Health Care Manag Sci (2011) 14:146–157
DOI 10.1007/s10729-011-9148-9

A probabilistic model for predicting the probability of no-show in hospital appointments

Adel Alaeddini · Kai Yang · Chandan Reddy · Susan Yu

Received: 29 July 2010 / Accepted: 18 January 2011 / Published online: 1 February 2011

Abstract The number of no-shows has a significant impact on the revenue, cost and resource utilization for almost all healthcare systems. In this study we develop a hybrid probabilistic model based on logistic regression and empirical Bayesian inference to predict the probability of no-shows in real time using both general patient social and demographic information and individual clinical appointments attendance records. The model also considers the effect of appointment date and clinic type. The effectiveness of the proposed approach is validated based on a patient dataset from a VA medical center. Such an accurate prediction model can be used to enable a precise selective overbooking strategy to reduce the negative effect of no-shows and to fill appointment slots while maintaining short wait times.

Keywords Logistic regression · Beta distribution · Bayesian inference · Healthcare · Scheduling

A. Alaeddini (✉) · K. Yang
Department of Industrial & Systems Engineering,
Wayne State University,
Detroit, MI 48202, USA
e-mail: adel.alaeddini@gmail.com

K. Yang
e-mail: ac4505@wayne.edu

C. Reddy
Department of Computer Science, Wayne State University,
Detroit, MI 48202, USA
e-mail: reddy@cs.wayne.edu

S. Yu
John D Dingell VA Medical Center,
Detroit, MI 48201, USA
e-mail: Susan.Yu@va.gov

1 Introduction

Scheduled but unattended appointment slots (no shows) cause significant disturbance on the smooth operation of almost all scheduling systems [3, 28]. In this paper, we consider the problem of effective scheduling by predicting no-shows accurately from the past data available. Specifically, we apply our model to healthcare data collected from medical centers. Medical healthcare centers can incur losses of hundreds of thousands of dollars yearly because of these no-shows.

No-show rates at medical centers can vary from as little as 3% to as much as 80% depending on the type of center and demographic information of the patients of the medical center [28, 30, 31]. According to Barron [1], eight studies at inner-city clinics, community health centers, and university medical centers indicate no-show rates of 10–30% while the estimated no-show rates for private practice are 2–15%. A recent study conducted at the national level [22] indicated that over 21% of all appointments made in health care systems may result in a no show.

While the reasons for these no-shows might vary from previous experience to personal behaviors, several practitioners and researchers have often neglected this important aspect of the scheduling problem. High rates of no-shows not only cause inconvenience to the hospital management but also have a significant impact on the revenue, cost and resource utilization for almost all healthcare systems. Hence, accurate prediction of no-show probability is a cornerstone for any scheduling systems and non-attendance reduction strategy [8, 10, 22, 23, 28].

In this paper, we develop a hybrid probabilistic model based on logistic regression and empirical Bayesian inference to predict the probability of no-shows in real-time. Our model uses both the general social and

demographic information of the individuals and their clinical appointment attendance records, as well as other variables such as the effect of appointment date and clinic type. Some of the critical factors that affect no-show rates will be investigated and modeled into the scheduling process: these factors include age, gender, race, population sector, and factors related to the previous appointment experience of the person such as number of previous appointments, their types and lead times. We will also consider the effect of personal behaviors such as previous appointment-keeping patterns in predicting no-shows, and build an empirical Bayesian paradigm for modeling this behavior.

Our robust and accurate scheduling system, including a hybrid prediction model, can be used to enable a precise selective overbooking strategy to reduce the negative effect of no-shows and to fill appointment slots while maintaining short wait times. The result of the proposed method can be used to develop more effective appointment scheduling [9, 16, 18, 19, 27]. It can be used for developing effective strategies such as selective overbooking for reducing the negative effect of no-shows and filling appointment slots while maintaining short wait times [25, 29, 34].

The rest of the paper is organized as follows: Section 2 summarizes related work in the literature. Section 3 describes relevant background for our algorithm. Section 4 discusses our proposed model for predicting no-show probabilities and also explains the optimization procedure used to improve the model predictions. Section 5 presents the results from applying the proposed model on a medical healthcare center appointment system. Finally, Section 6 concludes our work and presents some future extensions of the proposed model.

2 Relevant background

There are wide varieties of techniques that can be used for no-show probability estimation. Here, we will briefly discuss some of the related factors and quantitative methods studied in this domain.

2.1 Factors affecting no-shows

Several studies have discussed the effect of patients' personal information, such as age, gender, nationality, and population sector, on no-show probability [2, 16]. Some researchers have also investigated the relationship between no-show probability and factors related to the previous appointment experience of the person, such as the number of previous appointments, appointment lead times, wait times, appointment type, and service quality [2, 11, 13, 15, 17, 26]. A few studies also considered the effect of personal issues such as oversleeping or forgetting, health status, presence of a sick child or relative, and lack of transportation on missing appointments [6, 7]. We will consider many of these factors in our proposed model and also study the effect of personal behavior such as previous appointment-keeping pattern [12] in predicting no-shows.

2.2 Population-based models

Population-based techniques use a variety of methods drawn from statistics and machine learning, which can be used for predicting no-shows [12]. These methods use information from a whole population (dataset), in the form of set factors, to estimate the probability of a patient showing up for a scheduled appointment. Logistic regression is one of the most popular statistical methods in this category that is used for binomial regression, which can predict the probability of no-show by fitting numerical or categorical predictor variables in data to a logit function [21]. There has been some work using Tree- and rule-based models, which create if–then constructs to separate the data into increasingly homogeneous subsets, based on which the desired predictions of no-show can be found [16].

The problem with these population-based methods is that, although they provide a good initial estimate, because they do not differentiate between the behaviors of individual persons, they cannot update effectively, especially when a small dataset is used. Another problem with these methods is that once the model has been built, adding new data has very small effect on the result, especially if the size of initial dataset is much larger than the size of the new data. In Section 5, we will compare the performance of the above methods with the proposed method.

2.3 Individual-based models

Individual-based approaches are primarily time series and smoothing methods that are used for no-show rate prediction. These methods utilize past behaviors of individuals for estimating future no-show probability. Time series methods forecast future events such as no-shows based on past events by using stochastic models. There are different types of time series models, but the most common three classes are autoregressive (AR) models, integrated (I) models, and moving average (MA) models. These three classes depend linearly on previous data [5]. Combinations of these ideas produce autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models. The autoregressive fractionally integrated moving average (ARFIMA) model generalizes the former three. There are also model-free analyses such as wavelet transforms-based methods, which are not considered in this study [5].

Smoothing is an approximating function that attempts to capture important patterns in the data, while leaving out noise or other fine-scale structures/rapid phenomena. Many different algorithms are used in smoothing. Two of the most common algorithms are the moving average and local regression [32].

Bayesian inference is a method of statistical inference in which some kind of evidence or observations are used to update a previously-calculated probability such as improving the initial estimate of no-show probability [4]. To use Bayes' theorem, we need a prior distribution $g(p)$ that provides our initial prediction about the possible values of the parameter p before incorporating the data. The posterior distribution is proportional to prior distribution times likelihood $f(y|p)$:

$$g(p|y) \propto g(p) \times f(y|p) \tag{1}$$

If the prior is continuous the posterior distribution can be calculated as:

$$g(p|y) = \frac{g(p) \times f(y|p)}{\int_0^1 g(p) \times f(y|p) dp} \tag{2}$$

While individual-based methods are very fast and effective in modeling the behavioral (no-show) pattern of each individual, and work well with a small dataset, they do not use the information of the rest of the population, so they do not provide a reliable initial estimate of no-show probability, which is especially important in small datasets. In Section 5, we will compare the performance of the above methods with the proposed method.

As described above, each of the population-based and individual-based approaches has some advantages and disadvantages. However, no studies have employed these methods together to overcome their individual problems and improve their performance. In the next section, we develop a hybrid approach that combines logistic regression as a population-based approach along with Bayesian inference as an individual-based approach for our no-show prediction model. We will also compare the performance of the proposed approach with the representative algorithms from each of the population-based and individual-based approaches.

3 Preliminaries

In this section, we introduce important background related to the proposed algorithm. We describe the notations used in this paper and explain some basics of logistic regression and beta distribution, which are the core components of our algorithm. We also give more details about the Bayesian updating of the beta distribution, which is a vital component of modeling an individual's behavior.

3.1 Notation used

D_{GI}	Database of each individual's personal information
D_{NR}	Database of appointment information and attendance records of each person
$F(X_i,B)$	Logistic regression model
B	Vector of logistic regression parameters ($B=[\beta_0,\beta_1,...,\beta_k]$)
X_{ij}	Factors affecting no-show probability of patient i (independent variables in the logistic regression model), $\left(X_{ij} \in D_{GI} \vee D_{NR}, X_{ij} = [x_{ij0}, x_{ij1}, x_{ijk}]\right)$
Y_{ij}	Person i show/no-show record for appointment j, Y_{ij} = (0,1)
$p_{i0}(I_i = 1\|X_i)$	Prior probability of no-show
$\left(\alpha_{ij}^{pos}, \beta_{ij}^{pos}\right)$	Beta distribution posterior parameters
$\widehat{p}^{Model}$	Posterior probability of no-show
$\widehat{p}^{Emp}$	Empirical probability of no-show
n_i	Total number of appointments for person i
W_j	Weight for appointment j
T	Threshold for convergence of the objective
D	Improvement in the objective function
LG	Logistic regression model

3.2 Logistic regression

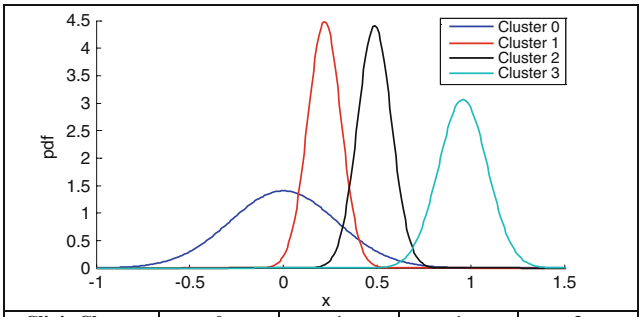
Logistic regression is a generalized linear model used for binomial regression, which predicts the probability of occurrence of an event by fitting numerical or categorical predictor variables in data to a logit function [24]:

$$Logit(p) = \log(p/1-p) \tag{3}$$

where $0 \leq p \leq 1$ and $(p/1-p)$ is the corresponding odds. The logistic function can be written as:

$$p = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+...+\beta_k x_k)}} \tag{4}$$

where p represents the probability of a particular outcome, given that set of explanatory variables and unknown regression coefficients $\beta_j, (0 < j < k)$ can be estimated



Clinic Cluster	0	1	1	3
μ	0	0.2197	0.4874	0.9595
σ	0.2843	0.0890	0.0904	0.1300

Fig. 1 The result of clustering the clinics based on their no-show probabilities

using maximum likelihood (MLE) methods common to all generalized linear models [21].

3.3 Beta distribution and Bayesian update

Beta distribution: *Beta*(α, β) represents a family of common continuous distributions defined on the interval [0,1] parameterized by two positive shape parameters, typically denoted by α and β with probability density function:

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad (5)$$

where Γ is the gamma function, and $\Gamma(\alpha+\beta)/\Gamma(\alpha)\Gamma(\beta)$ is a normalization constant to ensure that the total probability integrates to unity [14]. The beta distribution is the conjugate prior of the binomial distribution. From the Bayesian statistics viewpoint, a Beta distribution can be seen as the posterior distribution of the parameter p of a binomial distribution after observing α–1 independent events with probability p and β–1 with probability 1–p, if there is no other information regarding the distribution of p [4].

Table 2 Attendance record of a sample patient

Appointment No.	Appointment date	Clinic cluster	No-show
1	10/13/2009	1	0
2	10/29/2009	1	1
3	11/10/2009	0	0
4	11/17/2009	1	1
5	12/2/2009	2	1
6	12/8/2009	1	0
7	12/9/2009	2	0
8	12/23/2009	1	0
9	12/23/2009	1	1
10	12/29/2009	1	0
11	12/31/2009	0	1

3.4 Bayesian update of Binomial distribution

In Bayesian statistics, a Beta distribution [4] is a common choice for updating a prior estimate of the Binomial distribution parameter p because:

1. A Beta distribution is the conjugate prior of a Binomial distribution (See Section 3.3).
2. Unlike a Binomial distribution, a Beta distribution is a continuous distribution, which is much easier to work with in terms of inference and updating.
3. A Beta distribution has two parameters, which allows it to take different shapes, making it suitable for representing different types of priors.

If *Beta*(α,β) is used as a prior, based on the conjugacy property of Beta distribution, the posterior would be a new Beta posterior with parameters $\alpha' = \alpha + y$ and $\beta' = \beta + n - y$. In other words, Beta distribution can be updated

Table 1 Data structure and optimal value of the weighting factors

Analysis	Validation set	Appointment recency					Preceding non-workday		Clinic cluster			
		<1 week.	<1 month	<3 months	6<months	>6 months	Not-before holiday	Before holiday	Very important	…	…	Not important
Appointment time base	1	1	1	1	0.95	0.9	1	0.92	1	1	0.9	0.75
	2	1	0.96	0.91	0.88	0.81	1	1	1	0.91	0.84	0.62
	3	1	1	1	0.70	0.84	1	1	1	0.89	0.88	0.71
Patient base	1	1	1	0.72	0.72	0.62	1	1	1	0.95	0.79	0.51
	2	1	1	0.62	0.68	0.55	1	0.95	1	1	0.81	0.63
	3	1	0.97	0.71	0.66	0.51	1	0.98	1	1	0.89	0.58

simply by adding the number of successes y to α and the number of failures $n - y$ to β:

$$g(p|y) \sim Beta(\alpha + y, \chi + n - y)$$
$$g(p|y) = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y-\alpha-1}(1-p)^{n-y+\beta-1} \quad (6)$$

As discussed earlier, individual-based approaches like empirical Bayesian inference will not be able to provide an initial estimate of the prior distribution. Hence, before applying the Bayesian update, the parameters of the prior distribution should be initialized.

Bolstad [4] suggests choosing parameters that match the belief about the location (mean) and scale (standard deviation) of the original distribution. Hence, if an initial guess of parameter p is available, which in our study can be obtained from population-based approaches such as logistic regression, Beta distribution prior parameters can be computed by solving the following system of equations for α and β.

$$\begin{cases} p_i = \frac{\alpha}{\alpha+\beta} \\ \frac{p_i(1-p_i)}{n} = \sqrt{\frac{p_i(1-p_i)}{\alpha+\beta+1}} \end{cases} \quad (7)$$

The point estimate of the posterior parameter p of the binomial distribution would be the mean of Beta distribution $\frac{\alpha}{\alpha+\beta}$ of the updated Beta distribution.

Algorithm 1: No-show Prediction Algorithm

Input: Input data $\left(X_{ij}, Y_{ij}\right)$, Threshold parameter T

Output: Estimated no-show probability $\hat{p}^{Model}$, Beta distribution posterior parameters $\left(\alpha_{ij}^{pos}, \beta_{ij}^{pos}\right)$, Logistic regression estimated parameters $\hat{B}$

Procedure:

*/*Logistic regression*/*

$\hat{B} \leftarrow$ *Calculate MLE of(3.2) parameter*

$\hat{p}_{0ij}\left(Y_i = 1 \mid X_{ij}\right) \leftarrow F\left(X_{ij}, \hat{B}\right)$

$\left(\alpha_i^{pri}, \beta_i^{pri}\right) \leftarrow$ *Solve system of equation (7) with* $\hat{p}_{0i}\left(Y_i = 1 \mid X_i\right)$

*/*Weight optimization*/*

$$\hat{p}_i^{Emp} \leftarrow \frac{\sum_{j=1}^{m} Y_i^j}{m - l + 1}$$

Until equation (8) improvement $D < T$ *do*

W_j *set a value for appointments weights*

*/*Bayesian update*/*

$$\begin{cases} a_{ij}^{pos} \leftarrow a_i^{pri} + \sum_{i,1}^{i,j-1}\left(\prod_{\omega \in W} w_{ij\omega}\right) Y_{ij} \\ \beta_{ij}^{pos} \leftarrow \beta_i^{pri} + n_i - \sum_{i,1}^{i,j-1}\left(\prod_{k \in W} w_{ijk}\right) Y_{ij} \end{cases}$$

$$\hat{p}^{Model} \leftarrow \frac{\alpha_{ij}^{pos}}{\alpha_{ij}^{pos} + \beta_{ij}^{pos}}$$

$$\overline{p} \leftarrow \sum_{i=1}^{n}\left(\hat{p}_i^{Model} - \hat{p}_i^{Emp}\right)/n$$

$$S_p \leftarrow \frac{\sum_{i=1}^{n}\left(\hat{p}_i^{Model} - \hat{p}_i^{Emp}\right)^2 - \left[\left(\sum_{i=1}^{n}\left(\hat{p}_i^{Model} - \hat{p}_i^{Emp}\right)\right)^2 / n\right]}{n-1}$$

$$t_0 \leftarrow \frac{\overline{p}}{S_p / \sqrt{n}}$$

Return $\hat{p}^{Model}$

Table 3 Fitted logistic regression model for the sample patient

Sex	DOB	Marriage status	Medical service coverage	Zip code	Clinic cluster	Recency	preceding non-workday	Constant
71.691	−0.8600	6.51E-05	−0.13596	0.018	0.0015	0.482	0	3.0410

4 The proposed algorithm

Algorithm 1 illustrates the proposed approach, which can be categorized in three stages:

1. Initial no-show probability estimation
2. Bayesian update of the no-show estimate
3. Weight optimization

In the first stage, based on the dataset of individuals' personal information (D_{GI}) (such as gender, marital status, etc.) and their sequence of appointment information (e.g. previous attendance records (D_{NR})), a logistic regression model $F\left(X_{ij}, \widehat{B}\right)$ is formulated (*line 2*). Then, using logistic regression, an initial estimate of no-show probability is calculated, given by $\widehat{p}_{0i}(Y_i = 1|X_i)$. As discussed in Section 2, Logistic regression bundles the information of the complete population together and finds a reliable initial estimate of no-show ($\widehat{p}_{0i}$).

In the second stage, which is interlaced with the third stage, the initial estimate is used in a Bayesian update procedure to find the posterior no-show probability for each person. For this purpose, $\widehat{p}_{0i}$ is transformed into prior parameters of a Beta distribution $\left(\alpha_i^{pri}, \beta_i^{pri}\right)$ as shown in line 4. Next, using the attendance record of each person (Y_{ij}) the posterior parameters $(\alpha_i^{pos}, \beta_i^{pos})$ and posterior probability of no show $\widehat{p}^{Model}$ is calculated (lines 10 and 11). Again, as discussed in Section 2, the reason the Bayesian update procedure is applied to the output of the logistic regression is that showing up behavior of individual patients is typically not captured well by logistic regression. Additionally, updating regression parameters based on new data records is both more difficult and also only marginally effective (especially when the model is already constructed on a huge dataset) in comparison to Bayesian update. The reason is due to the use of empirical conjugate Bayesian models, which are easier to compute than classical regression because one would only need to keep track of the parameters dynamically based on new information rather than running the whole regression again.

In the third stage, appointments are weighted based on a subset of factors $W = [w_1, ..., w_\omega]$ *(line 8)* to increase the model performance in estimating the Empirical probability of no-show. An optimization procedure is used for finding the optimal value of the weights. The objective function of the model is to minimize the difference between the empirical and estimated probability of no-show:

$$\begin{aligned} &Max\ p - value_{paired\ t-test} = p\left(t_0 < -t_{\frac{\alpha}{2},n-1}\right) + p\left(t_0 > t_{\frac{\alpha}{2},n-1}\right) \\ &S.T: \\ &w_1, ..., w_\omega \in (0,1) \end{aligned} \tag{8}$$

Table 4 Bayesian update of beta distribution parameters

Appointment No.	Appointment date	clinic category	Weight			No-show	Weighted no-show	α	β	Estimated p
			preceding non-workday	Recency	Clinic cluster					
								0.3453	0.6547	0.345
12	1/25/2010	0	1	0.9	1	1	0.35	0.695	0.655	0.515
13	1/26/2010	1	1	0.9	0.9	0	0	0.695	1.655	0.296
14	2/2/2010	0	1	0.9	1	0	0	0.695	2.655	0.208
15	2/4/2010	2	1	0.9	0.75	0	0	0.695	3.655	0.160
16	2/6/2010	2	1	0.9	0.75	0	0	0.695	4.655	0.130
17	2/17/2010	0	1	0.9	1	0	0	0.695	5.655	0.109
18	2/18/2010	1	1	0.9	0.9	0	0	0.695	6.655	0.095
19	2/23/2010	0	1	0.9	1	0	0	0.695	7.655	0.083
20	3/2/2010	1	1	0.9	0.9	0	0	0.695	8.655	0.074
21	3/9/2010	0	1	0.9	1	1	0.35	1.045	8.655	0.108
22	3/16/2010	0	1	0.9	1	0	0	1.045	9.655	0.098
23	3/18/2010	2	1	0.9	0.75	0	0	1.045	10.655	0.089

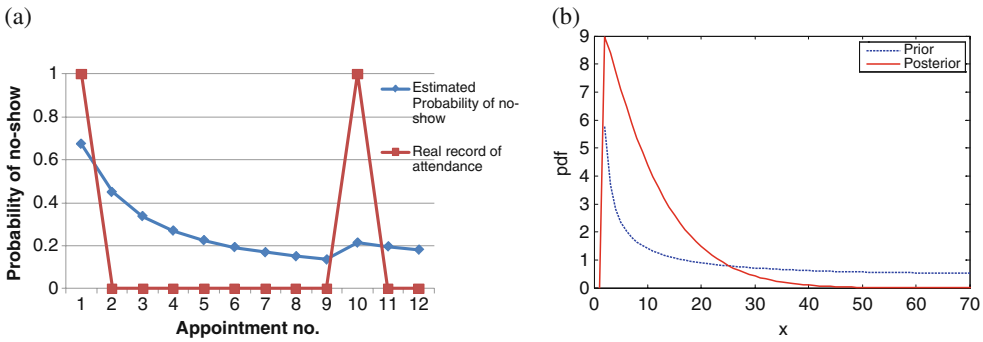


Fig. 2 Applying the proposed model for a sample patient: **a** real record of attendance and estimated probability of no-show using the proposed model, **b** prior and posterior distribution of no show

Where $w_1, ..., w_\omega$ are the weights to be optimized and p-value$_{paired\ t\text{-}test}$ is the p-value of a two-sided statistical hypothesis testing of the paired estimated p using the model and estimated p using the attendance records:

$$\begin{cases} H_0 : p_D^{Model} = p_D^{Emp} \\ H_1 : p_D^{Model} \neq p_D^{Emp} \end{cases} \tag{9}$$

It should be noted that mean squared error (MSE) can also be used as the objective function. However, the t-statistic used above not only contains MSE (S_p in the denominator of the t statistic is a linear function of MSE) (*line 14*), but also has a statistical distribution that makes it a better choice for our optimization model.

In (4.1), $t_{\frac{\alpha}{2},n-1}$ is the percentage of points or value of t random variables with $n-1$ degrees of freedom such that the probability that t_{n-1} exceeds this value is α, and $t_0 = \frac{\overline{p}}{S_p/\sqrt{n}}$ where $\overline{p} = \sum_{i=1}^{n} \left(\widehat{p}_i^{Model} - \widehat{p}_i^{Emp}\right)/n$ and S_p is calculated as follows:

$$S_p \leftarrow \frac{\sum_{i=1}^{n} \left(\widehat{p}_i^{Model} - \widehat{p}_i^{Emp}\right)^2 - \left[\left(\sum_{i=1}^{n} \left(\widehat{p}_i^{Model} - \widehat{p}_i^{Emp}\right)\right)^2 / n\right]}{n-1} \tag{10}$$

Where $\widehat{p}_i^{Emp}$ is the real rate of no-show for person i calculated as $\widehat{p}_i^{Emp} = \frac{\sum_{j=1}^{m} Y_i^j}{m-l+1}$, with Y_i^j as a binary (random) variable representing records of no-show/show of patient i for appointment j. Here, l is the index of first appointment in the validation dataset, which is discussed shortly, and m is the total number of appointments in the validation dataset for patient i. Also $\widehat{p}_i^{Model}$ is the estimated no-show probability calculated based on weighted appointments using the proposed model.

The optimization procedure is as follows: at every iteration, a vector of weights is assigned to the appointments in the validation dataset (line 8). The weighted appointments are then plugged into the Bayesian update mechanism for estimating the probability of no-show (lines 10 and 11). Next, the estimates of the proposed model and real attendance records are compared by forming a t-statistic (lines 12 to 14) and the p-value of the paired t-test, which shows the goodness of the assigned weights, is used for improving the initial set of weights (line 7) . This procedure continues until no improvement is observed. Then, the $\widehat{p}^{\,\mathrm{Model}}$ of the iteration resulting in the best value of the objective function is used as the no show estimate.

5 Experimental results

We applied our proposed model to healthcare data collected at the Veteran Affairs (VA) Medical Center in Detroit. We studied the performance of the proposed method along with a number of population- and individual-based algorithms

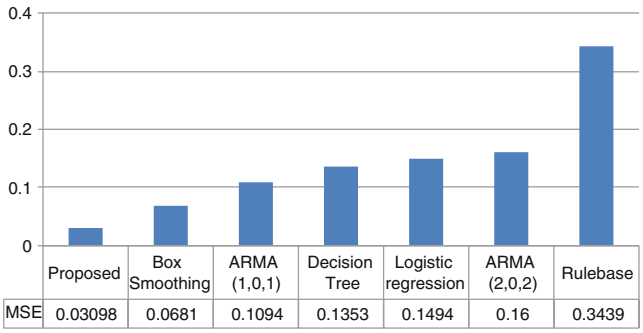


Fig. 3 Mean squared error (MSE) of different methods used for comparison

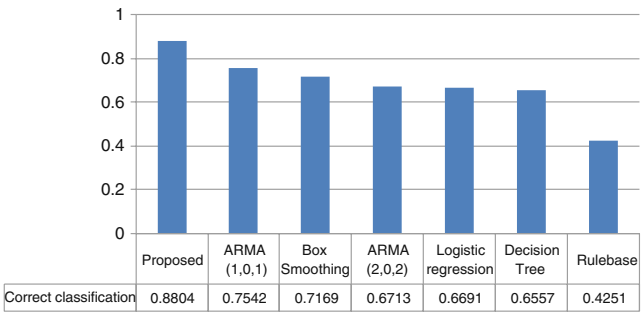


Fig. 4 Percentage of correct classification of different methods used for comparison

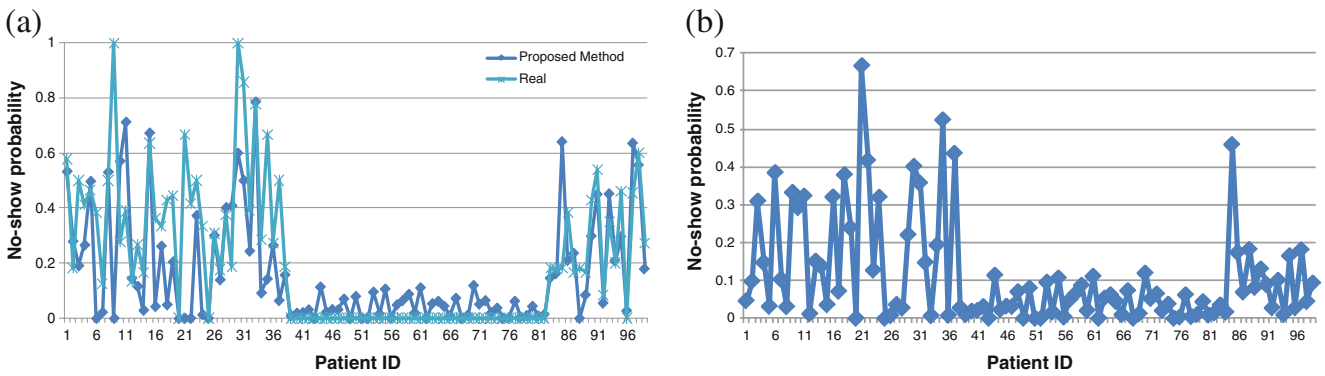


Fig. 5 Proposed approach performance over patients: **a** estimated vs empirical probability of no-show, **b** absolute difference of estimated and real no-show probability

on a database of 1,543 patient records with the following appointment information: (1) sex, (2) date of birth (DOB), (3) marital status, (4) medical service coverage, (4) address (zip code), (5) clinic and (6) prior attendance record in the hospital. We performed a threefold cross validation with approximately 500 records each for training, validation and testing.

This section is organized as follows: first we discuss pre-processing of the data. Next, we provide a stylized example of one patient record to illustrate how the model works. Finally, we discuss the results of applying the model to the dataset using two types of analysis: (1) patient-based analysis and (2) appointment time-based analysis.

5.1 Data pre-processing

The attributes in the dataset must be pre-processed before being used in the model. Specifically, pre-processing included dealing with missing attributes and eliminating co-linearity. In addition, due to the variety of clinics (more than 150 in our case), the accuracy of the logistic regression would be severely affected if this explanatory variable gets directly used in the model.

This problem is addressed by clustering similar clinics with respect to their no-show rates. While various types of clustering algorithms can be used for this purpose because the clinics are originally different in type, grouping them into a set of clusters will result in clusters with different density and dispersion. Such characteristics can be effectively modeled using Generalized Mixture Models (GMM). (See Appendix)

Figure 1 shows the result of clustering the clinics based on their probability of no-show using GMM. The final result has been verified by a team of experts.

Also, (1) appointment recency, (2) appointment preceding non-work days (Saturday, Sunday, and holidays), and (3) clinic cluster, are considered as weighting factors ($W = [w_1, ..., w_\omega]$. Regarding the first factor, it is reasonable that no-show records that occurred a long time ago do not carry the same weight as recent no-shows. This is based on the fact that patients may gradually or abruptly change their behavior, which should be reflected in the model. Regarding the second and third weighting factors, a preliminary study of the data revealed strong correlations between no-show rates and days close to holidays, and between no-show rates and clinic clusters. Hence, these factors are modeled into our approach.

The weights discussed above are arranged in a special data structure before being applied to the data. For the

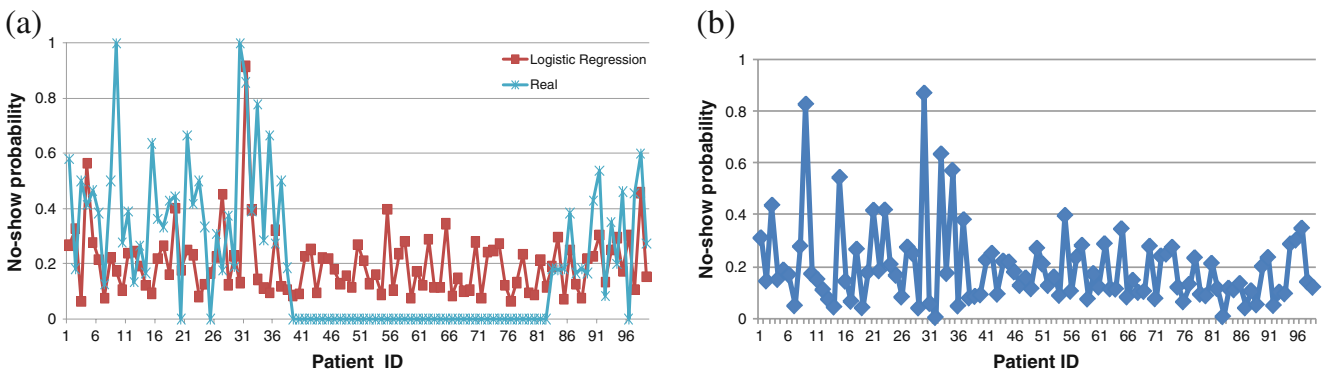


Fig. 6 Logistic regression performance over patients: **a** estimated versus empirical probability of no-show, **b** absolute difference of estimated and real no-show probability

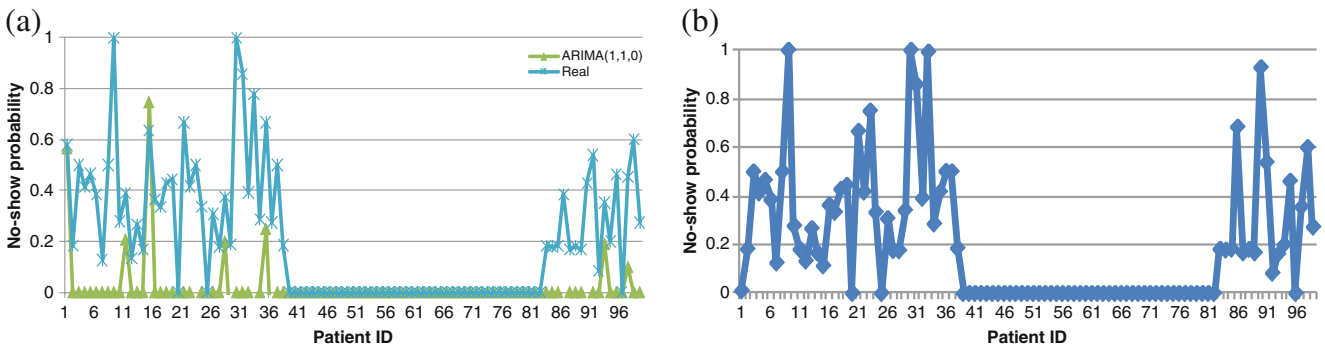


Fig. 7 ARMA(1,0,1) performance over patients: **a** estimated versus empirical probability of no-show, **b** absolute difference of estimated and real no-show probability

appointment recency, where more importance should be assigned to more recent appointments, a logarithmic time framework with five weights is used. For the appointment preceding non-work days, two weights are applied: one for Monday to Thursday and one for Friday and days before holidays. Finally, for clinic cluster, based on the groups derived using GMM, four weights are defined. Table 1 shows the final data structure and optimal values of the weights for the analyses performed in Sections 5.3 and 5.4, which were obtained by solving (4.1) using a Genetic Algorithm (GA) [20].

5.2 Applying the proposed model to a sample patient

To demonstrate how our approach works, we explain the procedure for a particular patient selected from the data. The patient was male, unmarried, less than 5% covered for medical service, and living in zip code 48235. Table 2 shows his appointment information as patterns of show/no-show from 10/13/2009 to 12/31/2009 (training data). Note that no-shows are represented by 1 while shows are represented by 0.

Using the patient's personal and appointment information as well as his previous attendance record, the parameters of the fitted logistic regression model are calculated as shown in Table 3.

Based on the estimated coefficients of logistic regression, the probability of his not showing up for the first appointment in the testing dataset (1/25/2010) is estimated as p=0.3453. This estimate is used for building the prior Beta(0.4353,0.6547) of the Bayesian updating procedure by solving Eq. 7. Table 4 illustrates the updated parameters of Beta distribution as well as the estimated probability of no-show after each appointment.

As graphically illustrated in Fig. 2a, the Bayesian update reacts quickly to each new data record, which means that the procedure can rapidly converge to the real distribution of no-show. Figure 2b compares the prior and posterior distributions of no-show probability before and after applying test data. As can be seen, probability density of the posterior distribution has been moved to the left which can be interpreted as reflecting a decreased probability of no-show.

5.3 Appointment-time based analysis

In this section, we compare the performance of the proposed model with a number of population- and individual-based algorithms based on time-based analysis. For this set of experiments, the training, validation and testing data are

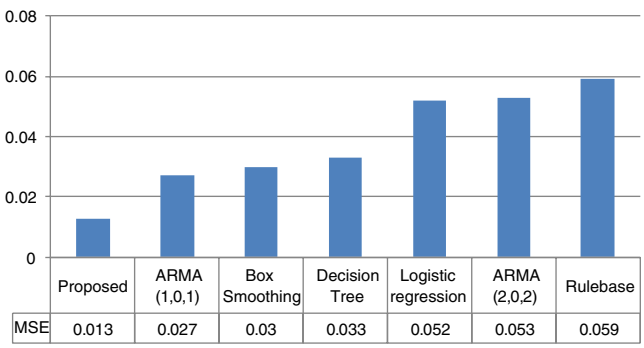


Fig. 8 Mean squared error (MSE) of different methods used for comparison

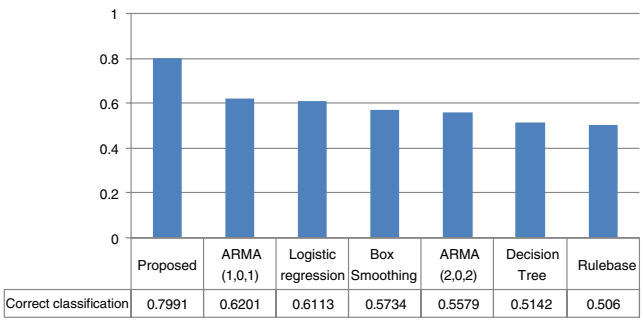


Fig. 9 Percent of correct classification of different methods used for comparison

defined as follows: appointments that occurred before 11/23/2009 have been used for training; appointments between 11/23/2009 and 2/1/2010 have been chosen for validation, and finally, appointments after 2/1/2010 have been considered for testing. The main reason for selecting the above dates is to have approximately 500 data records each in the training, validation and testing datasets.

The methods used in our comparison are as follows: Box smoothing, autoregressive moving average model (ARMA), decision tree, multiple logistic regression (with same predictors as used in the proposed model regression part) and rule base. For setting the parameters of the comparison methods, the size of the moving window in the Box method was studied over the range of 1 to 7 and the optimal size (5) was considered for the comparisons. For the ARMA model, two of the most common models in the literature, namely ARMA (1,0,1) and ARMA(2,0,2), have been considered. Also, J48 and PART algorithms are used for building the decision tree and rule base methods.

Figure 3 compares the mean squared error (MSE) of the methods. Based on the MSE measure, the proposed model performs clearly better than other methods, while the rule-based method has the largest error. As can be seen from the results, in general, individual-based methods outperform population-based methods, while bundling these methods together (as in our proposed method) significantly increases the performance.

Figure 4 also illustrates the percentage of correct classifications for each of the comparison methods, which was done by defining a cutoff value for the output of the methods (the cutoff value is optimized for each of the methods based on the validation dataset). The result of this analysis is very similar to Fig. 3, in which the proposed method performs better than other methods while individual-based methods in general have better performance than population-based methods.

Figures 5, 6 and 7 compare the empirical and estimated probability of no-show for the methods over different patients (the performance of other methods along with the source code is available upon request). As can be seen from Fig. 5a, the proposed approach often predicts the real pattern correctly. This is better illustrated in Fig. 5b, which shows the absolute difference between the estimated and empirical probability of no-show. Here, the mean difference is 0.1104, which is acceptably low. There are also a few cases with absolute difference larger than 0.5, which are related to patients with very few available data records.

Figure 6a illustrates the estimates from logistic regression, a population-based method. The estimates tend to have small fluctuations around an approximately fixed mean. Such result clearly shows that the regression models may not fully capture the difference among patients' personal behaviors. The absolute difference between the estimated and empirical probability of no-show, which is shown in Fig. 6b, also confirms similar results. Here, the mean of the differences is 0.1935 but the maximum difference is 0.8683, which is considerable. Such a result is very similar to other population-based methods discussed earlier.

Finally, Fig. 7a shows the results from the ARMA (1,0,1) model, which is a popular individual-based methods. ARMA is unsuccessful in predicting the no-show patterns for a large portion of patients with real no-show rates larger than zero. This can also be seen in Fig. 7b, which has several differences greater than 0.5 and a few differences equal to 1.

5.4 Patient-based analysis

We also compare our method to other methods proposed in the literature using patient-based analysis. For this purpose, out of 99 patients in the database, using three fold cross validation, approximately 33 patients each were randomly chosen for training, validation and testing. Figures 8 and 9 illustrate MSE and percent correct classification of the methods with results similar to the time-based analysis illustrated in the previous section.

The results from Figs. 2, 3, 4, 5, 6, 7, 8 and 9 clearly show the capability of the proposed model in estimating probability of non-attendance for both current and hypothetical patients of a health care system.

6 Conclusion and future work

Efficacy of any scheduling system depends highly on its ability to forecast and manage different types of disruptions and uncertainties. In this paper, we developed a probabilistic model based on logistic regression and Bayesian inference to estimate patients' no-show probability in real time. We also modeled the effect of appointment date and clinic on the proposed method. Next, based on real-world patient data collected from a Veterans Affairs medical hospital, we evaluated and showed the effectiveness of the approach. Our approach is computationally effective and easy to implement. Unlike population-based methods, it takes into account the individual behavior of patients. Also, in contrast to individual-based methods, it can put together information from the complete database to provide reliable initial estimates. The result of the proposed method can be used to develop more effective appointment scheduling systems and more precise overbooking strategies to reduce the negative effect of no-shows and fill appointment slots while maintaining short wait times.

One of the limitations of the current study is that it considers only one type of disruption, which is no-show, while other

cases such as cancelation and patient lateness can also have a large impact on the performance of the scheduling system. Such cases can be modeled using more sophisticated types of prior distributions, which will be explored in the future.

Appendix - Gaussian Mixture Models (GMM) and Expectation Maximization (EM) Algorithm

Gaussian Mixture Models (GMM) assume data points are drawn from a distribution that can be approximated by a mixture of Gaussian distributions. In this regard, assuming Q, the no-show rate of each clinic, is the feature vector, and k is the number of components (clinic clusters), the mixture model can be rewritten as:

$$p(Q|\{\Theta) = \sum_{i=1}^{k} a_i prob(Q|\theta_i) \tag{11}$$

Where $\{a_1, ..., a_k, \theta_1, ..., \theta_k\}$ is the collection of parameters with $0 \le a_i \le 1, \forall i = 1, 2, ..., k$ and $\sum_{i=1}^{k} a_i = 1$ and $p(Q|\theta_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{Q-\mu_i}{2\sigma_i^2}\right)$. Having as a set of n, i.i.d samples $Q = \left\{q^{(1)}, q^{(2)}, ..., q^{(n)}\right\}$ from the above model the log-likelihood function can be rewritten as:

$$\begin{aligned} &\log p(Q|\theta_i) = \\ &\log \prod_{j=1}^{n} p\left(q^{(j)}|\ \Theta\right) = \sum_{j=1}^{n} \log \sum_{i=1}^{k} \alpha_i p\left(q^{(j)}|\theta_j\right) \end{aligned} \tag{12}$$

Here, the goal is to find Θ that maximizes the log-likelihood function:

$$\widehat{\Theta}_{MLE} = \arg\ \max\{\log\ p(Q|\ \Theta)\} \tag{13}$$

The surface of the above likelihood function is highly nonlinear, and no closed form solution exists for the above likelihood function. One way to deal with this problem is by introducing a hidden variable Z:

$$\begin{aligned} &\log p(Q, Z|\theta_i) = \\ &\sum_{j=1}^{n} \sum_{i=1}^{k} z_i^{(j)} \log\left[\alpha_i p\left(q^{(j)}|z_i^{(j)}\theta_j\right)\right] \end{aligned} \tag{14}$$

and using Expectation Maximization (EM) algorithm as follows [33]:

i. Initializing parameters Θ
ii. Iterating the following until convergence:

$$E - Step:\ Q\left(\Theta|\Theta^{(t)}\right) = E_z \log\left[p(Q, Z|\Theta)|\Theta^{(t)}\right] \tag{15}$$

$$M - Step:\ \Theta^{(t+1)} = \arg\ \max Q\left(\Theta|\Theta^{(t)}\right) \tag{16}$$

References

1. Barron WM (1980) Failed appointments: who misses them, why they are missed, and what can be done. Prim Care 7(4):563–574
2. Bean AG, Talaga J (1995) Predicting appointment breaking. J Health Care Mark 15(1):29–34
3. BechM(2005)Theeconomicsofnon-attendanceandtheexpectedeffectof chargingafineonnon-attendees.HealthPolicy74(2):181–191
4. Bolstad WM (2007) Introduction to Bayesian statistics. Wiley-Interscience, New York
5. Brockwell P, Davis RA (2009) Time series: theory and methods. Springer Series in Statistics
6. Campbell JD, Chez RA, Queen TBA, Patron E (2000) The no-show rate in a high-risk obstetric clinic. J Women's Health Gend-Based Med 9(8):891–895
7. Cashman SB, Savageau JA, Savageau L, Celeste A, Ferguson W (2004) Patient health status and appointment keeping in an urban community health center. J Health Care Poor Underserved 15:474–488
8. Cayirli T, Veral E (2003) Outpatient scheduling in health care: a review of the literature. Prod Oper Manag 12(4):519–549
9. Chakraborty S, Muthuraman K, Mark L (2010) Sequential clinical scheduling with patient no-shows and general service time distributions. IIE Trans 42(5):354–366
10. Cote MJ (1999) Patient flow and resource utilization in an outpatient clinic. Socio-Econ Plann Sci 33:231–245
11. Cynthia TR, Nancy HG, Scott C, Donna SJ, Wilcox WD, Adolesc AP (1995) Patient appointment failures in pediatric resident continuity clinics. Pediatr Adolesc Med 149(6):693–695
12. Dove HG, Karen CS (1981) The usefulness of patients' individual characteristics in predicting no-shows in outpatient clinics. Med Care XIX(7):734–740
13. Dreihera J, Froimovicia M, Bibia Y, Vardya DA, Cicurela A, Cohen AD (2008) Nonattendance in obstetrics and gynecology patients. Gynecol Obstet Investig 66:40–43
14. Evans M, Hastings N, Peacock B (2000) Statistical distributions, 3rd edn. Wiley-Interscience, New York
15. Garuda SR, Javalgi RG, Talluri VS (1998) Tackling no-show behavior: a market driven approach. Health Mark Q 15(4):25–44
16. Glowacka KJ, Henry RM, May JH (2009) A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. J Oper Res Soc 60:1056–1068
17. Goldman L, Freidin R, Cook EF, Eigner J, Grich P (1982) A multivariate approach to the prediction of no-show behavior in a primary care center. Arch Intern Med 142:563–567
18. Gupta D, Denton B (2008) Appointment scheduling in health care: challenges and opportunities. IIE Trans 40:800–819
19. Hassin R, Mendel S (2008) Scheduling arrivals to queues: a single-server model with no-shows. Manage Sci 54(3):565–572
20. Haupt RL, Ellen S (2004) Practical genetic algorithm, 2nd edn. Wiley, New York
21. Hilbe JM (2009) Logistic regression models. Chapman & Hall/CRC Press
22. Hixon AL, Chapman RW, Nuovo J (1999) Failure to keep clinic appointments: implications for residency education and productivity. Fam Med 31(9):627–630
23. Ho C, Lau H (1992) Minimizing total cost in scheduling outpatient appointments. Manag Sci 38(2):1750–1764
24. Kleinbaum DG, Klein M (2002) Logistic regression a self-learning text, 2nd edn. Springer, New York
25. LaGangaLR,LawrenceSR(2007)Clinicoverbookingtoimprovepatient accessandincreaseproviderproductivity.DecisSci38:251–276
26. Lehmann TNO, Aebia A, Lehmann D, Balandraux OM, Stalder H (2007) Missed appointments at a Swiss university outpatient clinic. Public Health 121(10):790–799

27. Liu N, Ziya S, Kulkarni VG (2009) Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. Manuf Serv Oper Manag 12:347–364
28. Moore CG, Wilson-Witherspoon P, Probst JC (2001) Time and money: effects of no-shows at a family practice residency clinic. Fam Med 33(7):522–527
29. Muthuraman M, Lawley M, (2008) A stochastic overbooking model for outpatient clinical scheduling with no-shows I. 40(9):820–837
30. Nadkarni MM, Philbrick JT (2005) Free clinics: a national study. Am J Med Sci 330(1):25–31
31. Rust CT, Gallups NH, Clark S, Jones DS, Wilcox WD, Adolesc A (1995) Patient appointment failures in pediatric resident continuity clinics. Pediatr Adolesc Med 149(6):693–695
32. Simonoff JS (1996) Smoothing methods in statistics, Springer Series in Statistics. Springer, New York
33. Wang J (2009) Encyclopedia of data warehousing and mining, 2nd edn. Information Science Reference, Hershey
34. Bo Z, Turkcan A, Lin J (2010) Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. Ann Oper Res 178(1):121–144