Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–17

DOI: 10.1177/0962280216680242
smm.sagepub.com

SAGE

Statistical analysis of a low cost method for multiple disease prediction

Mohsen Bayati,[1,2] Sonia Bhaskar[2] and Andrea Montanari[2,3]

Abstract

Early identification of individuals at risk for chronic diseases is of significant clinical value. Early detection provides the opportunity to slow the pace of a condition, and thus help individuals to improve or maintain their quality of life. Additionally, it can lessen the financial burden on health insurers and self-insured employers. As a solution to mitigate the rise in chronic conditions and related costs, an increasing number of employers have recently begun using wellness programs, which typically involve an annual health risk assessment. Unfortunately, these risk assessments have low detection capability, as they should be low-cost and hence rely on collecting relatively few basic biomarkers. Thus one may ask, how can we select a low-cost set of biomarkers that would be the most predictive of multiple chronic diseases? In this paper, we propose a statistical data-driven method to address this challenge by minimizing the number of biomarkers in the screening procedure while maximizing the predictive power over a broad spectrum of diseases. Our solution uses multi-task learning and group dimensionality reduction from machine learning and statistics. We provide empirical validation of the proposed solution using data from two different electronic medical records systems, with comparisons over a statistical benchmark.

Keywords

clinical disease prediction, multitask learning, feature selection, group regularization

1 Introduction

The emergence of electronic health records (EHRs) has provided the applied statistics and machine learning community with tremendous amounts of data. This data, although not without its challenges, has provided several opportunities for research to enhance patient outcomes and decrease healthcare costs, among others. Much of the focus of this research and public policy efforts has been on disease prediction and prognosis, especially pertaining to chronic diseases.

Chronic diseases have been on the rise in the past decades. During the period 2000 to 2010, it is estimated that the percentage of adults aged 45 to 64 and 65 and over with two or more chronic diseases increased by 5.1% and 8.1%, respectively.[1] The increase in the incidence of chronic diseases has not just impacted the quality of life of this population, it has also had a significant impact on healthcare costs. In the US, from 1987 to 2009, 77.6% of Medicare's spending growth was due to beneficiaries with four or more chronic diseases. In particular, nearly 46% of Medicare spending in 2010 was on patients with six or more chronic diseases (less than 14% of beneficiaries) and the most common diseases among these patients were heart failure, chronic kidney disease, COPD, atrial fibrillation, and stroke.[2]

Hence, early identification of individuals who are at a higher risk of developing a chronic disease is of significant value, both clinically and economically. First, it can create opportunities for slowing down or even reversing the pace of the disease,[3] providing one way to help these individuals improve or maintain their current quality of life. For example, 80% of cases of cardiovascular disease and diabetes, and 40% of cancer cases can be treated

[1]Graduate School of Business, Stanford University, Stanford, USA
[2]Department of Electrical Engineering, Stanford University, Stanford, USA
[3]Department of Statistics, Stanford University, Stanford, USA

Corresponding author:
Mohsen Bayati, 655 Knight Way, E363, Stanford, CA 94305, USA.
Email: bayati@stanford.edu

successfully at an early stage.[4] This early stage intervention is also of significant cost value, as avoiding expensive medical procedures due to complications occurring during the standard care of chronically ill patients and other associated costs can lead to a reduction in healthcare spending.

One proposed solution to the rising numbers of individuals with chronic diseases and associated cost burden, supported by a 2010 study,[5] has been the introduction of incentive-based wellness programs by an increasing number of companies in developed countries. In 2013, 86% of all employers in US, which include Johnson & Johnson, Caterpillar, and Safeway, offered such incentivized programs.[6] These programs involve, among many other aspects, incentivizing employees to undergo a health risk assessment (HRA) in order to identify employees with risk factors for major chronic diseases. The HRA typically involves collecting basic lab results like lipid panels, blood pressure, age, gender, height and weight. This data is then used to calculate risk factors based on simple clinical rules.[7]

Although the potential impact of these wellness programs is anticipated to be high, a 2013 study conducted by the RAND corporation found that to enhance the overall benefit from these programs, the HRAs should be more accurate.[6] This supports further studies which also provide evidence that wellness programs cost more than the savings they generate.[8,9] In particular, a false positive (mistakenly assigning high risk to a healthy patient) leads to unnecessary intervention costs while a false negative (mistakenly assigning low risk to a high risk patient) would be a lost opportunity to avert a large future healthcare bill.

Currently HRAs are built on domain specific knowledge of diseases. The motivation for a more well-designed HRA is clear, but *how* can we design a more effective HRA? Statistical data-driven methods present one principled way to obtain a set of biomarkers for such an assessment which provides high disease prediction accuracy. The use of such methods for disease prediction has grown rapidly over the past few years. An admittedly incomplete list of such studies include: stroke prediction via Cox proportional hazard models,[10] relational functional gradient boosting in myocardial infarction prediction,[11] support vector machines and naive Bayes classifiers for cancer prediction,[12] neural networks for mortality prediction,[13] time series in infant mortality,[14] cardiac syndromes[15] or infectious disease prediction[16] and dimensionality reduction for reducing hospital readmissions.[17]

The majority of this work, however, has focused on individual disease prediction, rather than determining a set of biomarkers which are jointly predictive of several diseases. In principle, we could pool together the relevant biomarkers selected by each individual disease prediction model. However, as we will later show, this method will require many biomarkers and will be expensive. In this paper, we present a solution which uses methodology from group regularization[18] and multitask learning.[19,20] We will show how one can exploit the problem structure to not only design a model which has high predictive accuracy over multiple diseases, but also minimizes the number of biomarkers necessary to achieve this, and can hence be low cost. To the best of our knowledge, our paper is the first to focus on building a small, and thus low cost, universal set of features (biomarkers in our case) that can be used to predict multiple diseases. The closest study to this paper combines multitask learning with logistic regression for multiple disease prediction, however, their objective and model are different.[21] They do not aim to obtain a small set of biomarkers (hence their solution will be not be low cost) and their (latent) features are hard to interpret for medical professionals since each feature is a weighted combination of many basic biomarkers. Finally, we note that a subset of our results (Figures 1, 3a-b, and 4) were presented in 2015 American Medical Informatics (AMIA) conference in San Francisco, CA, USA.

The rest of the paper will be structured as follows. Section 2, covers the models and methods and the sources of data and preprocessing steps are explained in Section 3. The more technical discussion around how the models were trained and how statistical significance is assessed are discussed in Section 4. In Section 5, we validate our models on a large EHR data, and also study its sensitivity to different time horizons for diagnosis, and metrics of accuracy. Generalization of the results to a smaller external dataset is investigated in Section 6, where sensitivity to missing data imputation methods is also assessed. We conclude with a discussion of the methods, limitations, and potential future directions in Section 7.

2 Models and methods

In this section we introduce three approaches for disease prediction, single-task learning, multi-task learning, and OLR-M (short for Ordinary Logistic Regression with top M features) model. Since all of our models are extensions of the classical logistic regression, we start by a brief summary of logistic regression.

Given p-dimensional feature vectors $\boldsymbol{X}_i$ and labels $y_i \in \{0, 1\}$, logistic regression finds the optimal values of a p-dimensional parameter vector $\boldsymbol{\theta}$ and a scalar b such that the probability of observing y_i and $\boldsymbol{X}_i$ for all $i \in [n]$ is maximized. Throughout, we use notation $[n]$ for the set of integers $1, 2, \ldots, n$. The optimization problem for

finding the optimal $\boldsymbol{\theta}$ and b is given by

$$\left(\hat{\boldsymbol{\theta}}, \hat{b}\right) = \underset{\boldsymbol{\theta}, b}{\arg\max} \left[\sum_{i=1}^{n} \log P(y_i | \boldsymbol{X}_i; \boldsymbol{\theta}, b) \right] \tag{1}$$

where the conditional probability $P(y_i|\boldsymbol{X}_i; \boldsymbol{\theta}, b)$ is written in the form

$$P(y_i | \boldsymbol{X}_i; \boldsymbol{\theta}, b) = \frac{\exp\left[y_i(\boldsymbol{\theta}^\top \boldsymbol{X}_i + b)\right]}{1 + \exp(\boldsymbol{\theta}^\top \boldsymbol{X}_i + b)}$$

This form of $P(y_i|\boldsymbol{X}_i; \boldsymbol{\theta}, b)$ will be used throughout the chapter. We will also refer to the term $\sum_{i=1}^{n} \log P(y_i|\boldsymbol{X}_i; \boldsymbol{\theta}, b)$ as *log-likelihood*. In the context of disease prediction, $\boldsymbol{X}_i$ could represent the values of various biomarkers, and y_i could represent a label which indicates whether or not patient i develops a particular disease.

2.1 Single task learning

The single task learning model we will consider formulates the goal of maximizing the log-likelihood, while keeping the set of nonzero parameters small. This model consists of a logistic regression and a regularization term, the ℓ_1 or Lasso penalty.[22,23] In the context of disease prediction, for each disease $k \in [K]$, the model finds the set of parameters that maximizes the probability that a patient will be diagnosed with the disease given his or her vector of biomarkers. More specifically, for each disease k, we solve the optimization problem

$$\left(\hat{\boldsymbol{\theta}}_k, \widehat{b_k}\right)(\lambda) = \underset{\boldsymbol{\theta}_k, b_k}{\arg\max} \left[-\sum_{i=1}^{n} \log P\left(y_i^{(k)} | \boldsymbol{X}_i; \boldsymbol{\theta}_k, b_k\right) + \lambda \sum_{m=1}^{p} |\theta_{km}| \right] \tag{2}$$

The vector $\boldsymbol{\theta}_k$ and scalar b_k are the learned parameters for the model for disease k, with θ_{km} representing the mth element of $\boldsymbol{\theta}_k$. If θ_{km} is nonzero, then the model has selected the mth biomarker as relevant for prediction, and the magnitude of θ_{km} represents the importance of the mth biomarker to the prediction of disease k. The value of λ controls the number of values of m for which θ_{km} are nonzero and is optimized via cross-validation. When $\lambda = 0$, the model reduces to an ordinary logistic regression, which we will refer to as the OLR model.

2.2 Multitask learning model

The multitask learning model[19] formulates the goal of maximizing the log-likelihood and enforcing the set of biomarkers which is jointly predictive of all diseases of interest to be small, as follows. We will call them MTL model and STL model throughout. But if more than one model is mentioned we combine the reference to model. For example a sentence could be "All three models MTL, SLT, and OLR-M are accurate" We updated the definition of acronyms MTL and STL to reduce confusion.

This model consists of a logistic regression and a group-structured regularization term, the *group lasso* penalty.[18] In contrast to the single task learning (STL) model, multitask learning (MTL) model jointly learns the parameters for all disease predictions, by solving a single optimization problem given by

$$\left(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_K, \widehat{b_1}, \ldots, \widehat{b_K}\right)(\lambda) = \underset{\substack{b_1, \ldots, b_K \\ \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K}}{\arg\max} \left[-\sum_{k=1}^{K} \sum_{i=1}^{n} \log P\left(y_i^{(k)} | \boldsymbol{X}_i; \boldsymbol{\theta}_k, b_k\right) + \lambda \sum_{m=1}^{p} \sqrt{\sum_{k=1}^{K} \theta_{km}^2} \right] \tag{3}$$

where all variables are defined as before. The regularization term encourages biomarker m to be either nonzero or zero across all diseases together. In other words, for each m, θ_{km} is aimed to be nonzero or zero for all $k \in [K]$. Such co-variation of parameters through this group regularization can be interpreted as transfer of information between different disease prediction tasks during the training of the model. The model acts by learning the relevant biomarkers for multiple diseases simultaneously, while forcing these relevant biomarkers to be sparse,[19,24,25] resulting in a lower complexity model. We will see this lower complexity model translates to similar conclusions in cost in Section 5. Note that when $K = 1$ in MTL model, the penalty term reduces to the penalty term of STL model, and hence MTL model reduces to STL model.

It is important to note that the group lasso penalty can be used with more general empirical risk minimization terms, beyond the negative log-likelihood for logistic regression which is our focus in this paper. For example,

instead of logistic regression, group lasso could be used to implement MTL model for all generalized linear models. In particular, consider a general loss function $f(\boldsymbol{X}, y; \boldsymbol{\omega})$ for a feature vector $\boldsymbol{X}$ and label y where $\boldsymbol{\omega} \in \mathbb{R}^p$ is the unknown parameter that should be learned. Then the corresponding MTL model for our K task learning problem via group lasso would be given by the following optimization problem

$$(\widehat{\boldsymbol{\omega}}_1, \ldots, \widehat{\boldsymbol{\omega}}_K)(\lambda) = \underset{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_K}{\arg\max} \left[\sum_{k=1}^{K} \sum_{i=1}^{n} f\left(\boldsymbol{X}_i, y_i^{(k)}; \boldsymbol{\omega}_k\right) + \lambda \sum_{m=1}^{p} \sqrt{\sum_{k=1}^{K} \omega_{km}^2} \right] \tag{4}$$

2.3 OLR-M model

The OLR-M model is derived from the MTL model. After solving for the estimates $(\hat{\boldsymbol{\theta}}_k, \widehat{b_k})$, $k \in [K]$, for the MTL model with parameter λ, we retrain a truncated OLR model, as follows. For each value of λ, we determine the nonzero biomarkers, i.e. the values of m for which the quantity $\sqrt{\sum_{k=1}^{K} \hat{\theta}_{km}^2}$ was above a fixed threshold. We denote the number of nonzero biomarkers by M_λ, and give further details calculating M_λ in Section 4.1. We then retrain the model over only these M_λ biomarkers, separately for each disease prediction $k \in [K]$, using an ordinary logistic regression. This refitting is known to remove some of bias of the model due to regularization and improve performance. The optimization problem solved is given as

$$\left(\hat{\beta}_k, \widehat{c_k}\right)(\lambda) = \underset{\boldsymbol{\beta}_k, c_k}{\arg\max} \left[-\sum_{i=1}^{n} \log\, P(y_i | \boldsymbol{X}_i^{\lambda}; \boldsymbol{\beta}_k, c_k) \right] \tag{5}$$

where $\boldsymbol{\beta}_k$ refers to the truncated parameter vector of length M_λ, c_j is the intercept term, and $\boldsymbol{X}_i^{\lambda}$ corresponds to the truncated vector of M_λ biomarkers of patient i.

3 Data

Before validating the methods, we describe the datasets that include two sources of patient data and a separate source for the cost of laboratory exams.

3.1 Diseases of interest

We studied a total of nine diseases. These diseases have been suggested as a broad set of chronic diseases.[4,26,27] Their names followed by their abbreviations are: coronary artery disease (CAD), malignant cancer of any type (Cancer), congestive heart failure (CHF), chronic obstructive pulmonary disorder (COPD), diabetes (DB), dementia (DEM), peripheral vascular disease (PVD), renal failure (RF), and severe chronic liver disease (SCLD). These diseases are currently some of the most prevalent in the population. Within the Medicare beneficiaries, 31% have heart disease, followed by diabetes, renal disease, chronic obstructive pulmonary disorder, and cancer, with rates 28%, 15%, 12%, and 8% respectively.[2]

3.2 Patient data

We considered two patient populations, the Kaggle Practice Fusion dataset which is publicly available,[28] and patient records from Stanford healthcare (SHC), which will be referred to as the SHC dataset throughout. Our use of SHC data was approved by Institutional Review Board at Stanford University (protocol 20729). The SHC dataset contained 108,084 patients and 1313 available laboratory exams. The Kaggle dataset was much smaller, and had only 1096 patients with 258 available laboratory exams. The Kaggle dataset contains more patients, but only 1096 patients had laboratory exam data.

The patient data from the SHC dataset was extracted as follows. For this dataset, K is equal to 9. For each disease k for $k \in [K]$, we collected the biomarkers of patients during a one-month period, which occurred during 2010 to 2011. The biomarkers included age, gender, and the results of available laboratory exams. If there were multiple results for a laboratory exam during this month, they were averaged. If patients had results over multiple months, we took the first month for which results occurred. If the patient was diagnosed with disease k during or prior to this month, we removed the patient from the dataset. Therefore, we only considered patients who had never had any disease k, where $k \in [K]$, before the start of the month in which their results were taken.

We collected these biomarkers into a vector $\boldsymbol{X}_i \in \mathbb{R}^p$, $i \in [n]$, where X_{im} took the value of the result of biomarker m for patient i, n was the total number of patients collected, and p was the number of biomarkers. We assigned each patient a label $y_i^{(k)}$ for each disease, which we defined to be 1 if the patient had a positive diagnosis of disease k within one to thirteen months from the beginning of the month that the biomarker values were recorded, and 0 otherwise. We removed all patients that died during this time period. Therefore, our final SHC dataset included $n = 75,619$ patients. We also considered other time horizons, with diagnoses occurring within one to six months, one to eighteen months, and twelve to twenty four months, as a robustness check.

In the Kaggle dataset, due to lack of access to the exact time stamp for the laboratory exams or diagnosis codes, the vectors $\boldsymbol{X}_i$ were formed for each patient from the laboratory results and age information, without any use of temporal information. The label $y_i^{(k)}$ was determined by checking whether the patient had a positive diagnosis for disease k in the system. The diseases Cancer, DEM, RF, and SCLD were omitted from our analysis due to their scarcity in the data (hence $K = 5$ for the Kaggle dataset).

Finally, we standardized the observations so that the values of each biomarker have zero mean and variance one across the observations.

3.3 Imputation of missing values

For both datasets, the patients did not have results for all of the possible laboratory exams. To deal with the missing data entries, we used mean imputation and gave each missing value the average value of the laboratory exam, where the average was taken over all patients who had taken the exam. As a robustness check, in the model validation step, we repeated our analysis using hot deck (HD) multiple imputation and using multiple imputation by chained equations (MICE).[29]

In HD imputation, the missing data values are filled in by randomly sampling (with replacement) the available entries. For example, for each laboratory exam in the data matrix, we would populate the patients' entries which did not have values for those laboratory exams with randomly sampled entries of patients who did. Then the procedure is repeated multiple times to obtain a distribution for predictions of each model. While this method is an easy way to obtain multiple imputations and is non-parametric, it still distorts correlations and other measures of association between features.[30]

MICE is another method for multiple imputation that takes correlation between different features into account. One assumption necessary for using MICE is that the probability of missingness depends on the observed data but not on the missing data. For example, male participants may be less apt to a laboratory test that measures their cholesterol level, but it does not depend on their cholesterol level. MICE operates by a series of regression models, where each variable is modeled conditional on the other variables in the data.[31] A mean imputation is first used to fill in all missing values, and then the missing values of a variable are determined by a regression based on the other variables. The missing values are then filled in with these values. This procedure is cycled through several times, for all variables.

3.4 Cost data

Data for the cost analysis was taken from a publicly available website operated by Healthone Labs, which contains laboratory exam cost data.[32] Healthone Labs offers laboratory exams and has physical locations in all states in United States. Its website provides competitive prices of various laboratory exam packages, which are comprised of several laboratory exams. These prices were used to determine the cost of the set of laboratory exams which were selected by our prediction methods.

We emphasize that most laboratory exams are not sold individually but in packages. Thus, in order to find the cost of our screening procedure, we solved a set-covering problem, formulated as a mixed-integer optimization problem where we minimized the cost over the possible sets of packages but enforced the choice of packages to include all of the laboratory exams required by each disease prediction method. This problem can also be formulated as a convex linear programming problem, but since the size of the problem did not lead to a computational impasse, the mixed integer program was suitable for our needs. More specifically, we solved the optimization problem given by

$$\text{minimize} \quad \sum_{S \in \mathcal{S}} c_S q_S,$$

$$\text{subject to} \quad \sum_{S:e\in\mathcal{S}} q_S \geq 1 \quad \forall e \in \mathcal{U},$$
$$q_S \in \{0,1\} \quad \forall S \in \mathcal{S} \tag{6}$$

where $\mathcal{S}$ is the set of all packages, $\mathcal{U}$ is the set of laboratory exams for which we wish to obtain the total minimum cost, and q_S is an indicator variable which indicates whether a particular package $S \in \mathcal{S}$ is chosen for inclusion, and c_S is the cost of package $\mathcal{S}$.

We solved this optimization problem using the integer programming toolbox in MATLAB. We then calculated the cost of the laboratory exams by adding the prices of the set of packages $S \in \mathcal{S}$ for which $q_S = 1$ in the solution of equation (6).

4 Optimization and statistical significance

In this section, we first explain how the models of Section 2 were trained on the data, and then describe the statistical analysis performed to assess the significance of our results.

4.1 Learning the model parameters

The optimization package `MinFunc`,[33] which utilizes spectral projected gradient,[34] was used for solving the associated optimization problems described in Section 2.

The main performance metric for all prediction tasks is area under receiver operating characteristic (ROC) curve that is called AUC. Note that ROC curve is obtained by graphing classifier's true positive rate (TPR) against its false positive rate (FPR) for all possible discrimination thresholds. In other words, AUC measures how fast TPR grows compared to FPR as the threshold is varied.[35] AUC is sometimes called the c-statistic and is also equivalent to the Mann–Whitney U statistic (or Wilcoxon rank-sum test), for the median difference between the prediction scores in the two groups.[36] In addition to AUC, for robustness check, we compared model accuracy using positive predictive value and sensitivity as well.

For each value of λ, all models were trained and tested using five-fold cross validation when obtaining the AUCs for the SHC dataset. During training, due to the sparsity of positive examples, in each cross-validation fold the training data was sampled so that, for the SHC dataset, the ratio of positive to negative examples in the training set was 1 to 1. Since the Kaggle dataset was much smaller, we instead generated 1000 different realizations of the data via the bootstrap method, and obtained the AUCs using this data. The ratio of positive to negative samples during the training used for the Kaggle dataset was 1 to 3. All demonstrated AUC values in the figures are AUCs that were averaged over the test sets; they were averaged over the folds in the case of the SHC dataset, and over the 1000 different realizations for the Kaggle dataset. Since the AUC values on Kaggle data are optimistic due to the bootstrap procedure, they were corrected via the optimism method,[37] explained in Section 4.2 below. They were then averaged over the K diseases.

For each value of the regularization parameter λ, we considered biomarker m to have been selected by the model if for any disease k, it had a nonzero value of θ_{km}. We determined this as follows. For both the STL and MTL models, for a particular value of λ, the number of nonzero biomarkers was determined by counting the number of biomarkers m for which the quantity

$$\frac{\theta_{km}^2}{\sum_{m=1}^{p} \theta_{km}^2}$$

was greater than a threshold of $\tau = 10^{-6}$ for at least one value of $k \in [K]$. Thus each value of λ corresponds to a set of nonzero biomarkers, and a cost can be derived via solving the optimization problem given in equation (6).

4.2 Determining significance of results

We computed the P -values for the costs by splitting the SHC data into ten disjoint sets, and estimating a confidence interval for the resulting ten cost values using the standard t-distribution. This approach was chosen since the underlying distribution of the cost is difficult to obtain, and the independence of the results on the ten independent datasets facilitates computation of the standard error.

We obtained the P -values for comparing different AUC values on SHC dataset using the popular Delong method.[38] Due to the small size of the Kaggle dataset, we used a bootstrap method (in addition to Delong) to obtain AUC confidence intervals. However, since an estimate of the AUC obtained via bootstrap is optimistic, we adapted the optimism method[37] that is designed for mean-squared error, to correct the estimated AUC and confidence intervals via bootstrap again. Next, we will describe this *nested bootstrap* approach.

4.2.1 Confidence intervals via nested bootstrap

Let us denote the original Kaggle dataset by D_0. We obtained B bootstrapped datasets from D_0, referred to as D_b, $b \in [B]$. We then trained a given prediction method on all of these bootstrapped datasets and obtained models M_b, $b \in [B]$. We then calculated AUC_b by testing model M_b on D_b, and calculated $\mathrm{AUC}_{b,0}$ by testing model M_b on the original dataset D_0. The, optimism was then defined by

$$\widehat{\text{optimism}} = \frac{1}{B}\sum_{b=1}^{B}(\mathrm{AUC}_b - \mathrm{AUC}_{b,0}) \tag{7}$$

Then, using the same prediction method, we trained and tested a model M_0 on the original dataset D_0, to obtain AUC_0. Our final estimate of the AUC, denoted by $\widehat{\mathrm{AUC}}$ was calculated by

$$\widehat{\mathrm{AUC}} = \mathrm{AUC}_0 - \widehat{\text{optimism}} \tag{8}$$

In order to build confidence interval for $\widehat{\mathrm{AUC}}$, the above procedure was repeated B' times but each time instead of using D_0 as the original data set, a new bootstrapped version of D_0 was used. This led to B' values for $\widehat{\mathrm{AUC}}$ that were used with the $\mathbf{BC}_a$ method[37] to obtain a confidence interval. We used $B = 200$ and $B' = 1000$ for the experiments.

4.3 Computation environment

All computations were performed on a Dell Power Edge R710 server with two Intel Xeon CPUs (model X5570, 2.93Ghz, 8M Cache, Turbo HT) and 36GB Memory. Analysis of Sections 2, 3.4 and 4.1, and all plots were done using MATLAB version R2014b. All statistical validations; i.e. hypothesis testing, calculation of AUC confidence intervals (nested bootstrap and Delong), and multiple imputations (HD and MICE) were done in R version 3.2.0. Since Kaggle data set is publicly available, we have provided our source codes for Kaggle data in a public code repository.[39] The repository also contains instructions for MATLAB functions that were used to solve the optimization problems (3) and (6) on SHC data.

5 Model validation: SHC example

We now discuss the results obtained from validating our models using the data described in Section 3. Since we are interested in comparing the complexity and accuracy of the models, most of our results will focus on these two quantities. We will study sensitivity of our results to various perturbations in the data, such as time between diagnosis and results, as well as different imputation methods, and also look at our results under other metrics of accuracy. We will corroborate our results by reporting statistical significance within each comparison.

5.1 Accuracy versus model complexity and cost

We compare the prediction accuracy and model complexity, measured by the AUC and number of biomarkers, respectively, for all three models STL, MTL, and OLR-M. The average cross-validation AUC versus the number of biomarkers used or cost of biomarkers used are shown in Figures 1(a) and 1(b), respectively. Each of the three curves reports the average AUC versus the number of biomarkers or cost for one of the STL, MTL, and OLR-M models. The results in Figure 1(a) demonstrate that the MTL and OLR-M models achieve comparable accuracy to the STL model with a much smaller number of biomarkers. These results translate consistently when AUC is compared against cost in Figure 1(b). For each of the three models the point on the curve with the highest average AUC would correspond to the optimal tuning parameter for that model in terms of accuracy. In particular, for the OLR-M model the optimal value for the number of biomarkers is $M = 30$.

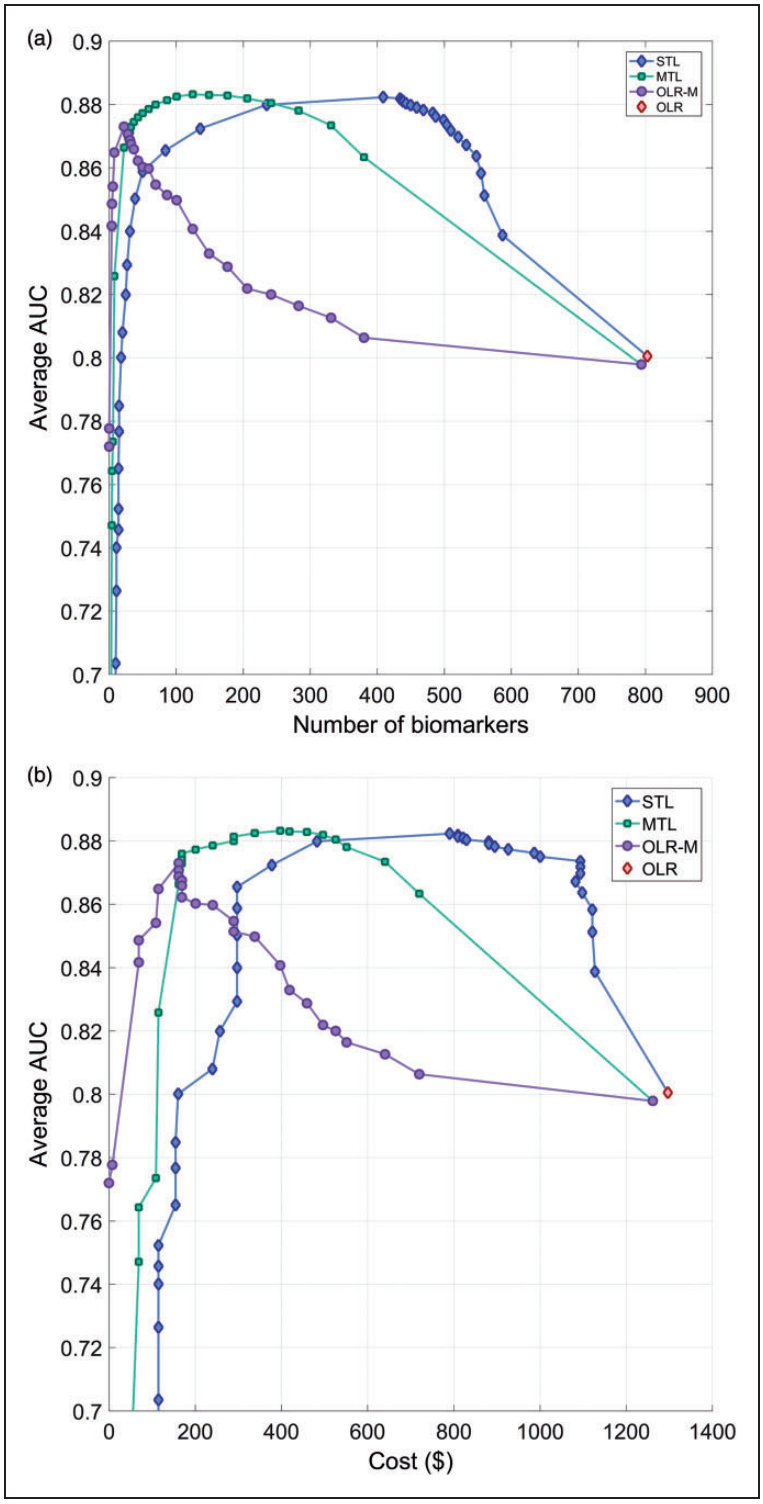

Figure 1. Comparison of AUCs and number of biomarkers selected by or associated cost of the MTL, STL, and OLR-M models. Points on the curves are obtained by changing the regularization parameter. (a) Number of features comparison. (b) Cost comparison.

Table 1 reports the *P*-values for testing the hypotheses that the MTL or OLR-M models are more accurate than the STL model for each disease. These relatively large *P*-values, obtained using the Delong method, demonstrate that there is not enough evidence to reject the null hypotheses; hence we cannot argue that the STL model has significantly better accuracy over all diseases than the MTL and OLR-M models.

Table 1. *P*-values for comparing AUCs of all approaches; H_0 is the null hypothesis.

	CAD	CANCER	CHF	COPD	DB	DEM	PVD	RF	SCLD
$H_0 : AUC_{MTL} > AUC_{STL}$	0.81	0.73	0.71	0.81	0.99	0.10	0.74	0.04	0.21
$H_0 : AUC_{OLR\text{-}M} > AUC_{STL}$	0.60	0.09	0.32	0.60	0.47	0.07	0.44	0.004	0.01

AUC: area under receiver operating characteristic curve; CAD: coronary artery disease; CHF: congestive heart failure; COPD: chronic obstructive pulmonary disorder; DB: diabetes; DEM: dementia; PVD: peripheral vascular disease; RF: renal failure; SCLD: severe chronic liver disease; MTL: multitask learning; STL: single task learning; OLR-M: Ordinary Logistic Regression with top M features.

Table 2. 95% confidence intervals for costs of all three models at their highest average AUC point.

Method	Cost ($)
STL	526.1 ± 50.294
MTL	342.7 ± 50.09
OLR-30	168.5 ± 72.97

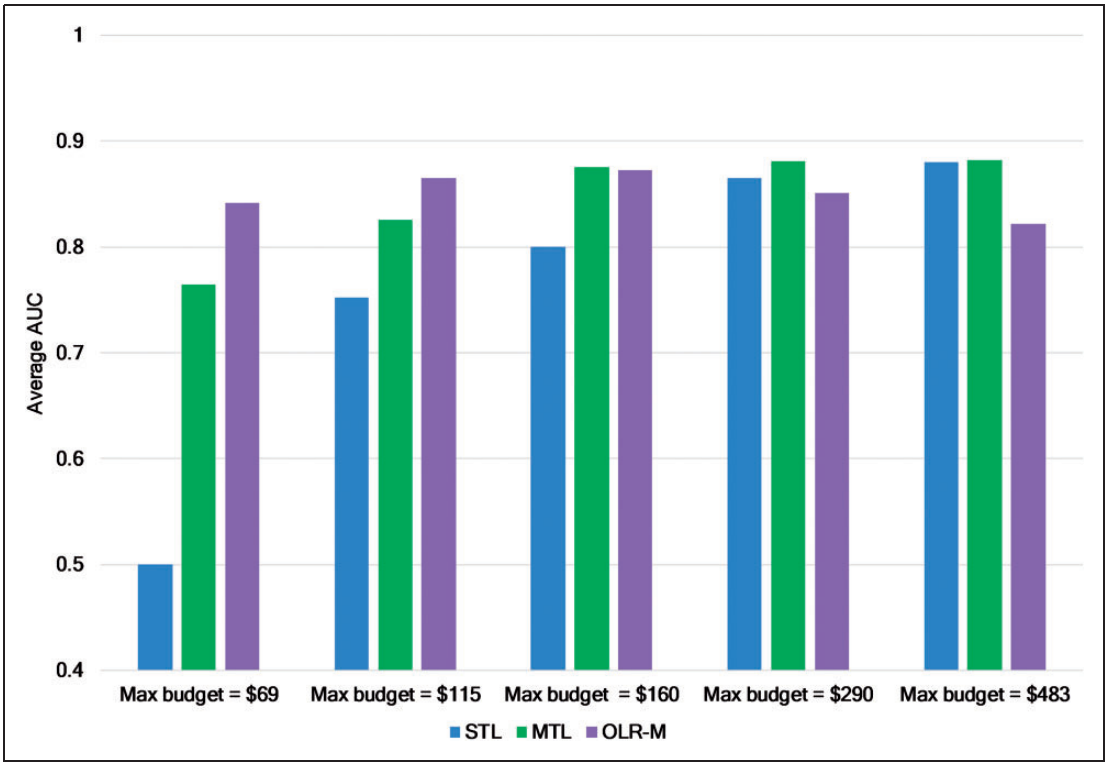

Figure 2. Comparison of accuracies under limited budgets of $69, $115, $160, $290, and $483.

Comparing the cost at the points with the highest AUC across all three curves on Figure 1(b), it is evident that the MTL and OLR-M models achieve a lower cost than the STL model, with little or no loss in prediction accuracy. This result is statistically significant, at a level $P < 0.01$; 95% confidence intervals are given in Table 2.

Under a limited budget, i.e. the maximum allowable cost for the biomarkers, a comparison of the accuracies for each of the MTL, OLR-M, and STL models is shown in Figure 2. The results are obtained when the maximum allowable budget is $69, $115, $160, $290, and $483. The OLR-M model is the most accurate model at $69 and

Table 3. List of the top features with their rank determined by the OLR-30 model.

Rank	Lab name
1	Creatinine
2	Age, gender
3	Hemoglobin A1C
4	Platelet count
5	Glucose
6	Alkaline phosphatase
7	QRSD interval
8	Prothrombin time/INR
9	Urea nitrogen
10	EGFR
11	Vitamin B12
12	HDL cholesterol
14	T-wave axis
15	Triglyceride
16	Globulin
17	RDW
18	Alpha-1 antitrypsin
19	Bilirubin

QRSD: Quick Response System Duration interval; EGFR: Estimated Glomerular Filtration Rate; HDL: High Density Lipoproteins; RDW: Red cell Distribution Width.

$115. As we increase the budget to $160, the MTL and OLR-M models nearly tie and both outperform the STL model. The STL model has better accuracy than the OLR-M model when the higher budget of $290 is considered and is the most accurate solution at $483.

5.2 Biomarkers selected for OLR-M model

As mentioned above, the value of M which achieved the highest average AUC (computed via cross-validation) was $M = 30$. The list of these top $M = 30$ biomarkers used by the OLR-M model is shown in Table 3. All of these biomarkers are associated with the diseases of interest in this paper, and as expected from the design of our method, many of them are associated with several diseases. For example, abnormal glucose levels are associated with signs of diabetes, cancer, renal disease, liver disease, and heart disease. Similarly, abnormal alkaline phosphatase levels can be associated with liver disease, cancer, as well as heart disease. A vitamin B12 test is often given to elderly patients who are experiencing neuropathy, but higher levels can also be associated with diabetes, liver disease, and heart problems.

In summary, our analysis selects biomarkers that are highly predictive of a broad spectrum of chronic diseases and, at the same time, allows for discrimination among them.

5.3 Sensitivity to length and shift of time window

We studied the effect of varying the time window over which we identify a patient as having a positive diagnosis. In the results presented in Section 5.1, the time window considered for the diagnosis was one to thirteen months after biomarkers were taken. We considered a shift of a year that corresponds to a time window of twelve to twenty-four months after the biomarkers were taken. We also studied the performance when the length of time horizon is varied by considering the interval one-to-six months and the interval one-to-eighteen months after the biomarkers were taken.

When the time interval is twelve to twenty-four months, we observed the same pattern as discussed in the previous section, although not as statistically significant as for the one-to-thirteen month horizon. The MTL and OLR-M models were still able to achieve comparable AUC to the STL model with fewer biomarkers and thus lower cost. When compared to the STL model, the difference in the cost for the MTL model at the expense of accuracy is not as significant as in the previously shown results but for the OLR-M model the difference is still significant. The Figure 3(a) presents this analysis. P-values for the differences in the AUCs are given in Table 4.

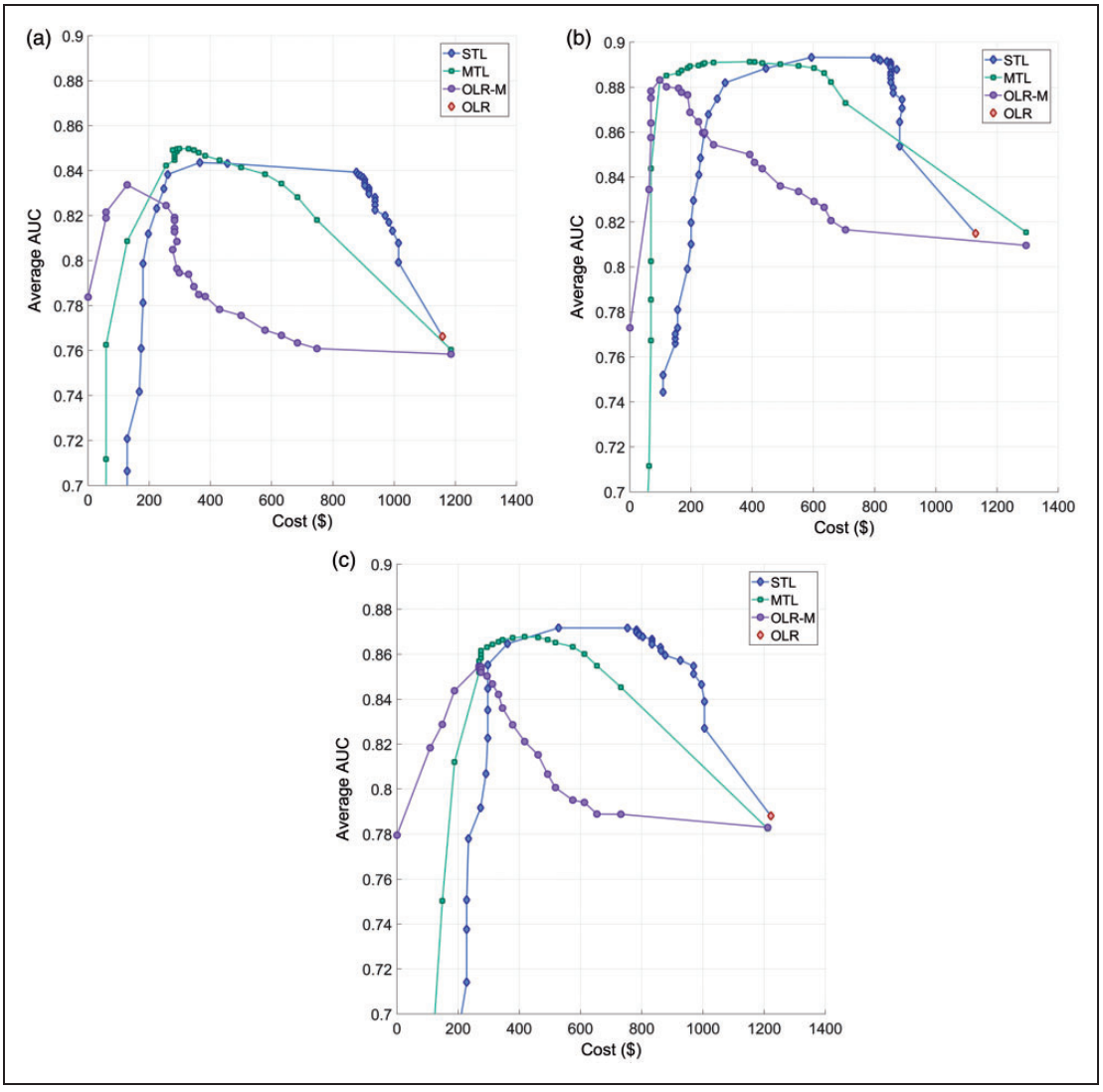

Figure 3. Comparison of costs and accuracy of models, for predicting the onset of diseases for shifted and varied lengths of the time horizon after the lab results. (a) Twelve to twenty four months after lab results. (b) Six month time horizon after lab results. (c) Eighteen month time horizon.

Table 4. *P*-values for comparing AUCs of all approaches; diagnosis between twelve to twenty four months.

	CAD	CANCER	CHF	COPD	DB	DEM	PVD	RF	SCLD
$H_0 : AUC_{MTL} > AUC_{STL}$	0.55	0.30	0.63	0.08	1.0	1.0	0.93	0.001	0.97
$H_0 : AUC_{OLR\text{-}M} > AUC_{STL}$	0.63	0.78	0.70	0.66	1.0	1.0	0.66	0.40	0.97

When the time horizon considered is one to six months, we also see a similar pattern. The MTL and OLR-M models are able to achieve a significantly lower cost model than the STL model as shown in Figure 3(b). The AUC values for all models for this time horizon has increased, which can be expected since the latest possible diagnosis is closer to the prediction time than for the one-to-thirteen month window. Figure 3(c) presents the results when the time horizon is one to eighteen months, in which the same effect is observed, but, again, less pronounced. *P*-values for the differences in the AUCs of the approaches is given in Tables 5 and 6, respectively.

Table 5. *P*-values for comparing AUCs of all approaches; six month time horizon.

	CAD	CANCER	CHF	COPD	DB	DEM	PVD	RF	SCLD
$H_0 : AUC_{MTL} > AUC_{STL}$	0.23	0.04	0.29	0.23	1.0	1.0	0.13	0.65	0.03
$H_0 : AUC_{OLR\text{-}M} > AUC_{STL}$	0.39	0.46	0.73	0.39	1.0	1.0	0.23	0.46	0.23

Table 6. *P*-values for comparing AUCs of all approaches; eighteen month time horizon.

	CAD	CANCER	CHF	COPD	DB	DEM	PVD	RF	SCLD
$H_0 : AUC_{MTL} > AUC_{STL}$	0.41	0.13	0.43	0.47	1.0	1.0	0.26	0.03	0.0003
$H_0 : AUC_{OLR\text{-}M} > AUC_{STL}$	0.78	0.75	0.69	0.83	1.0	1.0	0.40	0.07	0.23

5.4 Sensitivity to the metric of accuracy

For further validation, we evaluated the accuracy of the STL, MTL, and OLR-M models using two additional metrics: sensitivity and positive predictive value (PPV). We restricted our comparison to the optimal tuning parameter for each model, i.e. the highest AUC points for each curve in Figures 1(a) to 1(b) as discussed above. Similar to AUC, all approaches have comparable accuracy with respect to sensitivity (or true positive rate). With respect to positive predictive value, the MTL and OLR-M models are indistinguishable and closely follow the STL model. These findings are shown in Figures 4(a) and 4(b).

5.5 Visual discrimination ability

We studied the discriminating ability of all approaches for diabetes near the maximum allowable budget of \$115. At this cost, each model produces a score for each patient being diagnosed with diabetes. We show these scores for the positive and negative cases in a histogram. For an accurate representation, we have sub-sampled the negative cases so there are equal number of positive and negative cases. The results as shown in Figures 5(a), 5(b) and 5(c) show better differentiation by OLR-M and MTL models compared to the STL model.

6 Model validation – Kaggle example

We investigate the performance of the OLR-M model for $M = 30$ on a different patient population, the Kaggle dataset, as well as the performance under different imputation methods for the missing data. As aforementioned, the Kaggle dataset is a less ideal dataset, with fewer biomarker values available per patient. Recall from the previous section that the OLR-30 model consists of the biomarkers shown in Table 3, which were selected via cross-validation on the SHC dataset. The model was retrained on the Kaggle dataset using only these biomarkers.

6.1 Generalizability of OLR-30 to different patient populations

The OLR-30 model is able to achieve comparable accuracy to the STL model, with far fewer biomarkers, and thus has substantially lower cost. In particular, while the OLR-30 model uses 30 biomarkers, the STL model uses 64 biomarkers. Table 7 provides 95% confidence intervals for the AUC values for each disease on Kaggle data. As a robustness check, we obtained the intervals using two different methods.[37,38]

6.2 Sensitivity to other imputation methods

The above results were obtained when missing biomarker values were imputed via mean imputation, however other imputation methods exist. In addition, we may want to account for additional uncertainty in our estimates that results from having missing data. We focused on two methods, known as multiple imputation by chained equations (MICE)[29] and hot deck (HD),[30] which are described in Section 3.3. The goal

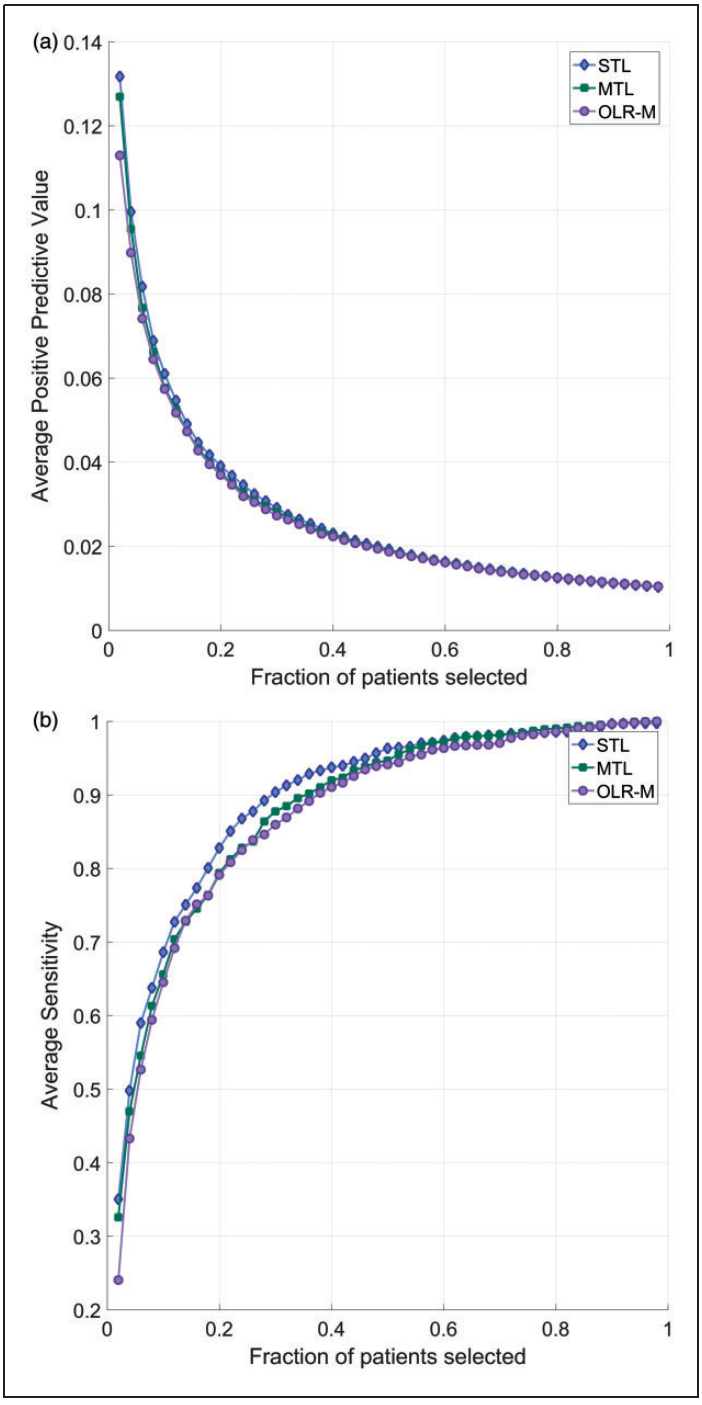

Figure 4. Comparison of predictive accuracy of models based on sensitivity and positive predictive value. Each point on *x*-axes serves as a fraction of patients (the fraction with the highest predicted risk averaged over all diseases) selected by the models as high risk. This threshold yields values for positive predictive value and sensitivity for each model and each disease. (a) Average sensitivity versus fraction of selected patients. (b) Average positive predictive value versus fraction of selected patients.

of this analysis was to check if the same pattern of results held when the dataset was imputed using multiple imputation methods. The results, as shown in Table 8, are consistent; the MTL and OLR-M models produce lower cost models with comparable accuracy to the STL model. We used 1000 imputed datasets to obtain our confidence intervals.

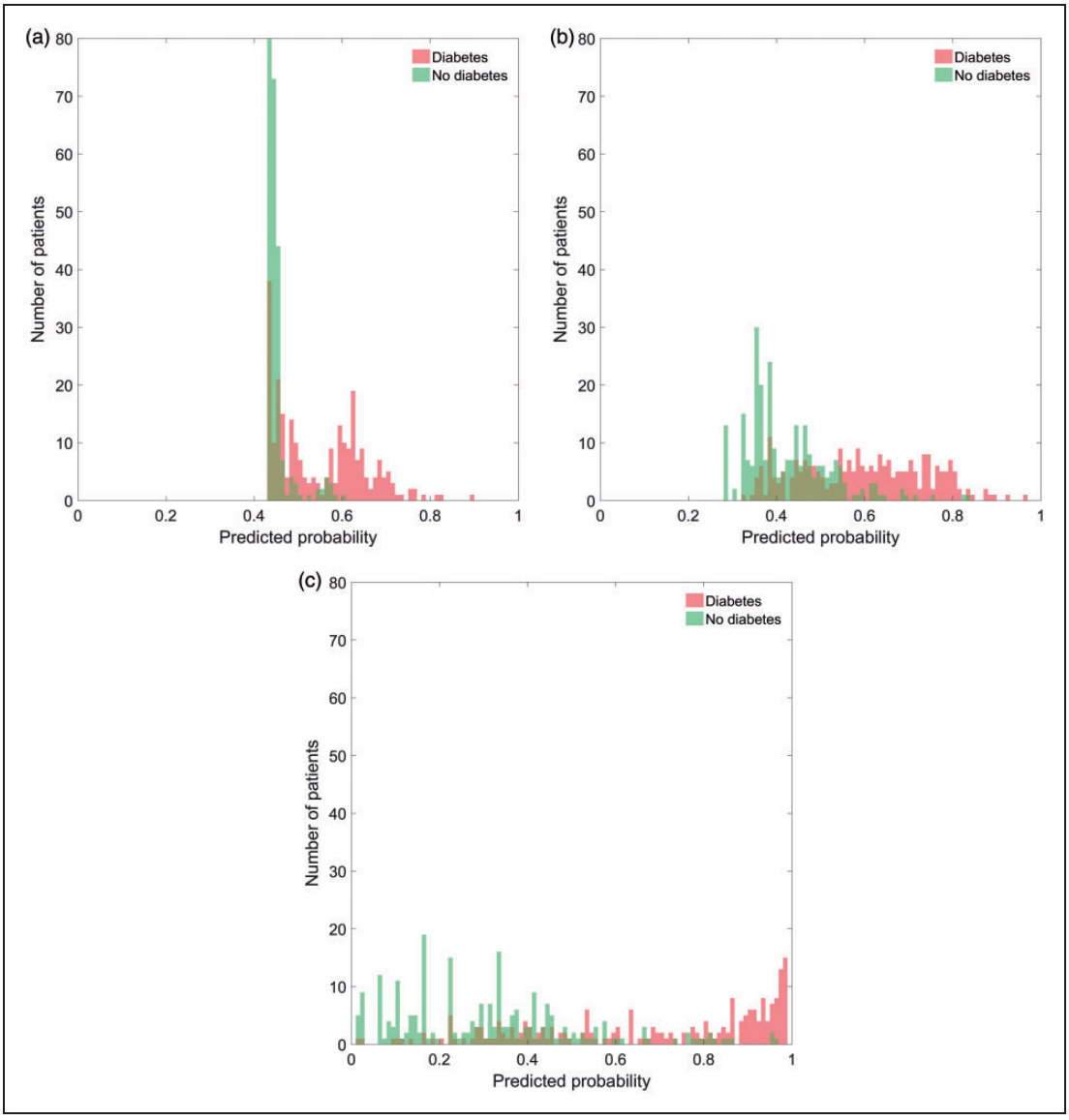

Figure 5. Visual comparison of discrimination in scores of all approaches at cost of $115. (a) STL: AUC = 0.81. (b) MTL: AUC = 0.83. (c) OLR-M: AUC = 0.87.

Table 7. 95% confidence intervals using both the optimism[37] and Delong methods,[38] for the AUC values of OLR-30 and STL approaches.

	Optimism		Delong	
Disease	OLR-30	STL	OLR-30	STL
CAD	(0.7339, 0.7979)	(0.7082, 0.8157)	(0.7015, 0.8056)	(0.6875, 0.7950)
CHF	(0.6557, 0.7246)	(0.5824, 0.7184)	(0.5459, 0.6547)	(0.5432, 0.6490)
COPD	(0.7259, 0.8738)	(0.6655, 0.7876)	(0.5305, 0.7680)	(0.5444, 0.7765)
DB	(0.7890, 0.8818)	(0.7377, 0.8748)	(0.7196, 0.8718)	(0.7195, 0.8655)
PVD	(0.8017, 0.9650)	(0.8685, 0.9627)	(0.8388, 0.9980)	(0.8430, 0.9935)

Table 8. 95% confidence intervals on the Kaggle dataset for different imputation methods.

	MICE		Hot deck	
Disease	OLR-30	STL	OLR-30	STL
CAD	0.6781 ± 0.028	0.7179 ± 0.0210	0.7133 ± 0.0252	0.7067 ± 0.0277
CHF	0.5772 ± 0.0303	0.5876 ± 0.0268	0.5748 ± 0.0328	0.5769 ± 0.0279
COPD	0.6522 ± 0.0602	0.6271 ± 0.0559	0.6563 ± 0.0669	0.6486 ± 0.0509
DB	0.6918 ± 0.0447	0.7117 ± 0.0412	0.7537 ± 0.0361	0.7367 ± 0.0323
PVD	0.7478 ± 0.0810	0.7915 ± 0.0467	0.7529 ± 0.0805	0.7886 ± 0.0577

7 Discussion

In this paper, we have presented a data-driven statistical approach to finding a subset of laboratory exams that accurately predicts the onset of any of nine diseases, and is low cost. We have achieved this via the MTL model which relies on multitask learning as well as group regularization, as well as the OLR-M model which uses MTL as a precursor. We have compared our methods to STL on the basis of accuracy, cost, and complexity. We have also studied the sensitivity of all methods to changes in time to diagnosis, imputation methods, different data, and other measures of accuracy.

While we would like to also compare our methods against common scoring rules used for individual diseases, we were unable to do so at this time because of the unavailability of necessary biomarkers in our data for these scores. For example, the Framingham score for predicting the ten-year risk of cardiovascular disease has an AUC of 0.78 in some populations.[40] The German diabetes score, predictive of diabetes, has an AUC of 0.70 in a particular population.[41] The Child–Pugh score for assessing risk of chronic liver disease, had a reported AUC of 0.86[42] in one study. While we would like to be able to evaluate these scores on our datasets, the datasets considered here are missing the necessary information, such as whether a patient was a smoker, in order to make proper comparisons. Hence, we limit our comparison to the state-of-the-art statistical method for disease prediction, namely the STL model, using age, gender, and comprehensive laboratory results.

Another limitation of this study is that we did not use other types of data such as diagnosis history, vital signs, etc. that can be found in most EHRs. The main reason for not using such data is that we did not have access to the cost of collecting those biomarkers. Obtaining a uniformly accepted cost data for collecting various biomarkers is quite challenging. We leave extending our results to include a richer set of biomarkers to future studies.

Acknowledgements

The authors are extremely grateful to L. Wein and S. Zenios for their encouragement and feedback on an early version of the manuscript; to Stanford Center for Clinical Informatics (SCCI) for assistance with data access and hosting a secure data analysis environment; and to our anonymous review team for guidance.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Bayati received support from Graduate School of Business at Stanford University and the National Science Foundation awards CCF:1216011 and CMMI:1451037. Montanari received support from National Science Foundation award CCF-1319979 and Air Force Office of Scientific Research grant FA9550-13-1-0036.

References

1. Freid VM, Bernstein AB and Bush MA. Multiple chronic conditions among adults aged 45 and over: trends over the past 10 years. *NCHS Data Brief* 2012; **100**: 1–8.

2. Bartee S, Wheatcroft G, Krometis J, et al. Chronic conditions among Medicare beneficiaries, chartbook. Technical Report. 2012. Baltimore: Centers for Medicare and Medicaid Services.
3. Bates DW, Saria S, Ohno–Machado L, et al. Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Aff (Millwood)* 2014; **33**(7): 1123–1131.
4. Morabia A and Abel T. The WHO report "Preventing chronic diseases: A vital investment" and us. Technical Report. 2006. Geneva, Switzerland: World Health Organization.
5. Baicker K, Cutler D and Song Z. Workplace wellness programs can generate savings. *Health Aff (Millwood)* 2010; **29**(2): 304–311.
6. Mattke S, Liu H, Caloyeras J, et al. Workplace wellness programs study: Final report. Technical Report. 2013. Santa Monica, California: Rand Corporation.
7. Walker G and Habboushe J. Medical calculators, equations, algorithms, and scores, 2005, www.mdcalc.com (accessed 1 June 2014).
8. Baxter S, Sanderson K, Venn AJ, et al. The relationship between return on investment and quality of study methodology in workplace health promotion programs. *Am J Health Promot* 2014; **28**(6): 347–363.
9. Caloyeras JP, Liu H, Exum E, et al. Managing manifest diseases, but not health risks, saved PepsiCo money over seven years. *Health Aff (Millwood)* 2014; **33**(1): 124–131.
10. Khosla A, Cao Y, Lin CC, et al. An integrated machine learning approach to stroke prediction. In: *Proceedings of the 16th international conference on knowledge discovery and data mining*, Washington, DC, USA, 25–28 July 2010, pp.183–192.
11. Weiss J, Natarajan S, Peissig P, et al. Machine learning for personalized medicine: Predicting primary myocardial infarction from electronic health records. *AI Magazine* 2012; **33**(4): 33–45.
12. Cruz JA and Wishart DS. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform* 2006; **2**: 59–77.
13. Nilsson J, Ohlsson M, Thulin L, et al. Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks. *J Thorac Cardiovasc Surg* 2006; **132**(1): 12–19.
14. Saria S, Rajani AK, Gould J, et al. Integration of early physiological responses predicts later illness severity in preterm infants. *Sci Transl Med* 2010; **2**(48): 48ra65 Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3564961/.
15. Syed Z, Leeds D, Curtis D, et al. A framework for the analysis of acoustical cardiac signals. *IEEE Trans Biomed Eng* 2007; **54**(4): 651–662.
16. Wiens J, Campbell WN, Franklin ES, et al. Learning data-driven patient risk stratification models for clostridium difficile. *Open Forum Infect Dis* 2014; **1**(2): ofu045. Available at: http://ofid.oxfordjournals.org/content/early/2014/06/18/ofid.ofu045.abstract.
17. Bayati M, Braverman M, Gillam M, et al. Data-driven decisions for reducing readmissions for heart failure: General methodology and case study. *PLoS ONE* 2014; **9**(10): e109264. Available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0109264.
18. Meier L, van de Geer S and Bühlmann P. The group lasso for logistic regression. *J R Stat Soc Series B Stat Methodol* 2008; **70**: 53–71.
19. Caruana R. Multitask learning. *Mach Learn* 1997; **28**(1): 41–75.
20. Pan S and Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2010; **22**(10): 1345–1359.
21. Wang X, Wang F and Hu J. A multi-task learning framework for joint disease risk prediction and comorbidity discovery. In: *Proceedings of 22nd international conference on pattern recognition*, Stockholm, Sweden, 24–28 August 2014, pp.220–225. New Jersey, USA: IEEE.
22. Tibshirani R. Regression shrinkage and selection with the lasso. *J R Stat Soc Series B Stat Methodol* 1996; **58**: 267–288.
23. Chen S, Donoho D and Saunders M. Atomic decomposition by basis pursuit. *SIAM J Sci Comput* 1998; **20**(1): 33–61.
24. Ando R and Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data. *J Mach Learn Res* 2005; **6**: 1817–1853.
25. Argyriou A, Evgeniou A and Pontil M. Multi-task feature learning. *Adv Neural Inf Process Syst* 2006; **19**: 41–48.
26. Dartmouth A. List of icd-9-cm codes by chronic disease category, http://www.dartmouthatlas.org/downloads/methods/chronic_disease_codes_2008.pdf (2008, accessed 1 June 2014).
27. Iezzoni L, Heeren T, Foley S, et al. Chronic conditions and risk of in-hospital death. *IEEE Trans Biomed Eng* 1994; **29**(4): 435–460.
28. Kaggle. Practice fusion diabetes classification, www.kaggle.com/c/pf2012-diabetes (2012, accessed 15 November 2013).
29. van Buuren S and Oudshoorn K. *Multivariate imputation by chained equations. MICE V1.0 user's manual*. The Netherlands: TNO Prevention and Health, 2000.
30. Schafer JL and Graham JW. Missing data: Our view of the state of the art. *Psychol Methods* 2002; **7**(2): 147–177.
31. Azur MJ, Stuart EA, Frangakis C, et al. Multiple imputation by chained equations: What is it and how does it work? *Int J Methods Psychiatr Res* 2011; **20**(1): 40–49.
32. Healthone. Lab tests data at healthone labs 2010, www.healthonelabs.com/ (accessed 1 June 2014).
33. Mark S. Function for unconstrained optimization of differentiable real-valued multivariate functions 2005, www.cs.ubc.ca/schmidtm/Software/minFunc.html (accessed 1 June 2014).

34. Birgin E, Martínez J and Raydan M. Nonmonotone spectral projected gradient methods on convex sets. *SIAM J Optim* 2000; **10**(4): 1196–1211.
35. Hastie T, Tibshirani R and Friedman J. *The elements of statistical learning*. New York: Springer, 2001.
36. Hanley JA and McNeil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 1982; **143**(1): 29–36.
37. Efron B and Tibshirani R. *An introduction to the bootstrap*. Boca Raton, USA: Chapman et Hall, 1993.
38. DeLong E, DeLong D and Clarke–Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 1988; **44**: 837–845.
39. Bayati M, Bhaskar S and Montanari A. Instructions and source codes for statistical analysis of a low cost method for multiple disease prediction 2016, https://github.com/mohsenbayati/lowcostprediction (accessed 21 October 2016).
40. Bitton A and Gaziano TA. The framingham heart study's impact on global risk assessment. *Prog Cardiovasc Dis* 2010; **53**(1): 68–78.
41. Hartwig S, Kuss O, Tiller D, et al. Validation of the German diabetes risk score within a population-based representative cohort. *Diabet Med* 2013; **30**(9): 1047–1053.
42. Chatzicostas C, Roussomoustakaki M, Notas G, et al. A comparison of Child–Pugh, APACHE II and APACHE III scoring systems in predicting hospital mortality of patients with liver cirrhosis. *BMC Gastroenterol* 2003; **3**: 7.