilization before/after their first behavioral health encounter

(Table 3) below displays the same information but with a coarse and nonlinear binning. Notably, while patients in the <50th percentile and 50-80 percentile groups most often remained in their cohorts, 95th and 99th percentile patients typically fell to lower percentile groups—a minority remained persistently high utilizers. This analysis also revealed a fair comparison for the "above x percentile" models—the baseline assumption that patients will continue using the same amount of health care before and after their first BH encounter.

Table 3: Percentile groups before/after BH encounter

		After						
Before	**Percentile**	**<50**	**50-80**	**80-90**	**90-95**	**95-99**	**>99**	**Total Before:**
	<50	2063	1215	278	73	25	8	**3662**
	50-80	1251	2185	743	197	86	25	**4487**
	80-90	302	919	639	291	128	35	**2314**
	90-95	97	324	333	230	179	49	**1212**
	95-99	26	140	228	191	181	73	**839**
	>99	5	28	43	54	73	42	**245**
	Total After:	**3744**	**4811**	**2264**	**1036**	**672**	**232**	**12759**

The overall effect of behavioral health visits on patient costs was quantified by converting the total amount of RVU's counted in the before- and after- BH groups using $35.8 per RVU. At that reimbursement rate, the pre-BH group cost $78.6 Million per year, while the post-BH group cost $73.3 Million. Also of note, the 2% of pre-referral patients who were in the 99th percentile category accounted for 16% of annual pre-BH cohort costs; the 2% of patients in the post-referral group whose utilization was above the 99th percentile accounted for 17% of post-BH costs. The BH group falling into non-BH percentiles 0-80 together accounted for 24% of spending before and 28% of spending after the first BH encounter, while comprising 67% and 71% of the BH patient population, respectively.

Categorization

Three category vectors were created and used as outcomes for prediction in our models. The positive categories for each vector were: 1. Post-BH utilization greater than 99th percentile. 2. Post-BH utilization greater than 95th percentile. 3. Post-BH utilization below pre-BH utilization by any amount.

Feature Selection

The two calculated features, utilization and visit frequency, and all of the structured data were exempted from feature selection and normalization. Throughout the feature selection process, each selection method was trained on the training data and then applied to both the training and test data.

Before feature selection, token vectors, bigram vectors, and trigram vectors were separately normalized along the sample axis. Next, n-grams that occurred in fewer than 3 or more than *(n_samples-3)* samples were removed from the feature set using a variance selection method. This reduced the feature space from over 17 million features to approximately 1 million. Finally, one sklearn *f_classif* selector, which chooses features based on their Anova F-value, was trained to select the best 100,000 features ("above percentile" model) or 20,000 features ("decreased utilization" model) from the total set of n-grams. Approximately 163,000 (83,000) features were included in each model.

Model Creation

Structured grid search was performed on a randomized test/train set to find the optimal settings and number of included n-grams for each model, detailed below. Two metrics were used to optimize grid search: Matthews Correlation Coefficient ("MCC") (Equation 2) and F1 Score ("F1") (Equation 3). The MCC is a metric that balances all four quadrants of the confusion matrix and is a conservative optimization approach. The F1 score reflects performance of the precision-recall curve and is less conservative than the MCC but more conservative than the Area Under the Receiver Operating Characteristic curve (AUROC).

$$\textit{Equation 2: } MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

TP, TN, FP, and FN are, respectively, True Positive, True Negative, False Positive, and False Negative

$$\textit{Equation 3: } F_1 = 2 * \frac{Precision*Recall}{Precision+Recall}$$

The number of trees in each forest was partially limited by the need to balance the demands of runtime and development time, especially with classifiers trained using the *max_features = None* argument. The number of included features was similarly limited by runtime and available memory: only about 200,000 maximum features could fit in memory when stored as a dense array.

Above Percentile Classifiers: Each Random Forest Classifier was trained using 5000 trees with samples weighted according to the inverse of class frequency and otherwise default settings.[18]

Increased/Decreased Utilization Classifier: 100 trees were calculated with balanced sample weights and the *max_features* parameter set to *None*, meaning that each tree would consider all possible features to determine optimal split.

Cross Validation

Stratified three-fold cross validation was performed to train and test each classifier. Feature selection methods were trained on each separate training fold and applied only to each corresponding test fold. Similarly, each classifier was created on the train fold and applied to the test fold. No information or model parameters were shared between folds of the cross-validation process. Each fold of the cross-validation required approximately 20 minutes for the "Above Percentile" classifier and approximately 40 minutes for the "Increased/Decreased" classifier. Runtime for an individual sample (patient) on a trained, loaded classifier is in the sub-seconds range.

Results

Predicting Decreased Utilization

(Figure 3) below shows the Receiver Operating Characteristic ("ROC") and Precision-Recall ("PR") curves corresponding to the classifier's ability to predict that a patient's utilization will decrease after their first BH clinic visit.

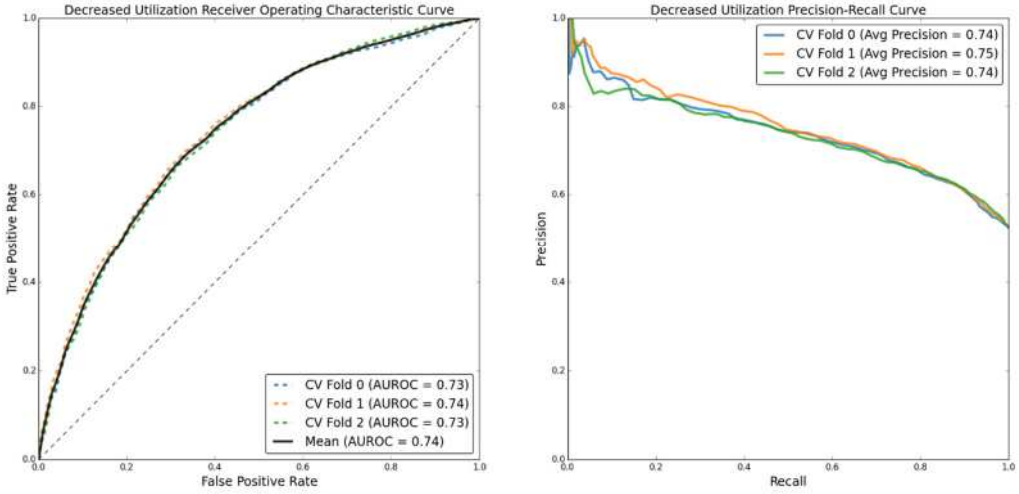

Figure 3: ROC and PR curves for predicting decreased utilization after BH encounter.

(Table 4) below details classifier performance metrics for each fold of the cross validation. The "Optimization" label details which performance metric was optimized to set the threshold value used in the classifier: the remaining metrics are reported at that given classifier confidence threshold. Both the Matthews Correlation Coefficient ("MCC") and, separately, the F1 score ("F1"; also called "F-measure") were calculated for each fold of the cross validation. With a mean Area Under the Receiver Operating Curve ("AUROC") of 0.74 and average precision—the precision-recall curve correlate to AUROC—of 0.74-0.75, the model performs much better than random chance.

Table 4: Predicting decreased utilization after BH encounter. MCC and F1 optimized performance metrics.

Fold	Optimization	MCC	F1	Precision	Recall	TPR	FPR	Threshold
1	MCC	0.36	0.70	0.69	0.72	0.72	0.36	0.52
	F1	0.33	0.73	0.62	0.88	0.88	0.59	0.38
2	MCC	0.37	0.72	0.67	0.77	0.77	0.42	0.47
	F1	0.34	0.73	0.62	0.87	0.87	0.58	0.36
3	MCC	0.35	0.72	0.65	0.82	0.82	0.49	0.45
	F1	0.33	0.73	0.62	0.89	0.89	0.61	0.37
Mean	MCC	0.36	0.71	0.67	0.77	0.77	0.42	0.48
	F1	0.33	0.73	0.62	0.88	0.88	0.59	0.37

Prediction of Utilization Above the 99th and 95th Percentiles

(Figure 4) shows ROC and PR curves for predicting post-BH patient utilization above the 99th percentile, with corresponding performance metrics in (Table 5). (Figure 5) and (Table 6) detail the same information for prediction of utilization above the 95th percentile.

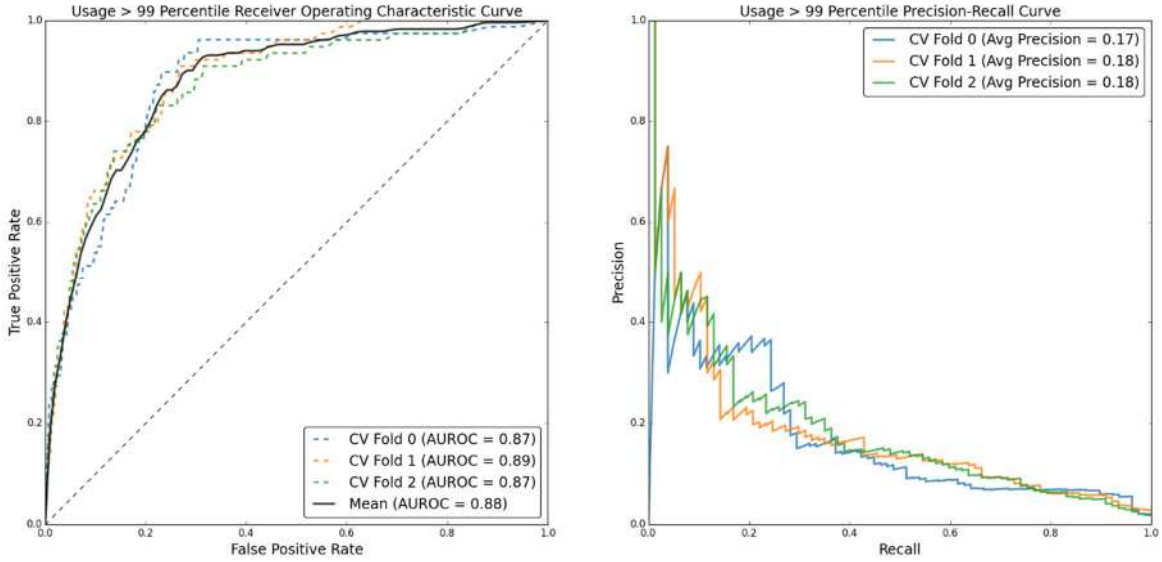

Figure 4: ROC and PR curves for predicting post-BH utilization greater than the non-BH 99th percentile.

Table 5: Predicting utilization above the 99th percentile. MCC and F1 optimized performance metrics

Fold	Optimization	MCC	F1	Precision	Recall	TPR	FPR	Threshold
1	MCC	0.288	0.292	0.358	0.244	0.244	0.008	0.133
	F1	0.288	0.292	0.358	0.244	0.244	0.008	0.133
2	MCC	0.252	0.205	0.121	0.636	0.636	0.085	0.039
	F1	0.251	0.245	0.170	0.429	0.429	0.039	0.058
3	MCC	0.260	0.273	0.240	0.312	0.312	0.018	0.116
	F1	0.260	0.273	0.240	0.312	0.312	0.018	0.116
Mean	MCC	0.266	0.257	0.240	0.397	0.397	0.037	0.096
	F1	0.266	0.270	0.256	0.328	0.328	0.022	0.102

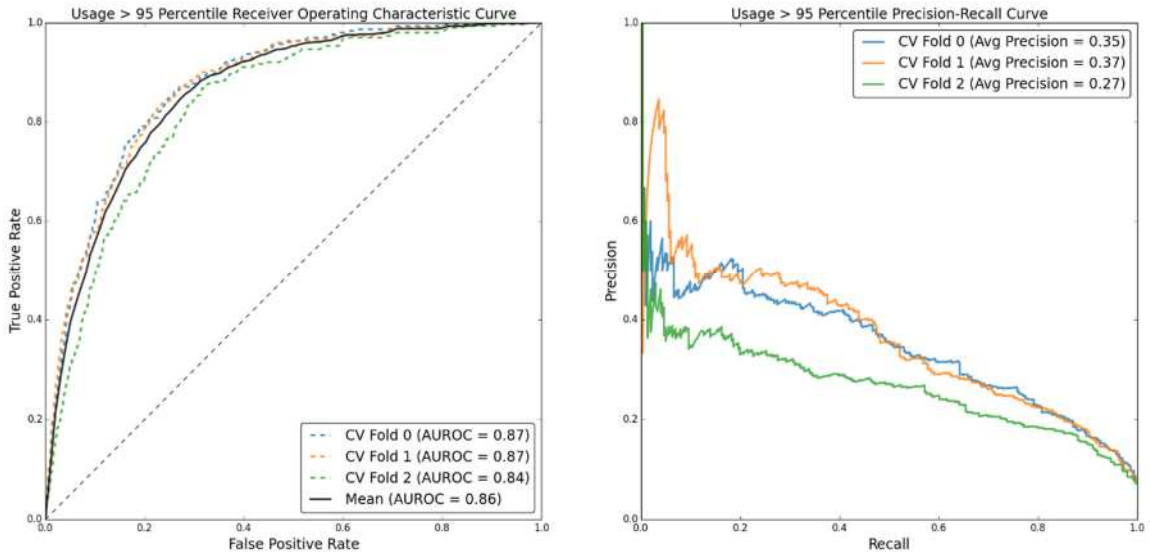

Figure 5: ROC and PR curves for predicting post-BH utilization greater than the non-BH 95th percentile

Table 6: Predicting utilization above the 95th percentile. MCC and F1 optimized performance metrics

Fold	Optimization	MCC	F1	Precision	Recall	TPR	FPR	Threshold
1	MCC	0.392	0.424	0.316	0.639	0.639	0.106	0.173
	F1	0.392	0.424	0.316	0.639	0.639	0.106	0.173
2	MCC	0.385	0.431	0.398	0.465	0.465	0.054	0.199
	F1	0.385	0.431	0.398	0.465	0.465	0.054	0.199
3	MCC	0.325	0.366	0.268	0.571	0.571	0.119	0.170
	F1	0.325	0.366	0.268	0.571	0.571	0.119	0.170
Mean	MCC	0.367	0.407	0.327	0.559	0.559	0.093	0.181
	F1	0.367	0.407	0.327	0.559	0.559	0.093	0.181

The maximal values of the mean rows were used to compare the classifiers with the base case assumption that a patient would continue using the same amount of health care before and after referral. The maxima were used because, unlike said basic assumption, a classifier model can give a range of values based on the individual sample being predicted. While the 99th percentile prediction model significantly outperforms the prior utilization assumption, the 95th percentile model is only a mild improvement, with its major additional contribution being higher recall.

Table 7: Comparison of RFC metrics to constant pre- and post- BH utilization prediction metrics ("Prior")

Percentile Cutoff	Model	MCC	F1	Precision	Recall
99%	RFC	0.266	0.270	0.256	0.397
	Prior	0.160	0.176	0.171	0.181
95%	RFC	0.367	0.407	0.327	0.559
	Prior	0.320	0.371	0.340	0.408

Discussion

Significance

This paper is novel in three ways: First, the studied intervention—a patient's first behavioral health clinic encounter—is both novel in its scope and aligns the studied intervention with resilience theories of behavioral health[19]; in other words, the importance of the intervention is that a patient sought and received behavioral health care, not the precise reason for that interaction. Second, the outcome of patient health care utilization as a primary endpoint, although not entirely novel by itself, has not been well studied using methods similar to this paper. Finally, in addition to structured administrative data, this model uses data derived from lab results and, less common in predictive healthcare research, free text from provider notes.

Using this robust data set and a random forest classifier, we achieved significant improvement over random chance and over prior probability for predicting patient health utilization after behavioral health referral. Furthermore, this

method achieved similar performance metrics to the latest models for predicting early hospital readmission.[10] (Table 8) below compares the classifier metrics reported in He's paper with the metrics in this paper. Ranges for He's paper reflect their lowest and highest reported performance metrics on two different patient cohorts and applied to both same-site and outside institution data sets, as well as the class balance of their data sets in comparison to ours.

Table 8: Comparison of He et al. early hospital readmission to RFC models from this paper

Model	F-Measure	Precision	Recall	AUC	Class Balance
Hospital Readmission	0.19-0.37	0.11-0.34	0.47-0.85	0.65-0.81	9%-16%
Decreased Utilization	0.71	0.67	0.77	0.74	53%
99th Percentile	0.27	0.26	0.40	0.88	2%
95th Percentile	0.41	0.33	0.56	0.86	5%

Because of the relatively unfiltered cohort and broadly defined behavioral health intervention, the decreased utilization classifier may be able to detect patients with behavioral health problems that increase their likelihood of seeking unnecessary care, but which can be hard to for clinicians to detect or appropriately refer.[20] If that is true, it may lead to reduced negative unintended treatment effects, benefiting the physical health of those patients and reducing the cost of unnecessary care.[21, 22] The high-utilization classifier may also be useful in determining which patients need additional supervision, or may be useful in future work evaluating the predictors of high medical costs.

Applications

We envision use of these classifiers in a clinical decision support role. Such a tool would be especially useful to primary care physicians in accountable care organizations and patient centered medical homes, who have a duty to optimize both health status and costs of their panel. Thanks to the high precision across the recall range of the decreased-utilization classifier, providers could see a useful patient-specific estimation of whether referral to a behavioral health provider will be likely to reduce that patient's health care costs. This model may also be able to detect patients who should be referred to behavioral health on an automated basis i.e. running in "the background" of the EMR. For example, patients with "confidence" values in the 0.8 to 1.0 range would have a very high chance of having decreased healthcare utilization following referral. Identification of a patient in this range could prompt an automated notification to their primary provider or to the provider of their next clinic visit, suggesting that the provider might consider whether that patient would benefit from behavioral health services. However, this should not imply that an increase in costs would not be justified by the other benefits of mental health care; additional considerations would be required in a clinical implementation setting.

Lessons

This project taught the student author about the use of appropriate metrics. AUROC seems to remain the standard metric for reporting classification performance in the medical literature. However, classification problems can be highly unbalanced, such as in the case of the 95th and 99th percentile models in this paper. While an AUROC approaching 0.9 may sound impressive, it does not tell the whole story. In a skewed class situation, a strong AUROC may not correlate to strong PR performance and may, if used as the main descriptive measure, mask poor performance.[23] Explicitly, it is possible, as demonstrated by this study's performance in classifying very-high utilization patients, to achieve a high ROC curve while having a low PR curve. Similarly, as in the case of the increased and decreased utilization model, while a high PR curve and strong F1 score are useful descriptors, they still ignore the influence of true negative predictions. Thus, the Matthews Correlation Coefficient, a binary classification analog of Pearson's r that includes all four quadrants of the confusion matrix, was selected as a primary performance metric.

Sometimes, a simple model can be as powerful as a complex model. The classifiers used in this study to predict very high utilization patients were only modestly able to outperform the extremely simplistic prediction that patients will use the same amount of health care before and after the intervention. Not only is this a humbling result, but it also raises an essential question: how much can health spending be predicted by a model? Moreover, were the modest gains in model performance due to poor feature and model selection, or do they reflect the inherent randomness of the measured outcome?

Limitations

One significant limitation of this paper is the fact that the models were tested without a final holdout group. While parameters from model training were not shared between training phases, the core model settings were based on grid search for optimal performance on random test/train slices of the data. Therefore, although test and train groups were randomized, it is possible that the model was over-fitted to the BWH cohort, with test sets acting as internal validation, not a true, fully unseen set of test data.

This study does not have a strong comparison model or baseline clinical prediction. While some studies use clinician chart review as a baseline for comparison, this would have been cost and time prohibitive for our current study. Future research regarding the model's ability to predict changes in cost and to detect patients who would benefit from behavioral health referral would ideally include a clinician comparison. Implementation of the model as a clinical decision support tool would also require modeled prospective implementation and true prospective implementation before being fully adopted.

A potentially significant limitation of the study is hidden in the setting. Brigham and Women's Hospital is a world-renowned tertiary care center. The Psychiatry department at BWH, thanks to its focus on psychosomatic medicine, is particularly focused on treatment of patients with a high amount of comorbid psychiatric, behavioral, and medical complexity. Referrals for more "typical" psychiatric problems or "more healthy" patients, due to finite in-house psychiatric services, are often made to providers outside the Partners system. For these reasons, there may be a selection bias to this cohort, which may limit the generalizability of this model outside the Brigham system. Furthermore, this setting of high medical and psychiatric complexity may limit the generalizability of conclusions regarding the cost-effectiveness of behavioral health services to BWH, where this study's data implies that behavioral health care may be mildly, but not significantly, cost saving.

Although speaking strongly to generalizability, this study used a completely unselected patient cohort and an outcome variable—costs—that can include a range of variables, including pure randomness, human behavior, and disease progression. In addition, the intervention—any sort of behavioral health clinic visit—is exceptionally broad and does not specify whether the studied patients already had mental health diagnoses or mental health treatments. This may have limited the predictive ability of the models. Additionally, it cannot be known from this study whether the intervention is causative of the observed and predicted changes, simply that there is a correlation. In other words, this model may be able to predict changes in healthcare utilization at other points in time, not just after a behavioral health visit.

Further Work

In order to address the limitation caused by the lack of a holdout group, a data request is currently being processed by RPDR for a set of patients from the rest of the Partners Healthcare system. This set of patients will be tested as a holdout group for the models trained on BWH patients. We are also interested in evaluating how many patients from the non-BH cohort will be classified by the model as being potential "reduced utilization" patients.

Thanks to the ability of random forest classifiers to output feature importance values, these models could be used to help discover why some patients use significantly higher levels of care after referral and could also help discover what specific features play a role in predicting beneficial outcomes of behavioral health referrals.[24, 25, 26]

Now that a pipeline has been established for analysis of RPDR-sourced data, this method can be easily modified to train on and predict other outcome variables. Future work will evaluate more specific outcomes, like reduction in glycated hemoglobin ("HbA1c") in diabetic patients following behavioral health encounters. This author also hopes that future work will be in collaboration with other Partners institutions, which may include access to summative clinical data like symptom surveys and validated risk scores; both could serve as additional interesting features and outcome variables. We also hope to utilize Partners high performance computing to annotate provider notes for the studied patients using MetaMap, which may be able to strengthen the signal gained from free text notes beyond our current "bag of n-grams" approach. This process would be a one-time computational cost that could be stored and shared with other projects.

Conclusion

Our study details the development of a random forest classifier that is able to predict patients who will have decreased health resource utilization following their first behavioral health visit, using structured administrative data, lab results, and free text notes as features. We are also able to identify a subset of patients who use very high amounts of resources following behavioral health referral. These models are both an improvement over comparison prior probabilities and compare favorably to similar studies. We hope that this method will be able to improve the timeliness and accuracy of referrals to behavioral health services, improving patient physical and mental health and reducing the financial costs associated with behavioral health comorbidity.

References

1. Nikhil Sahni, Anuraag Chigurupati, Dr. Marian Wrobel, et. al. Massachusetts 2013 Cost Trends Report [Internet]. Commonwealth of Massachusetts Health Policy Commission; 2013. Available from: http://www.mass.gov/anf/docs/hpc/2013-cost-trends-report-final.pdf
2. Schneider A, Hörlein E, Wartner E, Schumann I, Henningsen P, Linde K. Unlimited access to health care--impact of psychosomatic co-morbidity on utilisation in German general practices. BMC Fam Pract. 2011;12:51.
3. Ferrari S, Galeazzi GM, Mackinnon A, Rigatelli M. Frequent attenders in primary care: impact of medical, psychiatric and psychosomatic diagnoses. Psychother Psychosom. 2008;77(5):306-14.
4. Weich S, Lewis G, Donmall R, Mann A. Somatic presentation of psychiatric morbidity in general practice. Br J Gen Pract. 1995;45(392):143-7.
5. Wright A, McCoy AB, Henkin S, Kale A, Sittig DF. Use of a support vector machine for categorizing free-text notes: assessment of accuracy across two institutions. Journal of the American Medical Informatics Association. 2013 Sep 1;20(5):887–90.
6. Rochefort CM, Verma AD, Eguale T, Lee TC, Buckeridge DL. A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data. Journal of the American Medical Informatics Association. 2015 Jan 1;22(1):155–65.
7. Jaimes F, Farbiarz J, Alvarez D, Martínez C. Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room. Critical Care. 2005 Feb 17;9(2):R150.
8. Kessler RC, Warner CH, Ivany C, Petukhova MV, Rose S, Bromet EJ, et al. Predicting suicides after psychiatric hospitalization in US Army soldiers: the Army Study To Assess Risk and rEsilience in Servicemembers (Army STARRS). JAMA Psychiatry. 2015 Jan;72(1):49–57.
9. Ong MEH, Ng CHL, Goh K, Liu N, Koh ZX, Shahidah N, et al. Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. Critical Care. 2012 Jun 21;16(3):R108.
10. He D, Mathews SC, Kalloo AN, Hutfless S. Mining high-dimensional administrative claims data to predict early hospital readmissions. Journal of the American Medical Informatics Association. 2014 Mar 1;21(2):272–9.
11. Research Patient Data Registry (RPDR) | Research Information Services & Computing [Internet]. [cited 2015 Mar 11]. Available from: http://rc.partners.org/rpdr
12. PFS Relative Value Files - Centers for Medicare & Medicaid Services [Internet]. [cited 2015 Mar 12]. Available from: http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/PhysicianFeeSched/PFS-Relative-Value-Files.html
13. Download Anaconda Python Distribution [Internet]. [cited 2015 Mar 11]. Available from: http://continuum.io/downloads#py34
14. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011 Oct;12:2825−2830.
15. Breiman L. Random Forests. Machine Learning. 2001 Oct 1;45(1):5–32.
16. Porcelli P, Rafanelli C. Criteria for psychosomatic research (DCPR) in the medical setting. Curr Psychiatry Rep. 2010;12(3):246-54.
17. Accessing Text Corpora and Lexical Resources [Internet]. [cited 2015 Mar 12]. Available from: http://www.nltk.org/book/ch02.html
18. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.15.2 documentation [Internet]. [cited 2015 Mar 11]. Available from: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
19. Atkinson PA, Martin CR, Rankin J. Resilience revisited. J Psychiatr Ment Health Nurs. 2009 Mar;16(2):137–45.
20. Fazekas C, Matzer F, Greimel ER, et al. Psychosomatic medicine in primary care: influence of training. Wien Klin Wochenschr. 2009;121(13-14):446-53.
21. Bridges K, Goldberg D. Somatic presentation of depressive illness in primary care. J R Coll Gen Pract Occas Pap. 1987;(36):9-11.
22. Lloyd GG. Psychiatric syndromes with a somatic presentation. J Psychosom Res. 1986;30(2):113-20.
23. Fawcett T. ROC Graphs: Notes and Practical Considerations for Researchers. 2004.
24. Perlis RH, Iosifescu DV, Castro VM, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychol Med. 2012;42(1):41-50.
25. Wu ST, Sohn S, Ravikumar KE, et al. Automated chart review for asthma cohort identification using natural language processing: an exploratory study. Ann Allergy Asthma Immunol. 2013;111(5):364-9.
26. Denny JC, Choma NN, Peterson JF, et al. Natural language processing improves identification of colorectal cancer testing in the electronic medical record. Med Decis Making. 2012;32(1):188-97.