HHS Public Access
Author manuscript
*Methods Inf Med*. Author manuscript; available in PMC 2018 August 11.

Published in final edited form as:
*Methods Inf Med*. 2017 August 11; 56(4): 294–307. doi:10.3414/ME16-01-0112.

# A Multi-Way Multi-Task Learning Approach for Multinomial Logistic Regression: An Application in Joint Prediction of Appointment Miss-Opportunities across Multiple Clinics

**A. Alaeddini*** and **S. H. Hong**
Department of Mechanical Engineering, University of Texas at San Antonio, San Antonio, TX, 78249

## Summary

**Objectives—**Whether they have been engineered for it or not, most healthcare systems experience a variety of unexpected events such as appointment miss-opportunities that can have significant impact on their revenue, cost and resource utilization. In this paper, a multi-way multi-task learning model based on multinomial logistic regression is proposed to jointly predict the occurrence of different types of miss-opportunities at multiple clinics.

**Methods—**An extension of $L_1/L_2$ regularization is proposed to enable transfer of information among various types of miss-opportunities as well as different clinics. A proximal algorithm is developed to transform the convex but non-smooth likelihood function of the multi-way multi-task learning model into a convex and smooth optimization problem solvable using gradient descent algorithm.

**Results—**A dataset of real attendance records of patients at four different clinics of a VA medical center is used to verify the performance of the proposed multi-task learning approach. Additionally, a simulation study, investigating more general data situations is provided to highlight the specific aspects of the proposed approach. Various individual and integrated multinomial logistic regression models with/without LASSO penalty along with a number of other common classification algorithms are fitted and compared against the proposed multi-way multi-task learning approach. Fivefold cross validation is used to estimate comparing models parameters and their predictive accuracy. The multi-way multi-task learning framework enables the proposed approach to achieve a considerable rate of parameter shrinkage and superior prediction accuracy across various types of miss-opportunities and clinics.

**Conclusions—**The proposed approach provides an integrated structure to effectively transfer knowledge among different miss-opportunities and clinics to reduce model size, increase estimation efficacy, and more importantly improve predictions results. The proposed framework can be effectively applied to medical centers with multiple clinics, especially those suffering from information scarcity on some type of disruptions and/or clinics.

*Corresponding Author. Tel.: +1 210.458.8747; Fax:+1 210.458.6504. adel.alaeddini@utsa.edu.

**Keywords**

Multinomial logistic regression; Multi-way multi-task learning; Regularization; Miss-opportunities prediction; Proximal gradient descent

## 1. Introduction

The problem of miss-opportunities for scheduled appointments creates significant disruption in the smooth operation of almost all healthcare systems. For example, when patients do not show-up for their appointments or show up late, resources will be underutilized and other patients cannot get timely access to the care. Also, when scheduled patients cancel their appointments, they reduce the system (available) time to fill the schedule, which can cause even more problems than no-show. While, overbooking may help to some extent in these circumstances, it can result in congestion and patient frustration. Indeed, miss-opportunities negative effects on healthcare operations efficiency, patient satisfaction, and clinic costs, can reach hundreds of thousands of dollars yearly [1–5]. Hence, accurate prediction of no-show and cancellation probability is a cornerstone for any scheduling system and miss-opportunity reduction strategy [2, 5–9].

This paper presents a multi-way multi-task learning methodology based on multinomial logistic regression and an extension of $L_1/L_2$ regularization to jointly predict two major types of disruptions, namely no-show and cancellation across different clinics of medical centers, while transferring the information among miss-opportunities and clinics to improve the prediction results. The proposed approach can result in better appointment scheduling systems and more effective miss-opportunity reduction strategies [10–17].

The remainder of the paper is organized as follows: Section 2 discusses the related works in the literature. Section 3 explains various pieces of the proposed model. Section 4 evaluates the performance of the proposed prediction approach based on a real healthcare dataset from a VA medical center. Section 5 discusses the major features and some of the challenges using the proposed approach. Finally, Section 6 summarizes the experimental results and potential applications of the proposed approach.

## 2. Relevant background and problem formulation

While very few researchers consider the prediction of cancellation [10], numerous studies focus on no-show [18–22]. The probability of no-show in the literature fluctuates considerably for different diagnoses and demographics, stretching from almost zero up to 64% [3, 23–25]. Various reasons are reported for no-show ranging from conflict with other appointment to lack of transportation to forgetting the appointments [10, 26–30]. Several contributing factors are also associated with non-attendance behavior [9, 17, 31–34]. Some studies discuss the relationship between miss-opportunities and some specific health outcomes including primary care, psychiatric, and diabetes [35–39]. Several papers also discuss strategies to reduce appointment no-shows [40–47]. Meanwhile, the effectiveness of most of miss-opportunities reduction strategies depends on the accurate prediction of miss-opportunities in individual patients [9].

From a methodological perspective, most miss-opportunity prediction models use logistic regression [22]. There have also been researches based on classification and regression trees and rule-based [48]. Neural networks, Support Vector Machines (SVM), and Boosting methods are other data mining techniques that have been used for predicting miss-opportunities [10, 49]. Time series methods, filtering and smoothing algorithms are other examples of methods built on individual patients non/attendance records [10, 49].

While at least some of the existing approaches in the literature are shown to have acceptable performance in predicting a single type of miss-opportunity, most of them are not effective in modeling one or more of the following scenarios that occur commonly in healthcare: (1) multiple types of disruptions, i.e. no-show and cancellation, that are to be considered simultaneously, (2) multiple clinics in a medical center with significant miss-opportunities problem, (3) data/information scarcity for some types of disruptions and/or clinics, (4) too many (potential) factors suspected to affect miss-opportunities. The major problem with the existing approaches in dealing with above mentioned scenarios is that they cannot effectively transfer the information/knowledge from one type of disruption to another in order to improve the prediction results and help sparsity. Neither can they use the information/knowledge gained from one clinic (population) to advance the inference on the other/s. In the next section, a multi-way multi-task learning methodology is developed based on multinomial logistic regression and an extension of $L_1/L_2$ regularization to jointly predict two major types of disruptions, namely no-shows, and cancellations while transferring the information among various types of miss-opportunities as well as different clinics to improve the predictions.

## 3. The proposed multi-way multi-task learning

The proposed multi-way multi-task learning model is composed of three major components: (1) a multinomial logistic regression for joint estimation of no-show and cancellation, (2) a multi-way multi-task learning approach for transferring knowledge among different miss-opportunities and clinics, and (3) a proximal gradient descent algorithm for estimating the parameters of the proposed model (See Figure 1).

### 3.1 A multinomial logistic regression for joint estimation of no-show and cancellation

The proposed approach begins with developing a multinomial logistic regression for joint estimation of no-show and cancellation based on general social and demographical information of clinical population. Let $Y$ denotes an appointment outcome (attendance, no-show, cancellation) in a clinic, i.e. mental health ophthalmology, dental care, and dermatology, and $X$ denotes the associated risk factors (explanatory variables) including: gender, age, marital status, medical service coverage, distance to medical center, appointment time, and appointment weekday (See also Table 1). The multinomial logistic regression $F(X, B_k)$ then takes the form:

$$\begin{cases} P_k = \frac{\exp(XB_k)}{1+\sum_{k=1}^{K}\exp(XB_k)}, & k=1,2 \\ P_0 = \frac{1}{1+\sum_{k=1}^{K}\exp(XB_k)} \end{cases} \quad (1)$$

where $P_k$ denotes the $k^{\text{th}}$ event probability, and $k$ is in the index of the possible outcome (attendance, no-show, cancellation) and that $k$=0,1,2.

To learn the parameters of the multinomial logistic regression one can use a negated log-likelihood estimator $L(\mathbf{B})$, where $\mathbf{B}=[B_1, \ldots, B_K]^T$ is a $K \times P$ matrix ( $K=2$), whose $k^{\text{th}}$ row corresponds to the coefficients for the $k^{\text{th}}$ miss-opportunity (no-show, cancellation), and $p^{\text{th}}$ column corresponds to the coefficients of the $p^{\text{th}}$ factor in the model. In that case, the negated log-likelihood estimator for all the miss-opportunity probabilities is given by:

$$L(\mathbf{B}) = -log\left\{\prod_{k=1}^{K}\left\{\prod_{n=1}^{N}P\left(y_{kn}|X_n,\ B_k\right)\right\}\right\} = -\left\{\sum_{n=1}^{N}\left\{\sum_{k=1}^{K}y_{kn}X_nB_k - \log\sum_{k=1}^{K}\exp\left(X_nB_k\right)\right\}\right\} \quad (2)$$

To achieve a sparse solution to facilitate the interpretation of the model, a regularization term such as $\mathbf{B}_2$ can be considered in the model, which changes Equation (2) to the following loss function.

$$H(\mathbf{B}) = L(\mathbf{B}) + \lambda\sum_{k=1}^{K}\|B_k\|_2 \quad (3)$$

The second term in the above equation indicates a $L_2$ penalty that encourage sparsity, $B_{k2} = \sqrt{\sum_{p=1}^{P}\beta_{kp}^2}$, and $\lambda \geq 0$ is a tuning parameter that control the amount of shrinkage [50]. This optimization problem can be solved by gradient descent algorithm or any other appropriated mean.

### 3.2. A multi-way multi-task learning approach for transferring knowledge among different miss-opportunities and clinics

The probabilities of no-show and cancellation resulted from the proposed multinomial logistic regression are related and this information can be used to further to improve the predictions. This motivates the introduction of the multi-task learning paradigm that exploits the correlations amongst different miss-opportunities by learning them simultaneously rather than individually [51]. Several recent studies on multi-task learning have shown that the accuracy of predictive models can be improved when the tasks, e.g. classification problems, are related [52–55].

Exploiting the correlations amongst multiple tasks can be achieved by developing a group regularization penalty that enforces learning multinomial logistic regression for no-show and cancellation simultaneously rather than individually.

Reconsidering matrix $[\mathbf{B}]_{K \times P}$ of the multinomial logistic regression coefficients in Equations (1–2), Then a $L_1/L_2$ group penalty can be defined as follows:

$$\|\mathbf{B}\|_{L_1/L_2} = \sum_{p=1}^{P} \|B_p\|_2 \quad (4)$$

where $B_{p2} = \sqrt{\sum_{k=1}^{K} \beta_{kp}^2}$. In this case, $L_1$ penalty is applied over the $L_2$ norms of the vectors of regression coefficients, rather than individual elements of regression coefficients in regularization. Using this penalty, $L_1/L_2$ regularized multinomial logistic regression for joint prediction of multiple miss-opportunities estimates the matrix of regression coefficients $(\mathbf{B})$ by solving the following optimization problem:

$$\min_{\mathbf{B}} H(\mathbf{B}) = L(\mathbf{B}) + \lambda \|\mathbf{B}\|_{L_1/L_2} \quad (5)$$

where λ is the tuning parameter that determines the magnitude of penalization. Equation (5) can be viewed as a special group LASSO for multinomial logistic regression [56–58], in which groups are defined over individual risk factors, e.g. gender, at different miss-opportunities, namely no-show and cancellation. Consequently, the $L_1/L_2$ penalization shrinks the regression coefficients $\beta_p$ for the $p^{\text{th}}$ risk factor across all miss-opportunities to zero jointly, if that factor is not associated with any miss-opportunity, thus reducing the number of false positives. Conversely, if the factor is relevant to one or more miss-opportunities, i.e. no-show or cancellation, all elements in $\beta_p$ will be jointly set to have non-zero values, but the $L_2$ norm still allows the association strengths to be different across different types of miss-opportunities for the $p^{\text{th}}$ risk factor. Thus, the joint inference made by the $L_1/L_2$ penalty enables us to infer association between the risk factors by borrowing strength across different miss-opportunities and setting the corresponding regression coefficients jointly to non-zero values. One may notice that a large value of λ will set more columns $\beta_p$'s of $\mathbf{B}$ to zero.

In addition to the correlation among different types of miss-opportunities, there also exists correlation among specific miss-opportunities across various clinics in a medical center, i.e. mental health ophthalmology, dental care, and dermatology, that can be exploited to further improve the predictions. Such generalization will extend the matrix $\mathbf{B}$ of the multinomial logistic regression coefficients into a three dimensional array $[\mathbf{B}]_{K \times P \times H}$ where the $h^{\text{th}}$ dimension corresponds to the regression coefficients for the $h^{\text{th}}$ clinic (population) while $p$ and $k$ interpretation stays the same as before (See Figure 2).

Given the structure of the three dimensional array $[\mathbf{B}]_{K\times P\times H}$, an additional group regularization term needs be added to Equation (6) to capture the correlation among clinics.

$$\min_{\mathbf{B}} H(B) = L(\mathbf{B}) + \lambda_1 \sum_{h=1}^{H}\sum_{p=1}^{P} \|B_{ph}\|_2 + \lambda_2 \sum_{p=1}^{P}\sum_{k=1}^{K} \|B_{kp}\|_2 \quad (6)$$

where $\|B_{ph}\|_2 = \sqrt{\sum_{k=1}^{K} \beta_{kph}^2}$, and $\|B_{ph}\|_2 = \sqrt{\sum_{h=1}^{H} \beta_{kph}^2}$. The second and third terms of Equation (6) are $L_1/L_2$ regularization terms for exploiting the correlation among different miss-opportunities and clinics and influence different dimensions of the array $\mathbf{B}$ of regression parameters. Similar to Equation (5), Equation (6) can be viewed as a special group LASSO for multinomial logistic regression [56–58], in which two sets of groupings are defined over: (1) individual risk factors across different clinics, and (2) individual risk factors across different miss-opportunities. This multi-way regularization of the multinomial logistic regression likelihood function for joint learning of multiple tasks (miss-opportunities), i.e. no-show and cancelation, at multiple populations (clinics), i.e. mental health ophthalmology, dental care, and dermatology, etc. is referred to as "multi-way multi task learning" in this paper (See also Figure 3).

Various block-structured norms in the form of $L_1/L_q$, $q>0$, can be used to combine information from multiple tasks. For instance, in group LASSO, the grouping structure of the risk factors is assumed to be known, and the $L_q$ part of the $L_1/L_q$ norm is defined over regression coefficients for the members in each group, so that they are jointly selected or deselected. In multi-response regression, the $L_q$ part of the $L_1/L_q$ norm acts over the regression coefficients for all responses associated with each risk factor, and a risk factor is selected to be jointly influencing all of the responses. The proposed use of $L_1/L_2$ norm differs from these previous methods in that: (1) it takes advantage of the grouping structure among the risk factors/responses as well as samples, wherein each group corresponds to a miss-opportunity and/or clinic, and (2) it applies the $L_1/L_2$ regularization in a multi-way (two-way) rather than one way to affect different dimensions of array $\mathbf{B}$.

Obozinski et al. [59] found that for $k$ regressions, under certain conditions, the sample complexity for $L_1/L_2$, is up to $k$ times smaller than the LASSO sample complexity, with weak assumptions of shared support. Thus, we can expect under certain conditions, the proposed multi-way (two-way) $\mathrm{L}_1/\mathrm{L}_2$ app will approximately require up to $k/2$ times fewer samples than LASSO to obtain the correct set of regression coefficients.

### 3.3. Parameter estimation

The regression coefficients $\mathbf{B}$ are estimated by solving the optimization problem in Equation (6). The first tem of Equation (6), the log-likelihood of multinomial logistic regression, is a smooth convex function which can be easily minimized using first-order gradient algorithms. However, the second and third terms of Equation (6), $L_1/L_2$ penalties, are convex but not smooth at zero, and for this reason, methods based on the first-order gradient cannot be used directly for optimizing it. Here, a variation of Fast Iterative Thresholding Algorithm (FISTA) [60] is proposed to address the above problem. FISTA is an extension of proximal gradient descent algorithms for non-smooth functions where proximal gradient step is performed on the extrapolated point based on the previous two iterates to improve the convergence rate of the algorithm [61–65]. It should be noted that, instead of FISTA, one may use global optimization algorithms, such as simulated annealing [66] or pattern search [67], which don't require transformation of non-smooth function. In our simulation studies, these algorithms provide slightly less accurate estimates, but better computational complexity, which make them suitable for large scale problems.

**FISTA for multi-way multi-task learning multiple logistic regression—**Let $H$ represents the total number of clinics and $N_h$ denotes the number of patients in clinic *h*, with $N_h \in \{N_1, .., N_H\}$, $\sum_{h=1}^{H} N_h = N$; FISTA begins with rewriting the optimization problem in Equation (6) as:

$$\min_{\mathbf{B}} H(\mathbf{B}) = L(\mathbf{B}) + G(\mathbf{B}), \quad (7.a)$$

$$L(\mathbf{B}) = -\sum_{h=1}^{H}\left\{\sum_{n_h=1}^{N_h}\left\{\sum_{k=1}^{K} y_{kn_h} X_{N_h} B_{kh} - \log\left(1 + \sum_{k=1}^{K}\exp\left(X_{N_h} B_{kh}\right)\right)\right\}\right\} \quad (7.b)$$

$$\mathrm{G}(\mathbf{B}) = \lambda_1 \sum_{h=1}^{H}\sum_{p=1}^{P}\|B_{ph}\|_2 + \lambda_2 \sum_{p=1}^{P}\sum_{k=1}^{K}\|B_{pk}\|_2 \quad (7.c)$$

where $L(\mathbf{B})$ is the negated log-likelihood of multinomial regression for multiple clinics $h = \{1, \ldots, H\}$ which is a continuously differentiable with Lipschitz continuous gradient, and $G(\mathbf{B})$ is a set of continuous non-smooth convex group LASSO penalties with inexpensive prox-operator.

FISTA algorithm solves Equation (7) by generating a sequence $\mathbf{B}^{(m)}$ via:

$$\mathbf{B}^{*} = \mathbf{B}^{(\boldsymbol{m-1})} + \frac{m-2}{m+1}\left(\mathbf{B}^{(\boldsymbol{m-1})} - \mathbf{B}^{(\boldsymbol{m-2})}\right) \quad (8.a)$$

$$\mathbf{B}^{(m)} = \mathrm{prox}_{sg}\left(\mathbf{B}^{*} - s\nabla L(\mathbf{B}^{*})\right) \quad (8.b)$$

where the first step is an extrapolation step and the second step is the proximal gradient descent, and $s$ is the step size determined by the line search. In the remainder of this section, first proximal gradient descent and line search criterion are briefly explained and then integrated into the final algorithm for parameter estimation.

**Proximal gradient descent (PGD)—**PGD minimizes $G(B)$ plus a simple quadratic local model of $L(B)$ around $\mathbf{B}^*$ as follows:

$$\mathbf{B}^{(m)} = \mathrm{prox}_{sg}\left(\mathbf{B}^* - s\nabla L(\mathbf{B}^*)\right) \quad (9.a)$$

$$= \mathrm{argmin}_u \left( G(\mathbf{u}) + \frac{1}{2s} \| \mathbf{u} - \mathbf{B}^* + s\nabla L(\mathbf{B}^*) \|_2^2 \right) \quad (9.b)$$

$$= \mathrm{argmin}_u \left( G(\mathbf{u}) + L(\mathbf{B}^*) + \nabla L(\mathbf{B}^*)\ (\mathbf{u} - \mathbf{B}^*) + \frac{1}{2s} \| \mathbf{u} - \mathbf{B}^* \|_2^2 \right) \quad (9.c)$$

where $s$ the step size, which can be determined using line search. In Equation (9), $L(\mathbf{B})$, the negated log-likelihood of the multinomial logistic regression, is a differentiable function with gradient:

$$\nabla L(\mathbf{B}^*) = \begin{bmatrix} \left[\frac{\partial L(\mathbf{B}^*)}{\partial \beta_{111}}, \ldots, \frac{\partial L(\mathbf{B}^*)}{\partial \beta_{1P1}}\right] & \cdots & \left[\frac{\partial L(\mathbf{B}^*)}{\partial \beta_{11H}}, \ldots, \frac{\partial L(\mathbf{B}^*)}{\partial \beta_{1PH}}\right] \\ \vdots & \ddots & \vdots \\ \left[\frac{\partial L(\mathbf{B}^*)}{\partial \beta_{K11}}, \ldots, \frac{\partial L(\mathbf{B}^*)}{\partial \beta_{KP1}}\right] & \cdots & \left[\frac{\partial L(\mathbf{B}^*)}{\partial \beta_{K1H}}, \ldots, \frac{\partial L(\mathbf{B}^*)}{\partial \beta_{KPH}}\right] \end{bmatrix} \quad (10.a)$$

$$\frac{\partial L(\mathbf{B}^*)}{\partial \beta_{kph}} = -\sum_{n_h=1}^{N_h} \left\{ y_{khn_h} x_{pn_h} - \frac{\exp\left(\sum_{q=1}^{P} x_{qn_h}\beta_{kqh}\right)}{1+\exp\left(\sum_{q=1}^{P} x_{qn_h}\beta_{kqh}\right)} . x_{pn_h} \right\} \quad (10.b)$$

From sub-gradient definition of $\mathrm{prox}_{sg}$, $G(\mathbf{u})=0$, if and only if $\mathbf{B}$ minimizes $L(\mathbf{B}) + G(\mathbf{B})$.

**Line search criterion—**One natural and commonly used line search criterion is to require that the objective function value is monotonically decreasing. More specifically, one may propose to accept the step size $s^{(m)}$ at the outer iteration $k$ if the following monotone line search criterion is satisfied:

$$G\left(\mathbf{B}^{(m+1)}\right) \leq G\left(\mathbf{B}^{(m)}\right) - \frac{\sigma}{2} s^{(m)} \mathbf{B}^{(m+1)} - \mathbf{B}^{(m)2} \quad (11)$$

where $\boldsymbol{\sigma}$ is a constant in the interval $\sigma \in (0, 1)$ [68].

**Proposed FISTA algorithm—**Given the proximal gradient descent and line search criterion, the FISTA steps can be summarized in Algorithm 1. **Step 1** sets the parameters of the algorithm. **Step 2** initializes the starting point $\mathbf{B}$ and the outer loop counter of the algorithm. **Steps 3-10** represent the outer loop of the algorithm with **Step 4** initializing the step size. A good step size is critical for the fast convergence of the algorithm. In this paper, we propose to initialize the step size by adopting the Barzilai-Borwein (BB) rule [69], which uses a diagonal matrix $\mathrm{s}^{(m)} I$ to approximate the Hessian matrix $\nabla^2 L\left(\mathbf{B}^{(m)}\right)$ as follows:

$$\mathrm{s}^{(m)} = \frac{\left(\mathbf{B}^{(m)} - \mathbf{B}^{(m-1)}\right)^{Tr} \left(\nabla l\left(\mathbf{B}^{(m)}\right) - \nabla l\left(\mathbf{B}^{(m-1)}\right)\right)}{\left(\mathbf{B}^{(m)} - \mathbf{B}^{(m-1)}\right)^{Tr} \left(\mathbf{B}^{(m)} - \mathbf{B}^{(m-1)}\right)} \quad (12)$$

wherein $\mathbf{B}^{(m)}$ should be in the vectorized form. **Steps 5-8** constitute the inner loop of the algorithm with the proximal gradient descent in **Step 6**, step size updation in **Step 7**, and line search criterion check in **Step 8**. Finally **Step 9-10** increase the counter and check the stopping criterion of the outer iteration.

**Algorithm 1**

FISTA for Multi-Way Multi-Task Learning Multinomial Logistic Reg.

---

**1.** Choose parameters $\eta > 1$ and $s_{min},\ s_{max}$ with $0 < s_{min}, < s_{max}$

**2.** Initialize iteration counter $m \leftarrow 0$ and a bounded $\mathbf{B}^{(0)}$

**3. Repeat**

**4.** $\mathrm{s}^{(m)} \leftarrow$ Equation (12)

**5.** **Repeat**

**6.** $\mathbf{B}^{(m)} \leftarrow$ Equation (9)

**7.** $\mathrm{s}^{(m)} \leftarrow \eta \mathrm{s}^{(m)}$

**8.** **Until** Equation (11) satisfies

**9.** $m \leftarrow m+1$

**10. Until** some stopping criterion satisfied

---

# 4. Experimental results

In this section, first using a real dataset of attendance records at a VA medical center, the performance of the proposed multi-way multi-task learning approach is studied in comparison with classical and LASSO regularized multinomial logistic regression as well as a number of other predictive models including support vector machine, random forest, Bayesian network, and multilayer perceptron. Next, a simulation study, investigating more general data situations is provided to highlight the specific aspects of the proposed approach. Matlab (R2016.a) has been used for developing the computer code, which is published available as Online Appendix 1.

## 4.1. Real-world case study

The real-world case study is based on a set of 410 attendance records to predict no-show and cancellation events at four clinics, i.e. mental health, ophthalmology, dental care, and dermatology at a Veteran Affairs (VA) Medical Center (See Table 1).

Regarding the size of dataset, using a small dataset such as the one considered for this research can better demonstrate the performance of the proposed framework in the presence of data scarcity, which is a common problem in some healthcare areas [70, 71]. In addition, while it might be more desirable to have a larger dataset, acquiring a dataset containing multiple clinics with some common patients is challenging.

**4.1.1. Comparison with classical and LASSO regularized multinomial logistic regression—**Given the dataset of patient information, various models of multinomial logistic regression are fitted and compared against proposed multi-way multi-task learning approach. These methods include individually fitted multinomial logistic regression for each clinic with/without regularization, and integrated multinomial logistic regression wherein clinics are incorporated as a 3-level risk factor into the model with/without regularization (See Table 2). LASSO penalty is used for regularized models, while $L_1/L_2$ regularization is considered for transferring information among both clinics and miss-opportunities in the proposed approach. As shown in Table 2, from the sparsity stand point, the proposed multi-way multi-task learning model achieves a higher shrinkage ratio compared to the other models, because of regularizing the parameters in two directions, i.e. miss-opportunities and clinics.

Five-fold cross validation [72] with three training subset, one validation subset and one test subset in each of the five repetitions is used to train the parameters of comparing models, optimize the tuning parameters, and test their predictive accuracy. Figure 4 illustrates the contour plot of the validation Mean Squared Error (MSE) for the proposed multi-way multi-task learning approach as function of tuning parameters $(\lambda_1, \lambda_2)$. The large number of parameters in the model inflates the shrinkage penalty terms $(\mathrm{G}(\mathbf{B}))$, and therefore MSE can be minimized over a relatively small range of tuning parameters with a minimum at $(\lambda_1^*, \lambda_2^*) = (0.1, 0.1)$.

Figure 1 in Online Appendix 2 shows the effect of varying tuning parameters $(\lambda_1, \lambda_2)$ on the estimated parameters (excluding the intercept) and sparsity of the proposed multi-way multi-task learning approach for mental clinic. At the optimal level of tuning parameters, i.e. $(\lambda_1^*, \lambda_2^*) = (0.1, 0.1)$, the estimated parameters can be roughly grouped into three categories: (1) parameters shrunk to zero which show deselected risk factors, such as late afternoon, (2) parameters associated with risk factors having relatively small effect on the model outcome, such as gender, and finally (3) parameters associated with risk factors having significant effect on the outputs, such as distance to medical center. Similar grouping behaviors are also observed across other clinics and miss-opportunities. Another observation is similar shrinkage behavior of some of the risk factors across different miss-opportunities and/or clinics, such as "being divorced" which turn out to be insignificant for predicting no-shows at mental, ophthalmology, and dental care.

Table 3 illustrates the estimated parameters of the proposed multi-way multi-task learning approach for the optimal level tuning parameters, i.e. $(\lambda_1^*, \lambda_2^*) = (0.1, 0.1)$. The specific structure of $L_1/L_2$ regularization exploits the correlation among different miss-opportunities and different clinics for influencing different dimensions of the array $\mathbf{B}$ of regression parameters. This enables automatic individual and group selection/deselection of risk factors across both clinics and miss-opportunities.

In addition, Table 3 shows correlation among selected risk factors across different types of miss-opportunities, i.e. no-show and cancellation. It also shows some degrees of similarity among selected risk factors across different clinics, i.e. dental care and dermatology clinics. Such observation supports the idea (hypothesis) of existence of association among risk factors, miss-opportunities and different clinics. Therefore, a multi-way multi-task learning approach that elicits such relationship can indeed advance the predictive accuracy of the results.

Figures 5.a and 5.b demonstrate the correct classification rate of the comparing methods across the four clinics of the study, i.e. mental health, ophthalmology, dental care, dermatology, for predicting no-show and cancellation events. As shown in the Figure the proposed approach performs superior to its counterparts across all clinics and miss-opportunities. Meanwhile there is not much of a difference among the performance of other individual- and integrated- multinomial logistic models expect for the individual multinomial logistic model on predicting no-shows in clinic 3, i.e. dental care. In addition, the models with LASSO penalty perform on a par with the classical methods except for predicting no-show in clinic 4, i.e. dermatology, which has fewer records compared to other clinics.

Figure 2 in Online Appendix 2 provides the Receiver Operating Characteristic (ROC) curves of the comparing methods performance for no-show (blue), cancellation (green) and show-up (red) predictions across four clinics of the study. Similar to Figure 5 results, the proposed multi-way multi-task learning approach shows a consistent superior performance across all types of miss-opportunities and clinics compared to classical methods. Meanwhile, the difference in the predictive performance of the proposed approach gets considerably more distinctive for mental and ophthalmology clinics.

In addition, while for some classical methods the results supports better accuracy for prediction of show-up and no-show events compared to cancelation event, the results are more consistent across different types of miss-opportunities for the proposed approach. As discussed earlier in the paper, such improvement in the predictive performance is largely because of the information transfer between miss-opportunities and clinics using the specific structure of $L_1/L_2$ regularization.

**4.1.2. Comparison with other predictive models—**Here, the performance of the proposed approach is compared against a number of other common classification algorithms, the details of which is provided in Table 4. Like earlier analysis all methods are tuned and evaluated using the same dataset described in Section 4.1 and based on five-fold cross validation.

Figures 6.a and 6.b illustrate the correct classification rate of the comparing methods across the four clinics of the study for predicting no-show and cancellation events. With an exception of predicting cancellations in clinic 4, i.e. dermatology, the proposed approach performs uniformly better than all other methods across all clinics and miss-opportunities. After that, support vector machine, Bayesian network and random forest demonstrate comparable performances. However, MLP does not compare well with the other classifiers for none of the clinics or miss-opportunities.

Finally, Figure 3 in Online Appendix 2 provides the ROC curves of the comparing methods for predicting no-show (blue), cancellation (green) and show-up (red) predictions across four clinics of the study. Similar to Figure 6, the proposed multi-way multi-task learning approach demonstrate a consistent superior performance across most of miss-opportunities and clinics.

### 4.2. Simulation study

The simulation study is based on a set of 3,000 computer generated records $(N=3,000)$ to predict 4 types of events $(k=0,1,2,3)$, i.e. show-up, no-show, cancellation, tardiness, and 30 risk factors $(P=30)$, i.e. age, gender, etc. The 3,000 records have been evenly distributed among 6 population groups $(H=6)$, i.e. mental health, ophthalmology, etc., such that each population contains 500 records $(N_H=500)$. The simulation dataset consists of three major elements: (1) the three-dimensional matrix of risk factors $[X]_{N_H \times P \times H}$, (2) the three-dimensional matrix of regression coefficients $[\mathbf{B}]_{K \times P \times H}$, and (3) the matrix of observed events (responses) $[\mathrm{Y}]_{N_H \times H}$. The matrix of risk factors $(X)$ is generated based on independent and identically distributed (i.i.d.) random numbers from uniform distribution $U(0,1)$.

The individual elements of the matrix of regression coefficients $(\mathbf{B})$ are initially generated from normal distribution $N(0,6)$. To add sparsity to the regression coefficients of each population, $Bernoulli\,(0.25)$ distribution is used to randomly select some of the risk factors, an turn the associated regression coefficients to zero across all events in that population, i.e.

$B_{kph}=0,\ (\forall p \in \Omega\ ,\ \ k=0, ..., K)$, where $\Omega$ denotes the index of regression coefficients associated with selected risk factors. To add sparsity to the regression coefficients across all populations, $Bernoulli\ (0.25)$ distribution is used to randomly select some pairs of risk factors and events, then turn the associated regression coefficients to zero across all populations, i.e. $B_{kph}=0,\ (\forall\, k, p \in \Omega\ ,\ \ h=0, \ldots, H)$, where $\Omega$ denotes the index of regression coefficients for selected pairs of risk factors and events.

To add correlation among risk factors at each population, $Bernoulli\ (0.25)$ distribution is used to randomly select some of the risk factors, and scale the regression coefficients associated with each of them to have a sum of one across all events, i.e. $\sum_{k=1}^{K} B_{kph}=1,\ (\forall p \in \Omega\ )$, where $\Omega$ denotes the index of regression coefficients associated with selected risk factors. To add correlation among populations, $Bernoulli\ (0.25)$ distribution is used to randomly select pairs of risk factors and events, and scale the regression coefficients associated with each of the pairs to have a sum of one across all populations, i.e. $\sum_{h=1}^{H} B_{kph}=1,\ (\forall p \in \Omega\ )$, where $\Omega$ denotes the index of regression coefficients for selected pairs of risk factors and events. Figure 4 in Online Appendix 2 shows the heat map of the $\mathbf{B}$ matrix generated using above procedure.

Having $X$ and $\mathbf{B}$ matrices generated, equation (1) is used to calculate the probability of each event occurrence for all records, i.e. $P\ (y_{kn}|X_n,\ B_k)$. These probabilities are then turned into observed responses $(y_{kn})$ using random number generation from $ultinomial\ (p_1, \ldots, p_K)$ distribution.

Next, we use the simulated dataset to evaluate the performance of the proposed multi-way multi-task learning approach in comparison with some of the best performing predictive models in the previous section, including support vector machine, random forest, Bayesian network, and individual multinomial logistic regressions with LASSO regularization for each population. The details of the comparing methods and the analysis procedure are kept the same as the previous section.

Figures 7.a and 7.b demonstrate the correct classification rate of the comparing methods across different populations and events. Similar to the previous section, the proposed approach clearly outperforms all of its counterparts across different populations and events. The proposed approach also provides the most consistent correct classification rates across different populations and events. Meanwhile, there is not much of a difference among the performance of other methods expect for individual multinomial logistic regressions with LASSO for population 6 and event 1. The result of simulation study further demonstrates the improvements made by the proposed multi-way multi-task learning approach by transferring information among events and populations using the specific structure of $L_1/L_2$ Regularization.

## 5. Discussion

Most miss-opportunity prediction models focus on only one type of disruption, i.e. no-show, and one clinical population, i.e. primary care. They also require relatively large datasets for training the model parameters. This paper has focused on developing a joint model for both different types of disruptions and multiple clinics by: (1) Developing a multi-way multi-task learning approach for multinomial logistic regression based on extension of $L_1/L_2$ regularization, which takes into account the inter-relationships among, both various types of miss-opportunities as well as different clinics; and (2) Presenting a proximal algorithm for converting the non-smooth likelihood function of the multi-way multi-task learning model to a smooth optimization problem solvable via gradient descent algorithm. The model not only provides a single and compact framework for simultaneous estimation of no-show and cancellation across multiple clinics, but also offers better prediction accuracy compared to classical logistic regression models. Additionally, the transfer of information across different miss-opportunities and clinics which made possible by $L_1/L_2$ regularization enables the proposed approach to alleviate possible cases of data scarcity for one or more types of disruptions and/or clinics.

Meanwhile, the proposed multi-task learning approach implicitly assumes that no-show and cancellation are associated with similar factors. Nonetheless, while initial analysis has verified such assumption in the dataset of the study, no-show and cancellation may not necessarily have same predictors. The proposed approach also requires identification and optimization of two tuning parameters, i.e. $(\lambda_1^*, \lambda_2^*)$ , using cross validation in order to achieve appropriate level of sparsity and accurately estimate the model parameters. However, because of large size of the shrinkage penalty terms $(\mathrm{G}(\mathrm{B}))$, the tuning parameters $(\lambda_1^*, \lambda_2^*)$ can be efficiently searched over a relatively small area. Finally, the computational complexity of the proposed multi-way multi-task learning algorithm is considerably larger than classical and LASSO based multinomial logistic regression, as it involves optimization of non-smooth likelihood functions. However, the use of proposed FISTA algorithm that converts the $L_1/L_2$ regularized likelihood function into a convex and smooth optimization model can help to reduce the computational efforts and improve the estimated parameters. Besides, global optimization algorithms, such as simulated annealing [66] or pattern search [67], which don't require transformation of non-smooth function like FISTA, can also be considered to further reduce the computational complexity. In our simulation studies, these algorithms provide slightly less accurate estimates, but better computational complexity compared to FISTA, which make them suitable for large scale problems.

## 6. Conclusion and future work

Various types of miss-opportunities such as no-show and cancellation may happen to medical appointments, with significant negative impact on the revenue, cost and resource utilization. This paper proposes a multi-way multi-task learning model based on multinomial logistic regression and an extension of $L_1/L_2$ regularization to jointly predict the probability of different types of miss-opportunities across various clinics. It also presents a proximal algorithm to transform the likelihood function of the multi-way multi-task learning model, into a convex and smooth optimization problem solvable using gradient descent algorithm. Based on real patient dataset, the effectiveness of the proposed approach is demonstrated for estimating no-show and cancellation events at four different clinics of a Veterans Affairs medical center, which includes mental health, ophthalmology, dental care and dermatology. To investigate more general data situations, the performance of the proposed multi-way multi-task learning approach is also compared against a number of common predictive models in the literature using a simulation study. The proposed approach provides an integrated structure to effectively transfer knowledge among different miss-opportunities and clinics to reduce model size, increase estimation efficacy, and more importantly improve predictions results. The proposed framework can be effectively applied to medical centers with multiple clinics, especially those suffering from information scarcity on some type of disruptions and/or clinics.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## References


1. Rust CT, Gallups NH, Clark WS, Jones DS, Wilcox WD. Patient appointment failures in pediatric resident continuity clinics. Archives of pediatrics & adolescent medicine. 1995; 149(6):693–5. [PubMed: 7767428]
2. Moore CG, Wilson-Witherspoon P, Probst JC. Time and money: effects of no-shows at a family practice residency clinic. FAMILY MEDICINE-KANSAS CITY. 2001; 33(7):522–7.
3. Bech M. The economics of non-attendance and the expected effect of charging a fine on non-attendees. Health Policy. 2005; 74(2):181–91. [PubMed: 16153478]
4. Barron WM. Failed appointments. Who misses them, why they are missed, and what can be done. Primary care. 1980; 7(4):563–74. [PubMed: 7010402]
5. Hixon AL, Chapman RW, Nuovo J. Failure to keep clinic appointments: implications for residency education and productivity. FAMILY MEDICINE-KANSAS CITY-. 1999; 31:627–30.
6. Cayirli T, Veral E. OUTPATIENT SCHEDULING IN HEALTH CARE: A REVIEW OF LITERATURE*. Production and Operations Management. 2003; 12(4):519.
7. Cote MJ. Patient flow and resource utilization in an outpatient clinic. Socio-Economic Planning Sciences. 1999; 33(3):231–45.
8. Ho C-J, Lau H-S. Minimizing total cost in scheduling outpatient appointments. Management science. 1992; 38(12):1750–64.
9. Daggy J, Lawley M, Willis D, Thayer D, Suelzer C, DeLaurentis P-C, et al. Using no-show modeling to improve clinic performance. Health Informatics Journal. 2010; 16(4):246–59. [PubMed: 21216805]
10. Alaeddini A, Yang K, Reeves P, Reddy CK. A hybrid prediction model for no-shows and cancellations of outpatient appointments. IIE Transactions on Healthcare Systems Engineering. 2015; 5(1):14–32.
11. Chakraborty S, Muthuraman K, Lawley M. Sequential clinical scheduling with patient no-shows and general service time distributions. IIE Transactions. 2010; 42(5):354–66.
12. Gupta D, Denton B. Appointment scheduling in health care: Challenges and opportunities. IIE transactions. 2008; 40(9):800–19.
13. Hassin R, Mendel S. Scheduling arrivals to queues: A single-server model with no-shows. Management Science. 2008; 54(3):565–72.
14. LaGanga LR, Lawrence SR. Clinic overbooking to improve patient access and increase provider productivity*. Decision Sciences. 2007; 38(2):251–76.

15. Liu N, Ziya S, Kulkarni VG. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. Manufacturing & Service Operations Management. 2010; 12(2):347–64.
16. Muthuraman K, Lawley M. A stochastic overbooking model for outpatient clinical scheduling with no-shows. Iie Transactions. 2008; 40(9):820–37.
17. Zeng B, Turkcan A, Lin J, Lawley M. Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. Annals of Operations Research. 2010; 178(1):121–44.
18. Bowser DM, Utz S, Glick D, Harmon R. A systematic review of the relationship of diabetes mellitus, depression, and missed appointments in a low-income uninsured population. Archives of psychiatric nursing. 2010; 24(5):317–29. [PubMed: 20851323]
19. Denhaerynck K, Manhaeve D, Dobbels F, Garzoni D, Nolte C, De Geest S. Prevalence and consequences of nonadherence to hemodialysis regimens. American Journal of Critical Care. 2007; 16(3):222–35. [PubMed: 17460313]
20. George A, Rubin G. Non-attendance in general practice: a systematic review and its implications for access to primary health care. Family Practice. 2003; 20(2):178–84. [PubMed: 12651793]
21. Rowett M, Reda S, Makhoul S. Prompts to encourage appointment attendance for people with serious mental illness. Schizophrenia bulletin. 2010:sbq050.
22. Turkcan, A., Nuti, L., DeLaurentis, P-C., Tian, Z., Daggy, J., Zhang, L., et al. Handbook of Healthcare Operations Management. Springer; 2013. No-show modeling for adult ambulatory clinics; p. 251-88.
23. Mitchell AJ, Selmes T. A comparative survey of missed initial and follow-up appointments to psychiatric specialties in the United Kingdom. Psychiatric Services. 2007; 58(6):868–71. [PubMed: 17535950]
24. Cayirli T, Veral E, Rosen H. Designing appointment scheduling systems for ambulatory care services. Health care management science. 2006; 9(1):47–58. [PubMed: 16613016]
25. Cayirli T, Veral E, Rosen H. Assessment of patient classification in appointment system design. Production and Operations Management. 2008; 17(3):338–53.
26. Corfield L, Schizas A, Williams A, Noorani A. Non-attendance at the colorectal clinic: a prospective audit. Annals of the Royal College of Surgeons of England. 2008; 90(5):377. [PubMed: 18634730]
27. Gany F, Ramirez J, Chen S, Leng JC. Targeting social and economic correlates of cancer treatment appointment keeping among immigrant Chinese patients. Journal of Urban Health. 2011; 88(1): 98–103. [PubMed: 21246300]
28. Neal RD, Hussain-Gambles M, Allgar VL, Lawlor DA, Dempsey O. Reasons for and consequences of missed appointments in general practice in the UK: questionnaire survey and prospective review of medical records. BMC Family Practice. 2005; 6(1):47. [PubMed: 16274481]
29. Park WB, Kim JY, Kim S-H, Kim HB, Kim NJ, Oh M-D, et al. Self-reported reasons among HIV-infected patients for missing clinic appointments. International journal of STD & AIDS. 2008; 19(2):125–6. [PubMed: 18334069]
30. Sarnquist CC, Soni S, Hwang H, Topol BB, Mutima S, Maldonado YA. Rural HIV-infected women's access to medical care: ongoing needs in California. AIDS care. 2011; 23(7):792–6. [PubMed: 21287418]
31. Cashman SB, Savageau JA, Lemay CA, Ferguson W. Patient health status and appointment keeping in an urban community health center. Journal of health care for the poor and underserved. 2004; 15(3):474–88. [PubMed: 15453182]
32. Alafaireet P, Houghton H, Petroski G, Gong Y, Savage GT. Toward determining the structure of psychiatric visit nonadherence. The Journal of ambulatory care management. 2010; 33(2):108–16. [PubMed: 20228633]
33. Cohen A, Dreiher J, Vardy D, Weitzman D. Nonattendance in a dermatology clinic–a large sample analysis. Journal of the European Academy of Dermatology and Venereology. 2008; 22(10):1178–83. [PubMed: 18393960]
34. Lehmann T, Aebi A, Lehmann D, Olivet MB, Stalder H. Missed appointments at a Swiss university outpatient clinic. Public Health. 2007; 121(10):790–9. [PubMed: 17555782]

35. Schectman JM, Schorling JB, Voss JD. Appointment adherence and disparities in outcomes among patients with diabetes. Journal of general internal medicine. 2008; 23(10):1685–7. [PubMed: 18661189]

36. Murphy K, Edelstein H, Smith L, Clanon K, Schweitzer B, Reynolds L, et al. Treatment of HIV in outpatients with schizophrenia, schizoaffective disorder and bipolar disease at two county clinics. Community mental health journal. 2011; 47(6):668–71. [PubMed: 21384114]

37. Bigby J, Pappius E, Cook EF, Goldman L. Medical consequences of missed appointments. Archives of internal medicine. 1984; 144(6):1163–6. [PubMed: 6732374]

38. Nuti S, Seghieri C, Vainieri M, Zett S. Assessment and improvement of the Italian healthcare system: first evidence from a pilot national performance evaluation system. Journal of Healthcare Management, Forthcoming. 2012

39. Obialo CI, Bashir K, Goring S, Robinson B, Quarshie A, Al-Mahmoud A, et al. Dialysis" No-Show" on Saturdays: Implications of the Weekly Hemodialysis Schedules on Nonadherence and Outcomes. Journal of the National Medical Association. 2008; 100(4):412–9. [PubMed: 18481480]

40. Murray M, Tantau C. Same-day appointments: exploding the access paradigm. Family practice management. 2000; 7(8):45.

41. LaGanga LR. Lean service operations: reflections and new directions for capacity expansion in outpatient clinics. Journal of Operations Management. 2011; 29(5):422–33.

42. Fischman D. Applying lean six sigma methodologies to improve efficiency, timeliness of care, and quality of care in an internal medicine residency clinic. Quality Management in Healthcare. 2010; 19(3):201–10.

43. Garuda SR, Javalgi RG, Talluri VS. Tackling no-show behavior: a market-driven approach. Health Marketing Quarterly. 1998; 15(4):25–44.

44. Guse CE, Richardson L, Carle M, Schmidt K. The effect of exit-interview patient education on no-show rates at a family practice residency clinic. The Journal of the American Board of Family Practice. 2003; 16(5):399–404. [PubMed: 14645330]

45. Can S, Macfarlane T, O'Brien K. The use of postal reminders to reduce non-attendance at an orthodontic clinic: a randomised controlled trial. British dental journal. 2003; 195(4):199–201. [PubMed: 12970701]

46. Hardy K, O'Brien S, Furlong N. Quality improvement report: Information given to patients before appointments and its effect on non-attendance rate. BMJ: British Medical Journal. 2001; 323(7324):1298. [PubMed: 11731398]

47. Kopach R, DeLaurentis P-C, Lawley M, Muthuraman K, Ozsen L, Rardin R, et al. Effects of clinical characteristics on successful open access scheduling. Health Care Management Science. 2007; 10(2):111–24. [PubMed: 17608053]

48. Glowacka KJ, Henry RM, May JH. A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. Journal of the Operational Research Society. 2009; 60(8):1056–68.

49. Alaeddini A, Yang K, Reddy C, Yu S. A probabilistic model for predicting the probability of no-show in hospital appointments. Health care management science. 2011; 14(2):146–57. [PubMed: 21286819]

50. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer. 2005; 27(2):83–5.

51. Caruana, R. Multitask learning Learning to learn. Springer; 1998. p. 95-133.

52. Ando RK, Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data. The Journal of Machine Learning Research. 2005; 6:1817–53.

53. Chapelle, O.Shivaswamy, P.Vadrevu, S.Weinberger, K.Zhang, Y., Tseng, B., editors. Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2010. Multi-task learning for boosting with application to web search ranking.

54. Bakker B, Heskes T. Task clustering and gating for bayesian multitask learning. The Journal of Machine Learning Research. 2003; 4:83–99.

55. Chari R, Lockwood WW, Coe BP, Chu A, Macey D, Thomson A, et al. SIGMA: a system for integrative genomic microarray analysis of cancer genomes. BMC genomics. 2006; 7(1):324. [PubMed: 17192189]
56. Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008; 70(1):53–71.
57. Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso. 2010 arXiv preprint arXiv:10010736.
58. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of statistical software. 2010; 33(1):1. [PubMed: 20808728]
59. Obozinski, GR.Wainwright, MJ., Jordan, MI., editors. Advances in Neural Information Processing Systems. 2008. High-dimensional support union recovery in multivariate regression.
60. Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences. 2009; 2(1):183–202.
61. Tseng P. Approximation accuracy, gradient methods, and error bound for structured convex optimization. Mathematical Programming. 2010; 125(2):263–95.
62. Chen X, Lin Q, Kim S, Xing EP. An efficient proximal-gradient method for single and multi-task regression with structured sparsity. stat. 2010; 1050:26.
63. Chen X, Lin Q, Kim S, Carbonell JG, Xing EP. Smoothing proximal gradient method for general structured sparse regression. The Annals of Applied Statistics. 2012; 6(2):719–52.
64. Nesterov, Y., editor. Soviet Mathematics Doklady. 1983. A method of solving a convex programming problem with convergence rate O (1/k2).
65. Nesterov, Y. Introductory lectures on convex optimization. Springer Science & Business Media; 2004.
66. Aarts E, Korst J. Simulated annealing and Boltzmann machines. 1988
67. Vaz AIF, Vicente LN. A particle swarm pattern search method for bound constrained global optimization. Journal of Global Optimization. 2007; 39(2):197–219.
68. Wright SJ, Nowak RD, Figueiredo MA. Sparse reconstruction by separable approximation. Signal Processing, IEEE Transactions on. 2009; 57(7):2479–93.
69. Barzilai J, Borwein JM. Two-point step size gradient methods. IMA Journal of Numerical Analysis. 1988; 8(1):141–8.
70. Hand, DJ., Daly, F., McConway, K., Lunn, D., Ostrowski, E. A handbook of small data sets. CRC Press; 1993.
71. Steyerberg EW, Eijkemans MJ, Harrell FE, Habbema JDF. Prognostic modeling with logistic regression analysis in search of a sensible strategy in small data sets. Medical Decision Making. 2001; 21(1):45–56. [PubMed: 11206946]
72. James, G., Witten, D., Hastie, T., Tibshirani, R. An introduction to statistical learning. Springer; 2013.
73. Breiman L. Random forests. Machine learning. 2001; 45(1):5–32.
74. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine learning. 1997; 29(2-3):131–63.
75. Jensen, FV. An introduction to Bayesian networks. UCL press; London: 1996.
76. Haykin, SS., Haykin, SS., Haykin, SS., Haykin, SS. Neural networks and learning machines. Pearson Education; Upper Saddle River: 2009.
77. Platt JC. 12 fast training of support vector machines using sequential minimal optimization. Advances in kernel methods. 1999:185–208.
78. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to Platt's SMO algorithm for SVM classifier design. Neural Computation. 2001; 13(3):637–49.

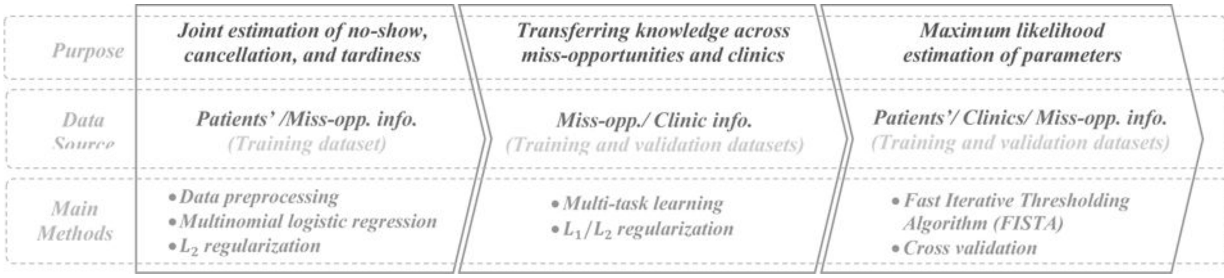



**Fig. 1.**

The general scheme of the proposed approach

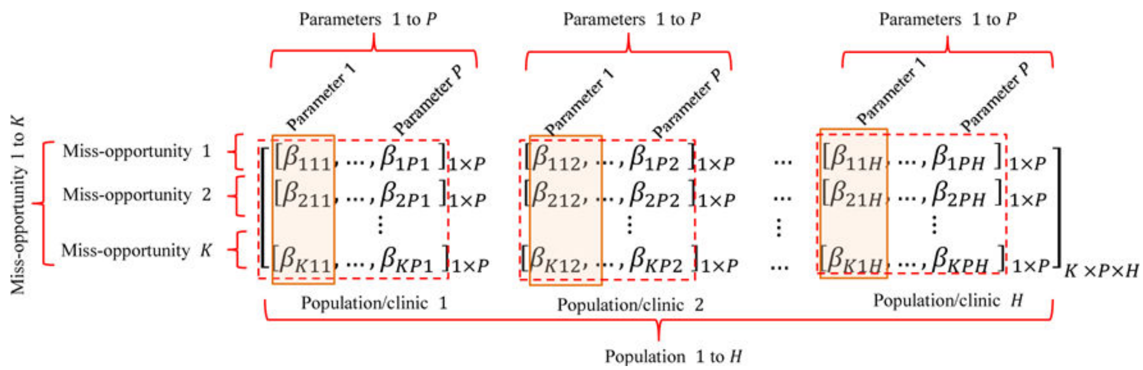



**Fig. 2.**

Illustration of the multinomial logistic regression coefficients $[\mathbf{B}]_{K \times P \times H}$ for modeling correlation among different miss-opportunities and different clinics

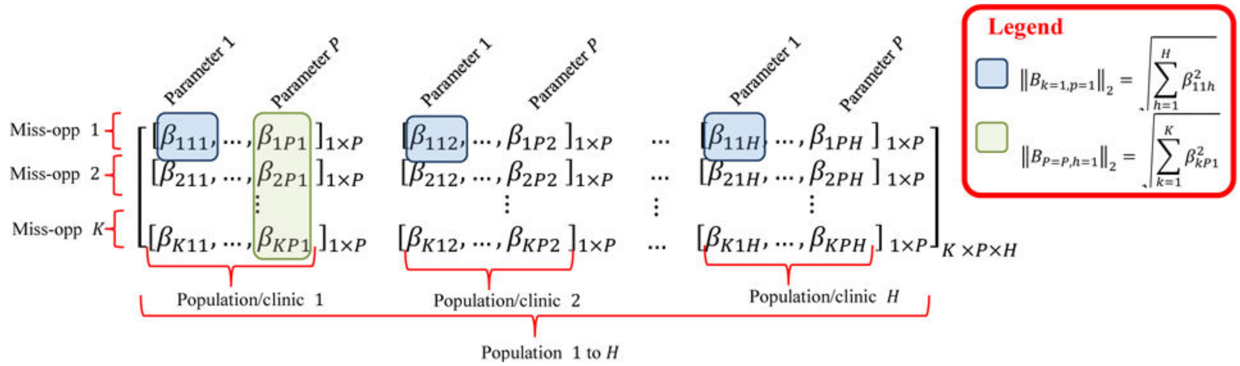



**Fig. 3.**

Illustration of multi-way regularization terms on specific parameters of multinomial logistic regression

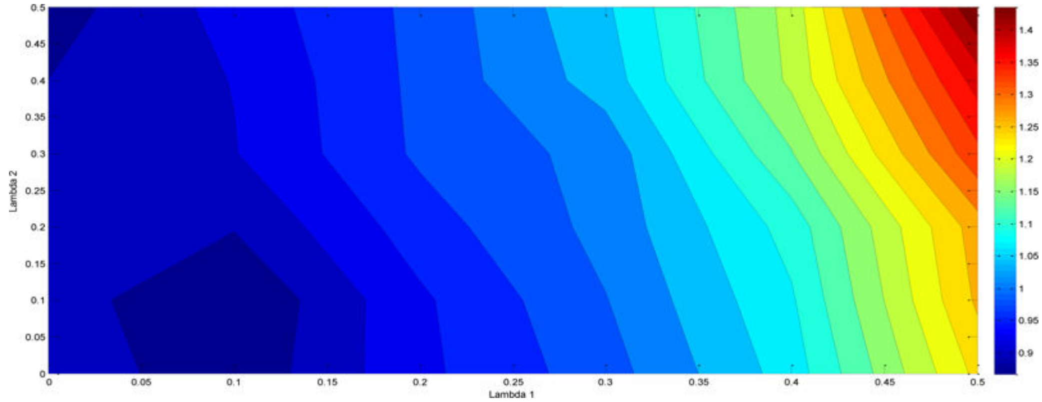



**Fig. 4.**

Contour plot of validation Mean Squared Error (MSE) for the proposed multi-way multi-task learning approach

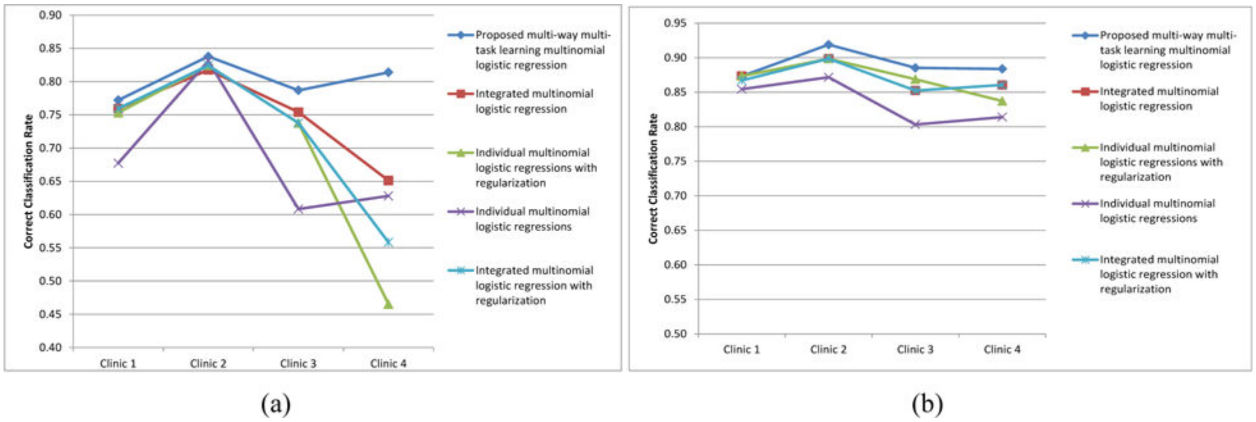
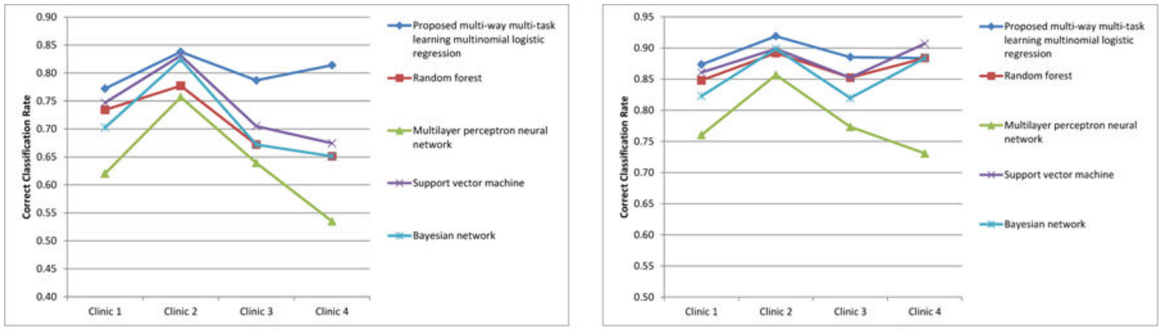



**Fig.5.**

Correct classification rate of comparing methods across different clinics for: (a) no-show prediction, (b) cancellation. The correct classification rates of each method across the four clinics are connected to provide better visual distinction.




(a) (b)

**Fig. 6.**

Correct classification rate of the comparing methods across different clinics for: (a) no-show prediction, (b) cancellation. The correct classification rates of each method across the four clinics are connected to provide better visual distinction.

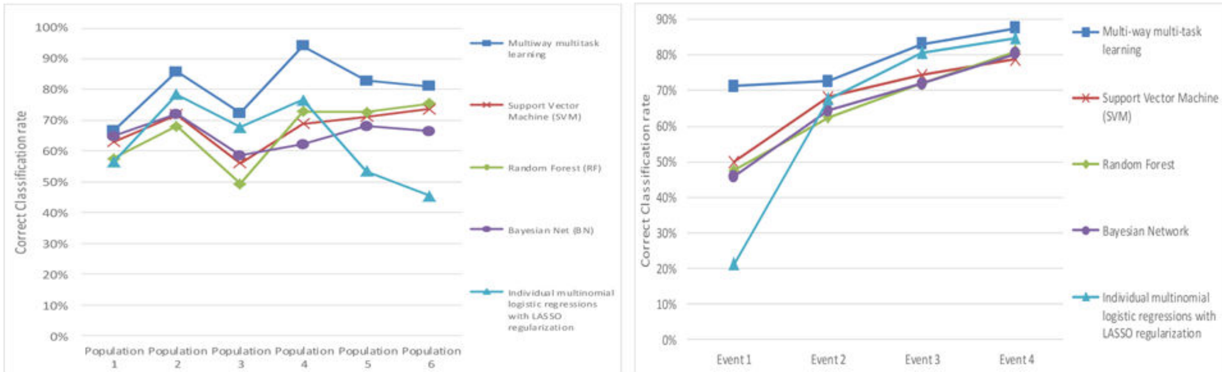



**Fig. 7.**

Correct classification rate of comparing methods in the simulation study across: (a) different populations, and (b) different events. The correct classification rates of each method are connected to provide better visual distinction.

## Table 1

Distribution of data across different clinics and risk factors

| | Clinic | Frequency | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Mental Health | | | Ophthalmology | | | Dental Care | | | Dermatology | | |
| | Miss-opportunity | No-show | Cancellation | Show-up | No-show | Cancellation | Show-up | No-show | Cancellation | Show-up | No-show | Cancellation | Show-up |
| Week day | Mon | 10 | 2 | 22 | 7 | 1 | 21 | 3 | 0 | 13 | 4 | 1 | 8 |
| | Tues | 5 | 3 | 15 | 7 | 3 | 25 | 4 | 2 | 6 | 8 | 3 | 9 |
| | Wed | 8 | 2 | 32 | 3 | 4 | 26 | 4 | 4 | 6 | 0 | 0 | 0 |
| | Thu | 7 | 9 | 20 | 6 | 4 | 23 | 1 | 1 | 6 | 1 | 1 | 0 |
| | Fri | 5 | 4 | 11 | 2 | 3 | 12 | 3 | 2 | 6 | 1 | 1 | 6 |
| | Sat | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Sun | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Daytime | Morning 8-11 | 15 | 8 | 34 | 15 | 11 | 68 | 7 | 2 | 19 | 2 | 1 | 6 |
| | late morning 11-12 | 4 | 2 | 14 | 3 | 0 | 9 | 1 | 2 | 5 | 3 | 1 | 1 |
| | Noon 12-13 | 2 | 2 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Early afternoon 13-15 | 10 | 6 | 48 | 5 | 2 | 26 | 4 | 2 | 5 | 5 | 3 | 7 |
| | Afternoon 15-16 | 4 | 2 | 4 | 2 | 2 | 5 | 3 | 1 | 6 | 4 | 1 | 9 |
| | Late afternoon 16-17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| | Evening >17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 |
| Distance to medical center (Average) | | 13.49 | 17.54 | 25.59 | 8.03 | 8.21 | 8.52 | 7.88 | 12.12 | 9.88 | 38.99 | 18.52 | 19.80 |
| Sex | Male | 28 | 13 | 76 | 24 | 15 | 104 | 13 | 8 | 34 | 12 | 5 | 13 |
| | Female | 7 | 7 | 27 | 1 | 0 | 4 | 2 | 1 | 3 | 2 | 1 | 10 |
| Age (Average) | | 50.82 | 50.70 | 50.37 | 66.8 | 64.46 | 67.52 | 52.66 | 47.00 | 58.24 | 57.42 | 54.50 | 57.69 |
| Marr. status | Divorced (m2) | 13 | 10 | 47 | 4 | 5 | 24 | 4 | 3 | 7 | 7 | 2 | 12 |
| | Never married (m3) | 6 | 4 | 23 | 9 | 5 | 27 | 3 | 0 | 7 | 5 | 2 | 4 |
| | Widowed (m4) | 7 | 3 | 19 | 3 | 2 | 2 | 5 | 1 | 7 | 1 | 2 | 3 |
| | Married (m1) | 9 | 3 | 14 | 9 | 3 | 55 | 3 | 5 | 16 | 1 | 0 | 4 |
| Poverty (Cov.) | (50-100%) s2 | 0 | 0 | 1 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| | (<50%) s3 | 8 | 9 | 39 | 9 | 3 | 55 | 5 | 3 | 5 | 10 | 2 | 10 |
| | (Non-service con.) s4 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 |
| | (NSC, pension) s5 | 4 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

| Clinic | Frequency | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mental Health | | | Ophthalmology | | | Dental Care | | | Dermatology | | |
| **Miss-opportunity** | **No-show** | **Cancellation** | **Show-up** | **No-show** | **Cancellation** | **Show-up** | **No-show** | **Cancellation** | **Show-up** | **No-show** | **Cancellation** | **Show-up** |
| **(Service con. <6) s6** | 5 | 1 | 19 | 8 | 4 | 29 | 6 | 0 | 11 | 3 | 1 | 5 |
| **(Service con. <5) s1** | 17 | 6 | 42 | 5 | 7 | 21 | 4 | 6 | 21 | 0 | 1 | 8 |

**Table 2**

Comparing methods information wherein individual model is indexed by $M$ and individual parameter is indexed by $P$

| No. | Approach | No. Initial Parameters | No. of final parameters (non-zero) | Shrinkage ratio | Individual model/par for each clinic | Individual model/par for each miss-opp. |
|---|---|---|---|---|---|---|
| 1 | Individual multinomial logistic regressions (for each clinic) | 192 (24 risk factors × 2 miss-opp × 4 clinics) | 192 | 0% | M | M |
| 2 | Individual multinomial logistic regressions with LASSO regularization | 192 (24 risk factors × 2 miss-opp × 4 clinics) | 174 | 9% | M | M |
| 3 | Integrated multinomial logistic regression (for all clinics) | 54 (27 risk factors × 2 miss-opp) | 54 | 0% | P | M |
| 4 | Integrated multinomial logistic regression with LASSO regularization | 54 (27 risk factors × 2 miss-opp) | 45 | 17% | P | M |
| 5 | Multi-way Multi-Task learning Multinomial logistic regression with $L_1/L_2$ regularization | 192 (24 risk factors × 2 miss-opp × 4 clinics) | 127 | 34% | M | M |

**Table 3**

Estimated parameters of the proposed multi-way multi-task learning approach for multinomial logistic regression, wherein zero parameters associated with deselected risk factors are shaded

| | Clinic | Mental Health | | Ophthalmology | | Dental Care | | Dermatology | |
|---|---|---|---|---|---|---|---|---|---|
| | Miss-opportunity | No-show | Cancellation | No-show | Cancellation | No-show | Cancellation | No-show | Cancellation |
| | Intercept | 0.0 | 0.0 | 0.0 | −1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Weekday | Mon | 0.8 | −0.4 | −1.0 | −3.0 | −0.5 | −2.5 | −1.0 | −1.5 |
| | Tues | 0.3 | −0.3 | −1.0 | −1.4 | 0.6 | 0.5 | 0.0 | −0.3 |
| | Wed | −0.1 | −1.0 | −1.8 | −0.8 | 0.8 | 1.0 | 0.0 | 0.0 |
| | Thu | 0.3 | 1.0 | −0.9 | −0.9 | −0.8 | 0.4 | 1.0 | 1.8 |
| | Fri | 0.5 | 1.3 | −1.5 | −1.0 | 0.0 | 0.3 | −1.8 | −0.3 |
| | Sat | 0.0 | 0.0 | −0.6 | −1.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| Daytime | Morning 8-11 | 0.0 | −0.5 | 0.0 | 0.0 | 0.3 | −1.0 | −0.5 | −0.3 |
| | late morning 11-12 | −0.5 | −0.8 | 0.5 | −1.6 | −0.3 | −0.3 | 0.3 | −0.3 |
| | Noon 12-13 | 0.3 | 0.8 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Early afternoon 13-15 | −1.5 | −2.0 | −0.4 | −1.0 | 1.0 | 0.3 | 0.8 | 0.8 |
| | Afternoon 15-16 | 0.3 | 0.1 | 0.5 | 1.0 | 0.3 | −1.0 | −0.5 | −1.5 |
| | Late afternoon 16-17 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| | Distance to medical center (mile) | -4.5 | −0.8 | −1.8 | −0.5 | −2.5 | 1.3 | 1.3 | −0.3 |
| | Sex | 0.3 | −0.3 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.3 |
| | Age | 0.0 | −1.8 | 0.0 | 0.0 | −1.3 | −2.5 | 0.0 | −0.5 |
| Marriage status | Divorced (m2) | 0.0 | 0.0 | 0.0 | 1.3 | 0.0 | −0.5 | 0.3 | 0.0 |
| | Never married (m3) | −0.3 | −0.3 | 0.0 | 1.3 | 0.3 | −2.8 | 0.8 | 1.0 |
| | Widowed (m4) | −1.3 | −0.8 | 2.1 | 1.8 | 0.4 | −1.3 | −0.8 | 0.5 |
| Poverty (Coverage) | (50-100%) s2 | 0.0 | 0.0 | −0.5 | 0.8 | −1.3 | 0.6 | −4.3 | −2.0 |
| | (<50%) s3 | −1.3 | −0.5 | 1.5 | −1.5 | 0.0 | 0.0 | 0.0 | 0.0 |
| | (Non-service connected) s4 | −1.0 | 0.5 | −0.5 | −1.5 | 0.0 | 1.8 | 0.0 | −1.3 |
| | (NSC, pension) s5 | 2.8 | 5.8 | 0.0 | 0.0 | 0.0 | 0.0 | 2.8 | 3.5 |
| | (service connected <6) s6 | 2.5 | 3.4 | −0.8 | 2.1 | 0.0 | 0.0 | 0.0 | 0.0 |

**Table 4**

Comparing methods information

| Row | Method | Parameters (estimated) | Reference | Note |
|---|---|---|---|---|
| 1 | Proposed multi-way multi-task learning framework | Regularization: $L_1/L_2$<br>Tuning par.: $(\lambda_1^*, \lambda_2^*) = (0.1, 0.1)$ | N/A | • Data splitting strategy: Five-fold cross validation |
| 2 | Random Forest (RF) | Confidence factor: 0.25 | [70] | • Algorithm: J48<br>• Data splitting strategy: Five-fold cross validation |
| 3 | Bayesian Net (BN) | Estimator: Simple ( $\alpha$=.5)<br>Search algorithm : K2 | [71,72] | • Data splitting strategy: Five-fold cross validation<br>• Search algorithm: hill climbing |
| 4 | Multilayer Perceptron Neural Net (MLP) | Hidden layers: a<br>Learning rate: .3<br>Momentum: 0.2 | [73] | • Data splitting strategy: Five-fold cross validation |
| 5 | Support Vector Machine (SVM) | Kernel fun.: Polynomial<br>Filter type: Normalize training data<br>Tolerance par: 0.001<br>Optimization algorithm: John Platt's Sequential Minimal Optimization | [74,75] | • Data splitting strategy: Five-fold cross validation |