ESSAYS IN APPOINTMENT MANAGEMENT

by

Shannon LaToya Harris

B.S. Systems Engineering, George Mason University, 2007

Submitted to the Graduate Faculty of

The Joseph M. Katz Graduate School of Business in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

THE JOSEPH M. KATZ GRADUATE SCHOOL OF BUSINESS

This dissertation was presented

by

Shannon LaToya Harris

It was defended on

April 20, 2016

and approved by

Luis G. Vargas, PhD, Professor

Jennifer Shang, PhD, Professor

Bjorn P. Berg, PhD, Researcher

Robert C. Hampshire, PhD, Assistant Research Professor

Dissertation Advisor: Jerrold May, PhD, Professor

ESSAYS IN APPOINTMENT MANAGEMENT

Shannon LaToya Harris, PhD

University of Pittsburgh, 2016

Patients who no-show or who cancel their outpatient clinic appointments can be disruptive to clinic operations. Scheduling strategies, such as slot overbooking or servicing patients during overtime slots, may assist with mitigating such disruptions. In the majority of scheduling models, no-shows and cancellations are considered together, or cancellations are not considered at all. In this dissertation, I propose novel prediction models to forecast the probability of no-show and cancellation for patients. I present analyses to show that no-shows and cancellations are two different types of patient behavior, and should be treated separately when scheduling a patient. Additionally, I develop a multi-day, online, overbooking model that incorporates no-show and cancellation probabilities, and outlines how patients should be optimally overbooked in an outpatient clinic schedule to increase clinic service reward. I find that past history is an indicator of future no-show behavior for patients attending outpatient clinics, and that only a limited look-back window is needed in order to gain insight into patient's future behavior. Advance appointment cancellations are more challenging to predict, and tend to occur at the beginning or at the end of an appointment's lifecycle. The optimal overbooking strategy is a function of both the no-show and the cancellation probabilities, and affects both the day on which an overbooking may occur, and the appointment slot in which the patient is overbooked.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank the many people who have helped me through the completion of my Ph.D and this dissertation. First, I would like to acknowledge my advisor, Jerry May. Jerry is honest, hardworking, and dedicated. He provided me with the academic, emotional, and personal support I needed to accomplish my goals over the past five years, and I am eternally grateful to him for his mentorship and kindness. Luis Vargas is also a valuable mentor. He challenged me intellectually, and was always available with advice and new ideas. I would also like to acknowledge the other members of my committee – Jen Shang, Bjorn Berg, and Robert Hampshire – who dedicated time to helping me develop the concepts in this dissertation. I am blessed to have had such a supportive committee, and their continued advice and mentorship is valuable to me.

I would not have been able to complete this journey without the unconditional love and support of my family – Robert, Maxine, and Robyn Harris. My mother was always there to lend an ear to help me talk through my research hurdles, my father always provided encouragement and praise, and Robyn was a constant friend and supporter. I love you all, and dedicate this dissertation to you.

I am thankful to many others who helped me during this journey. Frits Pil for his advice and mentorship, Dennis Galletta for his honesty and proofing abilities, Patrick Connally for always keeping me on track, and all of the other students in my program who talked me through coursework and research challenges. A very special thanks goes out to Carrie Woods. Carrie's mentorship and understanding of everything in life helped me almost daily. I am grateful I met her, and that I can call her a friend. I would also like to acknowledge all of my friends from home who stuck by me, even though I would disappear in my work for lengths of time. Also, special thanks to the friends I've made while in the program, who helped fill my time here in Pittsburgh with amazing experiences and memories.

1.0 INTRODUCTION

Timely patient access to healthcare systems is an on-going problem that is yet to be resolved (IOM 2015). Lengthy patient scheduling queues, and wait times at a clinic, may reduce patient satisfaction, and, perhaps, lead to “poorer health outcomes” (IOM 2015, p. 11). Patient behavior, such as no-shows and cancellations, can lead to schedule inefficiencies, such as underutilization of clinic resources or overtime. Examples of strategies used to mitigate the negative effect of appointment no-shows and cancellations include overbooking and the use of overtime slots. In this dissertation, I present models I have developed, in conjunction with my advisors Jerrold May and Luis Vargas, to predict no-show and cancellation probabilities, and to overbook patients in an outpatient specialty clinic. The models are motivated by the outpatient clinic scheduling practices of the Veterans Health Administration (VHA).

A patient not attending an appointment, a no-show, has been well studied in the outpatient scheduling literature. Cayirli and Veral (2003) list the prediction of no-shows as one of the three major decision levels in a scheduling system. Zeng et al. (2010) state that a no-show model that accurately captures patient behavior is the first step in developing an overbooking scheduling model. Cancellations are discussed less in scheduling literature, and are typically grouped with no-shows. Based upon our knowledge of cancellations, we posit that cancellations differ from no-shows, and should be considered separately. Given the gap in the scheduling literature that includes both no-shows and cancellations, I established a research goal to develop a no-show prediction model that can capture patient behavior, to perform a descriptive analysis on cancellations to determine if they differ from no-shows, to develop a predictive model for cancellations, and to incorporate both type of patient behaviors into an overbooking scheduling model.

Cancellations may be grouped into two categories: advance and late cancellations. The two types differ in their effects on the clinic schedule. Advance cancellations are appointments that are cancelled far enough in advance that the clinic may assume, with a high probability, that the appointment slot freed up by the cancellation may be reassigned to another patient. Late cancellations have a lesser probability of being reassigned, and are, at times, grouped with no-shows (Gupta and Denton 2008). In this dissertation, unless otherwise stated, cancellations always refers to advance cancellations.

In our analysis of patient no-show probability, we found that past attendance history is the most significant predictor of no-show probability. Examples of how past history has been incorporated into a no-show prediction model include using past history as an indicator variable which represents patient attendance for the last appointment (Glowacka et al. 2009), using prior no-show rate over a horizon (Daggy et al. 2010), or using count of previous no-shows (Huang and Hanauer 2014). In our no-show prediction model, we focus on refining the past history variable to potentially improve no-show probability prediction.

When developing the no-show history prediction model, we assume that the sequence of past no-shows, i.e., the order in which they occurred, is a significant factor in determining the no-show probability. Human beings tend to repeat behavioral patterns, but those patterns may change over time. More recent behavior is likely to be more salient than prior behavior, and, after some time, past behavior may no longer be relevant for predicting the future. We build a model that uses a patient's past sequence of successes and failures, over a limited historical horizon, in a regression-like approach, to predict the probability of a success on the next occurrence. Additionally, we develop a metric to determine the amount of past history necessary to make a prediction.

The results of our no-show model validate our assumptions concerning human behavior. We find that there is finite number of past appointments needed to predict no-shows, and that more recent behavior is more relevant than future behavior. The look-back window can be determined based upon a metric that considers the decrease in the sum of squared error between models. We find that within the look-back window, the sequence of past no-shows is relevant up to a point, then the count of no-shows becomes sufficient. The output of our model is a set of coefficients that provide an indication as to the rate at which past behavior becomes increasingly irrelevant.

Analysis of cancellation data revealed that cancellations are less habitual than no-shows, and, should be considered separately in appointment management analysis. A histogram of the number of people who cancel over the course of their appointment lead time reveals that the majority of cancellations tend to occur right after an appointment has been scheduled, or right before the appointment is to occur. Thus, *when* a cancellation occurs during the appointment lifecycle becomes an important factor. Predicting if a cancellation will occur and when it will occur requires predicting a binary and a continuous variable. Typical approaches for such problems are to predict each variable with a separate model. We develop a metric that allows us to predict both variables in a single model. We seek to create an efficient, singular model because this is preferable due to the dynamic nature of scheduling decisions in an outpatient clinic.

Our model is able to perform similarly to a conventional two-phase model approach, while also providing a consistent measure for predictions. The two most significant predictors of time to cancellation is the appointment lead time and past cancellation history. As the lead time of an appointment increases, a patient is more likely to cancel her appointment closer to when she called to make the appointment. Patients with more historical cancellations are more likely to cancel again, and also to cancel closer to when they made the appointment.

The overbooking scheduling model provides strategies to overbook up to two patients per day in a scheduling horizon. We incorporate clinic parameters, including indirect waiting, no-shows, and cancellations, to inform the overbooking decisions. We limit the model’s decision space to determine if and when a patient should be overbooked. The model is restricted to making overbooking decisions, because all other decisions are exogenous to the model. We assume the number of appointment slots is fixed, and that the length of each appointment is constant. In addition, demand for appointments exceeds appointment supply, and all available slots are already filled with patients. In the clinic we observed, clinic schedulers do not differentiate among patients based upon their unique probabilities of no-show and cancellation, so, in our model, we assume homogeneous no-show and cancellation probabilities.

The results of the overbooking scheduling model show that overbooking can benefit a clinic, when overbooking decisions are made in an informed manner. We define informed overbooking as the practice of providers to overbook based on the results of a prescribed analytical model that uses clinic parameters and patient behavior as inputs to direct decision-making. Evaluating scheduling decisions over a multi-day horizon, as opposed to just a single day, allows a clinic to better determine where a patient should be booked, and, under certain conditions, increase the amount of patients allowed into the clinic schedule.

The models presented in this dissertation contribute to the literature on healthcare appointment management in several ways. The no-show prediction model allows a clinic to make managerial decisions concerning the amount of data necessary to make predictions, and to determine how historical occurrences, within a finite window, contribute to future no-shows. The no-show prediction model is a function of the length of the past history considered, not of the number of observations in the data set, so is well-suited for large datasets. The advance cancellation model provides a novel alternative to a two-phase model, to predict if and when a patient will cancel an appointment. Cancellations in a healthcare context are not well studied in literature, and our approach allows for insight into how cancellations differ from no-shows. The overbooking model is novel in its inclusion of both no-shows and cancellations, while scheduling over a multi-day horizon. Our strategies allow a clinic to overbook up to two patients per day, in an informed manner, and potentially increase revenue while also increasing patient access. We show that overbooking is a function of both no-show and cancellation probabilities, and discuss how each of these probabilities effect overbooking decisions.

Currently the overbooking model focuses solely on overbooking patients, not the scheduling of the patients already in the schedule. This formulation is a first step in addressing patient access issues. We plan to develop a scheduling model that informs strategies for booking all patients. Additionally, we plan on extending the current model to include heterogeneous no-show and cancellation probabilities, based upon the output of the no-show and cancellation predictive models.

We thank the U.S. Department of Veterans Affairs for providing financial support with the University of Pittsburgh. This work is an outcome of a continuing partnership between the Katz Graduate School of Business and the Pittsburgh Veterans Engineering Resource Center (VERC). We also acknowledge financial support provided to Shannon L. Harris by the Fryrear Research Fellowship award through the Katz Graduate School of Business.

The remainder of this dissertation is organized as follows. Chapter 2 is the paper written on the no-show prediction model. Chapter 3 is the advance cancellation paper, and Chapter 4 is the overbooking model paper. Chapter 5 concludes, and Appendices are included with proofs and more details of topics discussed in each paper.

2.0 NO-SHOW HISTORY PREDICTIVE MODEL

We present a new model for predicting no-show behavior based solely on the binary representation of a patient’s historical attendance history. Our model is a parsimonious, pure predictive analytics technique, which combines regression-like modeling and functional approximation, using the sum of exponential functions, to produce probability estimates. It estimates parameters that can give insight into the way in which past behavior affects future behavior, and is important for clinic planning and scheduling decisions to improve patient service.

2.1 BACKGROUND

In this chapter, we present an analytical model for predicting the success of the next outcome of a binary time sequence, where the outcome, success or failure, is the result of human behavior. Our model is the result of the consideration of patients’ attendance or non-attendance at a wide variety of medical and surgical outpatient clinics, where outpatient attendance – one example of such a time sequence – is of significant concern. A patient not attending an appointment, a no-show, is disruptive to a clinic, may cause access and scheduling issues because of its effect on clinic capacity, and may increase the cost of clinic operation. Healthcare facilities have the potentially conflicting objectives of providing high-quality service and reducing costs, and the identification and reduction of no-shows assists with both of those objectives (Glowacka et al. 2009, LaGanga and Lawrence 2007). No-show rates vary, but have been reported to range from 3% to 80% (Rust et al. 1995).

The presence of no-shows has also impacted the healthcare scheduling literature. Outpatient clinics fall under a class of service operations that are affected by customer non-attendance (LaGanga and Lawrence 2012). Cayirli and Veral (2003) listed the prediction of no-shows as one of the three major decision levels in a scheduling system. Zeng et al. (2010) stated that a no-show model that accurately captures patient behavior is the first step in developing an overbooking scheduling schema. While the importance of identifying individual patient no-shows is recognized, scheduling models that incorporate the presence of no-shows typically use an average no-show rate for all scheduled appointments (LaGanga and Lawrence 2012, Zacharias and Pinedo 2014), or no-show probability based upon appointment lead time (Liu et al. 2010). LaGanga and Lawrence (2012) used a no-show rate that may differ for each day, and Zacharias and Pinedo (2014) assigned a high or low no-show rate for each patient. Both articles remarked that the schedules produced are improved by permitting heterogeneity in no-show rates. Berg et al. (2014) created an outpatient clinic scheduling model that allows for individual patient no-show probabilities, and found that their inclusion adds more volatility to the scheduling structure. The recognized importance of accurate no-show prediction, for operational planning and scheduling in healthcare and similar service environments, motivated us to build a predictive analytics model to do such predictions.

Prior modeling to predict no-shows ranges from rule-based methods (Glowacka et al. 2009) to logistic regression (Daggy et al. 2010, Huang and Hanauer 2014). The models typically include patient demographic variables, appointment characteristic variables, and a variable representing a patient’s past history. Glowacka et al. (2009) used an indicator variable which represents patient attendance for the past appointment. Additional representations include prior no-show rate over a horizon (Daggy et al. 2010), or count of previous no-shows (Huang and Hanauer 2014). In all models, the prior history variable is found to be significant. Our model focuses on modeling a patient’s past history, in an effort to refine the way it is included in a prediction model

Our model uses past sequences of successes and failures, over a limited historical horizon, in a regression-like approach, to predict the probability of a success on the next occurrence. Human beings tend to repeat behavioral patterns, but those patterns may change over time. More recent behavior is likely to be more salient than prior behavior, and, after some time, past behavior may no longer be relevant for predicting the future. We show how to estimate the parameters of such a model. Because the complexity of our methodology is a function of the length of the history included, not of the size of the data set, it is particularly useful for "Big Data" applications.

In this chapter, we focus on no-show predictions to inform planning and scheduling decisions, but our model is relevant in any service environment that is affected by customer non-attendance or non-participation. Examples of such applications are responses to charitable solicitations, such as the one considered by Fader et al. (2010), changes in employment (Mehran 1989), prediction of recessions (Startz 2008), and airline no-show rates (Lawrence et al. 2003), among others.

We numerically demonstrate the generalizability of our approach using two real data sets: one extracted from outpatient appointment records, and the other involving charitable solicitations. Our approach provides insight into the length of historical behavior that influences future behavior, and the relative importance of each of the observed outcomes in that historical record. In general, we found that the sequence of past successes is important for recent behavior, although the importance of the particular ordering of successes and failures may decrease as outcome recency decreases.

The remainder of this chapter is organized as follows. Section 2.2 includes a review of the related research. In Section 2.3, we present our model. Section 2.4 describes the datasets used for analysis, the model results, and comparisons. Section 2.5 has a discussion, summary, and directions for possible further research.

2.2 LITERATURE

Predicting no-shows based upon past historical values involves the analysis of binary data sequences, so we provide a brief review of the literature on that topic. Approaches to modeling binary data include Markov models (Cox 1981, Berchtold and Raftery 2002) and moving average approaches that incorporate generalized linear models (Zeger and Qaqish 1988, Li 1994, Startz 2008). The simplest Markov models consider only the current state in describing future behavior. If it is assumed, as in our model, that outcomes earlier than the present one are necessary for accurate prediction of future occurrences, a higher-order Markov chain can be developed (Cox 1981), which is subject to the curse of dimensionality (Startz 2008, Prinzie and Poel 2006).

Cox (1981) provided a review of the literature on time series, and proposed several examples of "observation-driven" models, in which the conditional expectation of the present depends explicitly on past data. He proposed that binary data be analyzed using an observation-driven linear logistic regression model, which Startz (2008) termed *BAR*(*p*). The *BAR*(*p*) model uses a logit model, and has *p+1* parameters – a constant value and *p* lagged values. The *BAR*(*p*) model is attractive, as it is linear, and its parameters may be estimated using logistic regression, but Cox stated that it may not be suitable for data with long-range effects. Zeger and Qaqish (1988) extended the *BAR*(*p*) technique, an extension that Startz labeled the *BARX*(*p*) model, to include cross-terms for all *p* lagged values, and a potential for substituting covariate terms for the constant value. Startz stated that, while the *BARX*(*p*) model provides a starting point for moving away from traditional Markov models, it does not perform well when transitions are on the "edge of permissible space" (Startz 2008), that is, transition probabilities that are 0 or 1. Li (1994) proposed another variant of the *BAR*(*p*) model, the *BARMA*(*p*,*q*) model, which adds moving average terms. The *BARMA*(*p*,*q*) model is a focus of (Startz 2008). Startz found that the *BARMA*(*p*,*q*) model performs better than traditional Markov models when predicting U.S. recessions. We build on the autoregressive nature of those models, and include the use of exponential sums to enhance predictions.

Our formulation differs from a typical autoregressive model in the distribution of the errors, model evaluation techniques, assumptions, and the amount of data needed for model evaluation. A *BAR*(*p*) model, as described in Startz (2008), is similar to a logistic regression where the errors are assumed to follow a logistic distribution. The parameters are estimated using techniques such as quasi-maximum likelihood, with no closed form solutions for the coefficients. The data are collected sequentially through time, and the model assumes that the data are equi-spaced. Our model is analogous with a least squares regression, where the errors are assumed to be normally distributed and the coefficients can be directly solved. We do not consider the spacing of the collected data. Additionally, a *BAR*(*p*) model requires more data to generate parameter estimates. Box et al., (2011) state that a minimum of fifty data points is preferred when building an autoregressive model. For applications such as outpatient appointment no-shows and charity donor solicitations, obtaining fifty historical data points for each person is highly restrictive, and thus many people would be excluded from the model. Our model requires a person to have one more data point than the lag number being modeled. This allows for people with one occurrence to still be included in the analysis. Because the use of a *BAR*(*p*) or a *BARMA*(*p*,*q*) model would exclude the majority of our dataset due to the length of data needed to tune the model, we do not directly compare the results of our model with these models.

Two additional models that predict binary data without an exponential increase of parameters are the Mixture Transition Distribution (MTD) model, introduced in Raftery (1985), and the beta-geometric/beta-Bernoulli (BG/BB) model, from Fader et al. (2010). We refer to those models in depth because of their salience, and because of their application to service industries.

The MTD model seeks to predict the next outcome of a binary variable based upon past history. It produces an *m x m* transition probability matrix (TPM), where *m* denotes the number of states, and a vector of lag parameters that allows for each lag to be weighted separately. Probabilities of success are calculated by multiplying the transition probability at each lag with the lag weight, and adding across all lags. This approach is more parsimonious than is a Markov model; the number of parameters is $m(m\text{-}1)+(l\text{-}1)$, where *l* is the number of lags in the model. The MTD parameters can be solved using approaches such as maximum likelihood estimation algorithms, minimum χ^2estimation, or expectation-minimization (EM) algorithms. Extensions to the MTD model are discussed in Berchtold and Raftery (2002), and allow for the use of distinct transition matrices for each past occurrence (MTDg model), and an infinite length history (Mehran, 1989). Applications of the MTD model include employment data (Mehran 1989), financial services (Prinzie and Poel 2006), and non-Gaussian time series (Berchtold and Raftery 2002).

We seek to improve on the MTD approach in several ways. First, due to the iterative nature of the algorithms required to solve for the MTD's parameters, in some cases, an optimal solution may not be reached (Berchtold and Raftery 2002). We believe that a guarantee of optimality is attractive in a prediction setting, especially when predictions may be used to induce operational change. In Section 2.3.2.1, we demonstrate the optimality and uniqueness properties of the solution to our model. Second, the MTD model is not easily implementable. A software program, MARCH, is available online (Berchtold 2005). However, the performance of MARCH deteriorates as the dataset size increases. For example, running the program on a dataset of

473,144 records with nine lags required one hour of CPU time on a high-end desktop computer, indicating that the time involved might be prohibitive on data sets as large as our complete outpatient data file, which has over five million records. “Big Data” applications may well involve even more than five million records. We believe that it is advantageous to create an approach that can be implemented using spreadsheet software, and for which the computation time is not a function of the size of the data set.

Fader et al. (2010) proposed a Bayesian technique, which they termed the beta-geometric/beta-Bernoulli (BG/BB) model, for responses to solicitations by charities. Their approach assumes that the probability of a success (meaning a donation) and the probability of a “death” (a donor becoming permanently inactive) are heterogeneous, and follow a beta distribution. The model uses a binary representation of giving history to tune the model. It assumes that historical sequences with the same number of successes (frequency) and the same last success (recency) produce the same probability for the next outcome. That is, if the history of the system is written with the most recent trial on the left and the least recent on the right, the sequences 11100 and 10101 produce the same probability of success on the next trial.

The BG/BB model is attractive, because the number of parameters to be estimated is the same for any value of k – two for the beta-geometric element and two for the beta-Bernoulli element – and because of its concise representation of the binary time series sequences. Our technique differs from the BG/BB model in two significant aspects. One, our model incorporates structures for capturing situations in which the impact on the future of past occurrences diminishes with increasing time, with greater impact for the more recent occurrences. The BG/BB model is not able to detect such effects. For example, in the sequences mentioned in the above paragraph, if the impact of a success on the second occurrence is large compared to the impact of a success on the fifth, our approach would predict that the sequence 11100 is more likely to be followed by a success than is the sequence 10101. The BG/BB model must assign equal follow-up-success likelihood to both. Two, our formalism incorporates structures that allow for the direct interpretation of the relative effect at each lag, which we believe is integral to a prediction model. The BG/BB model does not have such structures.

2.3 MODEL DEVELOPMENT

Based on our review of the literature on patient attendance in healthcare applications, and on a study of our outpatient data, we anchor our model on two assumptions. While these assumptions may seem implicit in a model that predicts future behavior, we believe that building a model grounded on these assumptions allows us to tailor the model to human behavioral applications such as appointment no-shows.

Assumption 1: Past history is an important determinant of future no-show behavior.

A plethora of literature exists on how patient demographics or appointment characteristics affect no-show behavior. Variables typically identified as being significant include age, gender, appointment lead/delay time, and the number of previous appointments (Bean and Talaga 1995, Garuda et al. 1998). Although an individual's past attendance history has been found to be the most significant determinant of future no-show behavior (Goffman et al. 2015, Garuda et al. 1998, Daggy et al. 2010*)*, past history is not usually represented explicitly. It is typically operationalized as an input variable, usually as an indicator variable for most recent appointment status (Glowacka et al. 2009), or as the fraction of appointments that have been no-shows (Daggy et al. 2010). A predictive model for outpatient no-shows, such as the one described in Goffman et al. (2016), is based on modeling components beyond those incorporated in our model.

Assumption 2: The sequence of past no-shows, i.e., the order in which they occurred, may be a significant factor in determining the probability of the next no-show.

A more parsimonious model might assume that sequences can be grouped based upon total number of successes (no-shows) or time of last success (no-show). Current research has found that the number of previously made appointments assists in predicting no–shows (Cosgrove 1990), with no mention of the ordering of the successes through time. From preliminary analysis of our dataset, we find that the ability of the model to allow for varying importance of at least the most recent lags is essential to model accuracy. As an example, for patients with 5 appointments and 3 successes, with a success on the most recent occurrence (the digit on the far left of sequence), the no-show probability ranges from 0.343 for sequence 10101 to 0.449 for sequence 11100.

2.3.1 Model

Approximating an arbitrary function by a sum of exponential distributions is an established concept (see, for example, Beylkin and Monzon 2005, Beylkin and Monzon 2010, Gatuschi 2012). It has been shown that a finite linear combination of exponential functions constitutes a dense set in the space of continuous functions, and may be used to represent many physical processes such as exponential decay (Pereyra and Scherer 2010) and hospital length of stay (Vasilakis and Marshall 2005, Xie et al. 2005). To model *k* historical sequences with exponential functions, we begin with a general exponential function as in (2.1)

$$f\left(k\right)=\sum_{j=0}^{k}z_{j}e^{-\lambda_{j}k}, \tag{2.1}$$

where $z_j \in \mathbb{R}$ are decay amplitudes and the λ_j are decay rates. If we seek to model with an intercept term, Equation (2.1) becomes

$$f\left(k\right)=z_{0}+\sum_{j=1}^{k}z_{j}e^{-\lambda_{j}k}. \tag{2.2}$$

Solving for z_0, z_j, and λ_j values leads to a nonlinear least squares problem. Several algorithms have been developed to solve for the parameters, using techniques such as singular value decomposition (SVD) and ordinary least squares (OLS), both of which lead to good approximations (Pereyra and Scherer 2010). Because we seek to model probabilities with a model that can be solved to optimality, we work with a modified version of Equation (2.2).

Our objective is to predict the probability of success on the next occurrence of a Bernoulli process that has a non-constant probability of success. The prediction is based solely on a fixed window of past occurrences of the process; the width of the window is denoted by *k*. Because there are two possible outcomes at each time period, and there are *k* prior time periods, there are $i=2^k$ possible *k*-period sequences of zeroes and ones. We denote a success at time period *t* by $X_t=1$ and a failure at time period *t* by $X_t=0$. The predicted probability of a success at time *t*, given the history of the successes and failures over the *k* prior time periods is denoted by

$$\hat{p}_{ik} = \hat{P}\left(X_t = 1 \mid X_{t-1}, X_{t-2}, X_{t-3}, \ldots X_{t-k}\right). \tag{2.3}$$

We want to estimate $\hat{p}_{ik}$ using a sum of exponential functions, as in Equation (2.2). We assign the decay amplitudes, z_j of Equation (2.2), as the zeroes and ones that represent the past history of the process. We denote those as $x_{ijk}, j = 1, \ldots, k$, which represent a success or failure on the j^{th} past occurrence of the i^{th} historical sequence when sequences are of length *k*. In Equation (2.2) the decay rates at each lag, λ_j, are multiplied by the lag number, *k*, which produces λ_j estimates that are scaled by the lag number. We remove this relationship, to allow the λ_j to be on the same scale and to be directly comparable. Considering the adjustments to Equation (2.2) just described, we estimate $\hat{p}_{ik}$ by

$$\hat{p}_{ik} = z_{0k} + \sum_{j=1}^{k} e^{-\lambda_{jk}} x_{ijk} \tag{2.4}$$

For each possible sequence in the history of length *k*, denote by v_{ik} the proportion of the observations that have the historical sequence *i*, and by p_{ik} the proportion of those observations that were followed by a success on the next occurrence. For a given value of *k*, to solve for the intercept z_{0k} and the $\lambda_{jk}, j = 1, .., k$, we use the technique of weighted least squares, and minimize F_k, where F_k is given by:

$$F_k = \sum_{i=1}^{2^k} v_{ik} \left(p_{ik} - z_{0k} - \sum_{j=1}^{k} e^{-\lambda_{jk}} x_{ijk} \right)^2 \tag{2.5}$$

The squared errors are weighted by the v_{ik} to account for the frequency of each sequence in the dataset, and in order to permit successes and failures in the population to have different likelihoods. For a look-back window of width *k*, there are *k+1* variables to solve for in the model of Equation (2.5). To ensure that the model produces values that may be interpreted as probabilities for all sequences, the minimization of (2.5) is performed subject to the following constraints:

$$z_{0k} \geq 0 \tag{2.6}$$

$$0 \le z_{0k} + \sum_{j=1}^{k} e^{-\lambda_{jk}} \le 1 \tag{2.7}$$

Equation (2.5) allows each lag to have a unique coefficient, and, therefore has *k+1* decision variables. For some datasets, it could be optimal to have fewer decision variables, and allow for the coefficients, after some point in the look-back window, to be identical. Modeling the data in this way indicates that, after some point in the past, the total number of successes is sufficient to provide insight; the ordering of the successes is not necessary. This allows for the estimation of fewer decision variables, and for the data to be divided into fewer data sequences. To account for groups of lags that have the same effect, we add an additional constraint to (2.5) that allows a block of the λ_{jk} values to be equal. We refer to the number of distinct λ_{jk} values as *k'*, where *k'* is a value between 1 and *k*. For each *k*, we generate *k* models denoted by $F_{k',k}$. For example, when *k=3*, we predict the three models shown in (2.8).

$$\begin{aligned}
F_{1,3} &= \sum_{i=1}^{2^3} v_{i3}\left(p_{i3} - z_{03} - e^{-\lambda_{13}} \sum_{j=1}^{3} x_{ij3} \right)^2 \\
F_{2,3} &= \sum_{i=1}^{2^3} v_{i3}\left(p_{i3} - z_{03} - e^{-\lambda_{13}} x_{i13} - e^{-\lambda_{23}} \sum_{j=2}^{3} x_{ij3} \right)^2 \\
F_{3,3} &= \sum_{i=1}^{2^3} v_{i3}\left(p_{i3} - z_{03} - \sum_{j=1}^{3} e^{-\lambda_{j3}} x_{ij3} \right)^2
\end{aligned} \tag{2.8}$$

When *k'* is equal to *k*, $F_{k',k}$ is equal to F_k, and all the λ_{jk} are distinct. When *k'* is equal to one, all the λ_{jk} are equal, and a success at any lag contributes the same amount to the predicted probability of a success at time *t*. To determine the optimal *k'* value for any value of *k*, we use a BIC equation tailored for regression, where

$$BIC = n \ln\left(\frac{SSE}{n} \right) + \left(k + 2 \right) \ln\left(n \right) \tag{2.9}$$

(Burnham and Anderson 2002). The coefficient of ln(*n*) is *k+2*, to account for the estimation of the intercept and the estimation of the model variance. We use the BIC metric because the Weighted Sum of Squared Error (WSSE) is not adjusted for the number coefficients that have been estimated, so WSSE decreases monotonically as *k'* increases. We refer to the model of

Equations (2.5), (2.6) and (2.7) as Sums of Exponentials for Regression (SUMER), because our $\hat{p}_{ik}$ values are estimated using a regression model, where the coefficients of the regression are modeled by exponential functions.

2.3.2 Key Results

2.3.2.1 Optimality

For any given value of *k*, taking the first partial derivatives of (2.5) with respect to z_{0k} and to the $e^{-\lambda_{jk}}, j = 1,..,k$, and setting them equal to zero, yields a system of *k+1* linear equations in *k+1* unknowns. To solve for z_{0k} and $e^{-\lambda_{jk}}, j = 1,..,k$, in general, it is necessary to solve a linear system of the form $A_k s_k = b_k$, where s_k is the coefficient vector, and A_k is the Hessian of the objective function F_k in (2.5). The matrix A_k is positive definite (Theorem 2.1, below), so A_k is invertible. Thus, the system of linear equations can be solved by Cramer's Rule, and $s_k = A_k^{-1} b_k$. Details of Cramer's Rule and model formulation can be found in Appendix A.

Because constraints (2.6) and (2.7) are linear, SUMER is a convex optimization problem. When A_k is positive definite, the objective function of (2.5) is strictly convex, and s_k is a unique global minimizer (Griva et al. 2009). Theorem 2.1 below shows that when at least one sequence is represented in a dataset, A_k is positive definite. Because the reduced parameter models are equivalent to adding equality constraints to the original model, proof of uniqueness of the original model holds for all reduced parameter models.

Theorem 2.1. If $v_{ik} > 0$ for at least one *i*, then the matrix A_k is positive definite for all values of *k*.

Proof. SUMER can be written in matrix form as in Equation (2.10).

$$F_k = V_k \left(P_k - X_k s_k \right) \tag{2.10}$$

In this formulation, V_k is a $\left(2^k \times 2^k\right)$ diagonal matrix with the sequence counts along the diagonal, P_k is a vector of observed values, s_k is a coefficient vector, and X_k is a $\left(2^k\right) \times \left(k+1\right)$

design matrix with the first column containing all ones, and subsequent columns containing the binary sequences of length k. In general, a $(n \times n)$, symmetric matrix, *H*, is positive definite, if, for a vector $w \neq 0$, $w'Hw > 0$ (Horn and Johnson 2012).

The Hessian of F_k, $A_k = X_k{}'V_k X_k$, is a $(k+1)\times(k+1)$, symmetric matrix. For $w \neq 0$, $w'\left(X_k{}'V_k X_k\right)w = \left(X_k w\right)' V_k X_k w = \left\|V_k^{1/2} X_k w\right\|_2^2 > 0$, when V_k is non-zero. Therefore, the Hessian of F_k is positive definite for all values of k.

Corollary 2.1. All s_k are global minimizers of F_k.

Proof. Follows because the Hessian matrix A_k is positive definite for all *k*.

2.3.2.2 Parameter Interpretation

Because the x_{ijk} values are binary, SUMER estimates the probability of success at time period *t* by adding exponential terms for every time period at which there was a success. We chose to use the exponential distribution to model the coefficients because the parameters and model are easily interpretable. The value of z_{0k} is the predicted probability of success on the next occurrence, when all past occurrences have been failures. Intuitively, as *k* increases, z_{0k} should decrease, and approach zero from above, i.e., $z_{0k} \underset{k\to\infty}{\longrightarrow} 0^+$. Thus, as the history of all failures becomes longer, the lower the predicted probability of success should be on the next occurrence, as the person has established a more consistent pattern of failures.

The terms $e^{-\lambda_{jk}}, j = 1,..,k$ are comparable to the typical "beta" coefficients in a regression. They represent the change in the probability of success between a success and a failure at lag *j* when all other lags are held constant. If we assume that more recent behavior is likely to be more salient than prior behavior, we would expect $e^{-\lambda_{jk}}$ to monotonically decrease as *j* increases. A monotonic decrease would indicate that more recent successes have a greater impact on the probability of success at the next occurrence. Because $e^{-\lambda_{jk}}$ can never be negative, a success at lag *j* will always increase the probability of a success on the next occurrence;

failures contribute no value to the probability of success. Modeling when this assumption does not hold is an extension of the model outlined in Section 2.3.2.3.

While we can gain historical insight from the $e^{-\lambda_{jk}}$ values, we can gain additional insight by interpreting the decay rates, λ_{jk}. For a given value of *k*, λ_{jk} represents the rate at which a success at time *j* produces a success at time *t*. By Equation (2.7), λ_{jk} will always be greater than or equal to zero, because $e^{-\lambda_{jk}}$ is greater than one for $\lambda_{jk} < 0$. The greater the value of λ_{jk}, the faster the effect of a success at lag *j* decays and approaches zero. So, for larger λ_{jk}, lag *j* is less relevant to the outcome at time *t*. There is a different vector of rates for each look-back window. The rate vectors for any two values of *k* are assumed to be independent, and are modeled with separate exponential distributions. For a fixed value of *j*, as the length of the look-back window increases, we expect the coefficient $e^{-\lambda_{jk}}$ to decrease, because there are more historical time periods contributing to the probability estimate, that is, for a given value of *j*, we expect λ_{jk} to increase as *k* increases.

2.3.2.3 Model for Handling Special Datatypes

SUMER, as described above, incorporates the assumption that past successes (ones) in the historical sequences will have a positive effect on the probability of success in the future. SUMER generates a probability prediction by adding the predicted probability for the sequence of all past failures, z_{0k}, with the coefficients, which will always be positive. If, as an example, the probability of success for the all failure sequence is greater than the probability of success for the all success sequence i.e., $p_{1,k} > p_{2^k,k}$, then SUMER must try to estimate $z_{0k} > \left(z_{0k} + \sum_{j=1}^{k} e^{-\lambda_{jk}} x_{ijk} \right)$, which will produce estimates for all sequences equal to z_{0k}. Datasets for which it would be necessary to predict $p_{1,k} > p_{2^k,k}$ are datasets where there is a “ping-pong effect”, and a success (one) at a lag reduces the baseline probability of success, z_{0k}, as opposed to increasing it. A donation dataset with donors who contribute sporadically, rather than consistently, might cause such an effect. For such persons, once they have contributed, they do not contribute again until, perhaps, their charitable budget has been replenished.

To permit SUMER to accommodate such datasets, an additional set of parameters is applied at each lag to allow lags to have a negative effect. Equation (2.11) shows the extended model with the additional parameters, α_{jk}.

$$\sum_{i=1}^{2^k} v_{ik}\left(p_{ik} - z_{0k} - \sum_{j=1}^{k} \alpha_{jk} e^{-\lambda_{jk}} x_{ijk}\right)^2 \tag{2.11}$$

Additional constraints, $-1 \leq \alpha_{jk} \leq 1$ and $e^{-\lambda_{jk}} > 0$, are required to estimate the parameters. The constraints on the α_{jk} are added to bound the parameter space. Because they are bounded to be between -1 and 1, we can interpret them as proportions, as described below. The second constraint is analogous to Equation (2.7) above. Given that an additional set of parameters, α_{jk}, have been added, the first-order conditions are now nonlinear functions. We solve the model using an iterative algorithm.

Table 2.1 displays a sample dataset, along with the BG/BB, MTDg, SUMER and Extended SUMER predictions. The addition of the α_{jk} parameters noticeably improves SUMER's performance, and permits probability estimates less than z_{0k} to be generated. The extension to SUMER is therefore important in permitting it to model "ping-pong" behavior.

Table 2.1. Extended SUMER Analysis on Sample Data

Sequence	Counts	Actual Probability	SUMER Prediction	Extended SUMER Prediction	MTDg Prediction	BG/BB Prediction
000	7,026	0.800	0.608	0.800	0.800	0.029
001	559	0.240	0.608	0.240	0.240	0.163
010	281	0.730	0.608	0.730	0.730	0.264
011	535	0.170	0.608	0.170	0.170	0.356
100	295	0.729	0.608	0.730	0.730	0.304
101	286	0.171	0.608	0.170	0.170	0.570
110	392	0.661	0.608	0.660	0.660	0.570
111	1,730	0.100	0.608	0.100	0.100	0.835

Table 2.2 lists the parameters generated from each model. The original SUMER model assigns a coefficient of zero to each of the past occurrences. If we interpret $e^{-\lambda_{jk}}$ as the amount the

baseline probability changes with a success at lag *j*, α_{jk} is the percentage of $e^{-\lambda_{jk}}$ that is used to change the baseline probability. When $\alpha_{jk} = -1$, 100% of the predicted $e^{-\lambda_{jk}}$ value decreases the baseline probability when there is a success at lag *j*. When $-1 < \alpha_{jk} \le 0$, then $(100 * \alpha_{jk})\%$ of the predicted $e^{-\lambda_{jk}}$ value decreases the baseline probability when there is a success at lag *j*. When $0 < \alpha_{jk} \le 1$, a success at lag *j* still increases the baseline probability at the rate λ_{jk}. Because the model is designed with this case in mind, the increase will typically be modeled with the rate parameter, λ_{jk}, and the associated α_{jk} will be one. For example, for datasets where the α_{jk} parameter is not necessary, the generated solution vector for λ_{jk} using the extension is identical to the solution vector when SUMER is utilized, and all $\alpha_{jk} = 1$.

Table 2.2. Parameters for Model Extension

	λ_{13}	λ_{23}	λ_{33}	z_{03}	$e^{-\lambda_{13}}$	$e^{-\lambda_{23}}$	$e^{-\lambda_{33}}$	α_{13}	α_{23}	α_{33}
SUMER	20.00	19.76	20.00	0.608	0	0	0			
Extended SUMER	1.818	2.234	0.000	0.799	0.162	0.107	1	-0.431	-0.655	-0.559

SUMER is a generalization of Extended SUMER, with all α_{jk} assumed to be one. An analyst can determine if Extended SUMER is necessary by analyzing the table p_{ik} values. When p_{1k} is greater than any other probability, Extended SUMER should be implemented.

2.3.3 Specification of *k*

If the same amount of historical data is available for all cases in the data, then, after a *k'* model has been chosen within each *k*, it is possible to compare the chosen models across *k* to determine the optimal width of the look-back window. In this section, we present an approach to select that value of *k*. We assume that the optimal *k* value will be chosen based upon analysis of a training dataset. For new cases with historical sequences of length less than *k*, the approach of this section could be used to rank-order the attractiveness of the models that are feasible for that case. At least one historical sequence is needed to use SUMER as an analytical tool. New patients with no history could be assigned an average probability of success, or a probability of success based upon patient demographics and appointment characteristics.

In Section 2.3.1, we showed how to choose the optimal *k′* for each *k*. In general, we use the BIC metric of Equation (2.9) to determine the optimal width of the look-back window, *k*. We also develop a heuristic to choose *k* when the BIC values do not achieve a minimum value in the interior of the range of look-back windows considered.

Using a BIC criterion, the optimal value of *k* is the smallest *k* value for which $BIC_k \le BIC_{k+1}$. Our empirical experience shows that for very large values of *n*, the small decrease in SSE that is achieved by increasing *k*, results in a small decrease in BIC, and BIC steadily decreases across all *k*. Thus, it is preferable to choose *k* based on achieving a sufficiently large decrease in SSE, rather than by requiring only a decrease in BIC. Such an approach is also used in clustering routines to determine the optimal number of clusters (Chiu et al. 2001).

From Equation (2.9), $BIC_k < BIC_{k+1}$, when $\ln\left(SSE_k\right) - \ln\left(SSE_{k+1}\right) < \frac{\ln(n)}{n}$, or when $\frac{SSE_k}{SSE_{k+1}} < \exp\left(\frac{\ln(n)}{n}\right)$. When *n* is large, $\frac{\ln(n)}{n}$ becomes very small, and $\exp\left(\frac{\ln(n)}{n}\right) \cong 1$. In this situation, even a minimal decrease in SSE will result in a decrease in BIC, and the SSE is expected to decrease with an increase in *k*. To implement the heuristic, we stop increasing *k* when there is not a sufficient change in SSE when going from *k* to *k+1*, that is, when $\frac{SSE_k}{SSE_{k+1}} < 1 + \delta$, where $0 \le \delta \le 1$ is a user-specified parameter.

2.4 NUMERICAL COMPARISONS

We used data sets from two different service operational environments to assess the performance of SUMER on real data. The first data set is the charity donor data from Fader et al. (2010) which was used to validate the BG/BB model. The second data set was extracted from the attendance of outpatients at a Veterans Health Administration (VHA) facility. We denote the charity donor data by DO and the outpatient dataset by OP.

2.4.1 SUMER Results for DO

DO includes data collected from 1996-2006, and includes 11,104 records. For each solicitation, a donation is coded as 1 and a non-donation as 0. The data were split into training and test datasets. The training dataset consists of historical donation activity from 1996-2004 to predict the donation activity in 2005, so a look-back window may range from one to nine. The model was tuned on the training dataset, an optimal *k* value between one and nine was chosen, and the model was tested on the activity of 2006 based upon the optimal- *k* past occurrences.

First, models were run on the training data for all *k'* and *k* combinations to determine the preferred model. With a maximum *k* value of nine, there were $\frac{9(10)}{2} = 45$ models to tune. SUMER was programmed in Wolfram Mathematica 10, and solutions for the 45 models were found in 49.67 seconds. Solutions for a single model were calculated in less than 2 seconds. Table 2.3 shows the optimal *k'* value chosen at each *k* and the calculated BIC value. The $BIC_k \le BIC_{k+1}$ at *k=8*; thus, the optimal width of the look-back window for DO is eight.

Table 2.3. DO BIC values for optimal k' for a look-back window of size k

k	*k'*	BIC	*k*	*k'*	BIC
1	1	-41,505.73	6	5	-51,393.22
2	2	-48,080.27	7	5	-51,551.08
3	3	-49,841.52	**8**	**5**	**-51,649.01**
4	3	-50,866.05	9	6	-51,595.51
5	5	-51,219.33			

Table 2.4 lists the decay rate (λ_{jk}) and the coefficient ($e^{-\lambda_{jk}}$) values for the model when *k=8* and *k'=5* (see Appendix A for the variable values from all models). For lags five through eight, the total number of donations, and less the ordering of the donations, is significant for predicting the

probability of success for the next occurrence. As an example, the probabilities of success for sequences 10000110 and 10000011 would both be equal to 0.4286, because they have two donations in years represented by lags five through eight. The intercept value indicates that if a person has failed to donate during the eight prior years, the probability that he or she will donate in the current year is 0.0073. In contrast, a person who donated in *each* of the past eight years will donate in the current year with probability 0.8428. That value is calculated by adding the coefficient values across all lags.

Table 2.4. Decay rates and coefficients generated from SUMER for k=8 for DO

	constant	**lag 1**	**lag 2**	**lag 3**	**lag 4**	**lag 5-8**
Decay rates		0.9840	1.5303	2.4345	2.7700	3.7400
Coefficients	0.0073	0.3738	0.2165	0.0876	0.0627	0.0238

As expected, the coefficients decrease as *j* increases. That decrease causes the most recent lags to have more of an effect on the outcome at time *t*. Additionally, holding the lag value constant, the λ_{jk} increase as *k* increases, up to *k=8* (this can be seen in the values in Appendix A). Lag 1 is the most influential lag, with a coefficient value that is 44% of the largest possible probability, and almost twice as large as the coefficient value at lag 2. The probability of giving decreases from 0.8428 to 0.4690 when a person did not donate the previous year. A person is 3.33 times less likely to give in the next solicitation if he or she gave in lag years three through eight, but not in lag years one and two. Lags five through eight, collectively, contribute only 11% to the largest possible probability, thus soliciting people who have only donated within that timeframe would prove to be less fruitful. As a final insight, we find that, while the frequency of successes is sufficient for analysis within lags five through eight, this does not hold throughout lags two through four. This is evidenced when looking at the predicted probabilities for sequences 11100000 and 10000011, which are 0.6852 and 0.4286, respectively. While both of these sequences contain a donation in the most influential lag, knowledge of the timing of the other two donations increases the probability for sequence 11100000 by 59.8%.

To evaluate the model's performance, we use the model at the optimal value of *k*, and calculate the probabilities of donation in 2006 based upon 1998-2005 historical data, and the area under the ROC curve (AUC). We compare the AUC for our model with the AUC derived from the empirical probabilities used to tune the model. We provide this comparison because the empirical probability values are easily calculated, they are typically utilized when a model is not available, and they may provide a contrast between utilizing a descriptive analytics technique and a predictive analytics technique. To compare AUCs, we use the nonparametric approach of DeLong et al. (1988), which detects differences among two or more models based on the areas under their ROC curves. We use the DeLong et al. method because the ROC curves for the models are correlated, as they were applied to the same dataset.

The AUC for SUMER equals 0.9160, and the AUC for the empirical probabilities equals 0.9018. Using the DeLong method, the SUMER AUC is statistically greater than the empirical AUC (p<0.0001), thus SUMER is the preferred prediction model for the DO dataset.

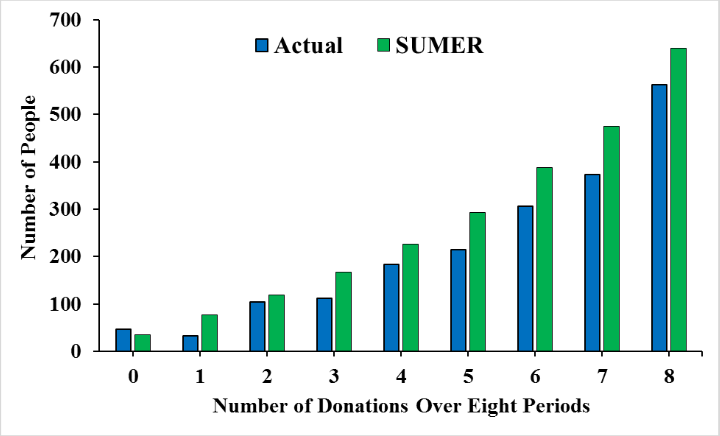


Figure 2.1. Predicted versus Actual Number of Donations for the DO Dataset

As an additional evaluation of model performance, we compare the expected number of people who are predicted to make zero to eight donations, as calculated from SUMER at $k=8$ and $k'=5$, with the actual number from the test dataset. Figure 2.1 illustrates the comparison. The pattern of the actual distribution indicates that, as the number of donations increases, the number of people who donate increases. Given the properties of a regression model, the total expected donations will equal the total actual donations. SUMER estimates balance across all donation levels.

2.4.2 SUMER Results for OP

OP is derived from the show/no-show behavior of patients at a Veterans Health Administration (VHA) facility from Fiscal Year 2007 to Fiscal Year 2012. A maximum of sixteen past appointments for each patient were tallied, with a total of 4,760,733 appointment sequences generated. The MTDg model required more than 72 hours to estimate parameters for the model with all 4,760,733 records. Thus, a subsample of 473,144 sequences was used to train the all models, to allow for comparison with the MTDg model. The training dataset consists of appointments one through fourteen to predict the no-show on the fifteenth appointment. We tested the model on the no-show realization of the sixteenth appointment. For a maximum *k=14*, there were $\frac{14(15)}{2}=105$ models to run. The program took 2,506 seconds in Wolfram Mathematica 10 for all 105 models and under 3 seconds for the model when *k=9*. Table 2.5 lists the optimal *k'* value chosen at each *k* and the calculated BIC value.

Table 2.5. OP BIC values for optimal k' for a look-back window of size k

k	***k'***	**BIC**	***δ***	***k***	***k'***	**BIC**	***δ***
1	1	-2,527,199.72	0.1579	8	5	-2,813,135.15	0.0547
2	2	-2,596,550.86	0.1143	**9**	**5**	**-2,838,317.65**	**0.0436**
3	3	-2,647,762.47	0.0909	10	7	-2,858,500.91	0.0375
4	3	-2,688,913.79	0.0768	11	8	-2,875,912.33	0.0378
5	5	-2,723,880.86	0.0739	12	12	-2,893,412.70	0.0305
6	4	-2,757,648.72	0.0637	13	10	-2,907,634.92	0.0257
7	7	-2,786,807.25	0.0572	14	11	-2,919,617.95	

The BIC values in Table 2.5 are all negative, and decrease steadily as *k* increases from one to fourteen. To determine the optimal value of *k*, we use the heuristic from Section 2.3.3 and calculate $\delta = 473{,}144 * 0.0000001 = .0473$. At that value, we select *k=9*, with *k'=5*. Similar to the DO data, lags five through nine have the same increase on the probability of success. Table 2.6 lists the decay rates and the coefficients for the preferred model.

Table 2.6. Decay rates and coefficients generated from SUMER for k=9 for OP

	constant	lag 1	lag 2	lag 3	lag 4	lag 5-9
Decay rates		1.7399	2.8001	3.0003	3.1275	3.2202
Coefficients	0.0421	0.1755	0.0608	0.0498	0.0438	0.0399

There appear to be similar trends in the coefficients in both datasets. In OP, appointments that are more recent have a greater effect on the probability of a no-show at the next appointment than do less recent appointments. The probability of a success following a history of all failures – a patient showing up for all past appointments – is greater in OP than it is in DO, and the probability of a success following a history of all successes sequence is less in OP than it is in DO. The most recent occurrence is also significant for OP. If it is a success, it contributes 30.7% to the maximum possible probability of a success on the next occurrence. The second most recent outcome has less weight; the probability of a success on the next occurrence is only 1.7 times less if a patient has no-showed for all appointments as compared with no-showing for all but the most recent two. Lags two through nine have similar coefficients, ranging from 7% to 10% of the total possible probability on the next occurrence. As a result, sequence difference in occurrences two to nine time periods previous do not result in noticeably different predicted probabilities of success on the next occurrence. For example, the increase in probability between sequences 111000000 and 100000011 is 10.3%, even though the timing of the successes, except for the most recent one, is as different as possible, and the sequences have the same number of successes.

The AUC values for SUMER and for an empirical probability table on the sixteenth appointment, based upon behavior of the seventh through fifteenth appointments, are 0.7064 and 0.7066, respectively. The empirical table has a greater AUC by 0.000195; but that difference is not statistically significant at the $\alpha=0.5$ level. Thus, we conclude that SUMER is preferred to an empirical table for this dataset also, given the insight provided by the parameter values.

Figure 2.2 depicts the number of people who are predicted to have a particular number of no-shows over nine periods versus the number of people who actually had that number of no-shows, for the OP dataset. The overall pattern in Figure 2.2 is the opposite of the pattern in the DO dataset; the number of people who no-show is inversely related to the number of no-shows.

SUMER, with *k=9* and *k'=5*, follows this pattern and balances out the expected number no-shows across the nine periods.

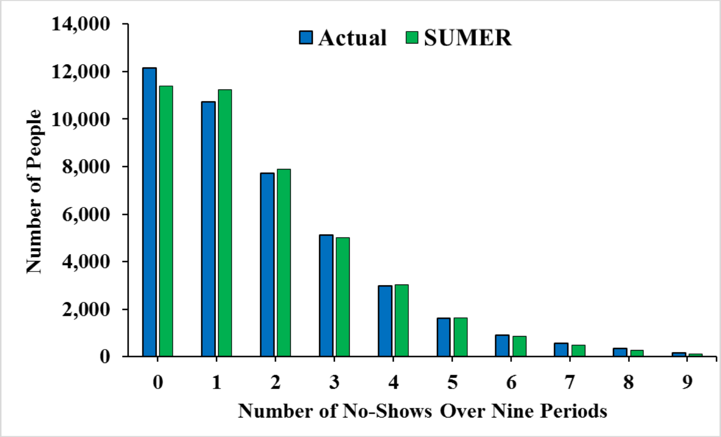


Figure 2.2. Predicted versus Actual Number of No-Shows for the OP Dataset

2.4.3 SUMER Parameter Analysis

As expected, the decay rates and coefficients for the two datasets differ. There are several differences in the datasets that can cause these contrasts. First, the success rate is greater in the DO data set than it is in the OP data set. For DO, the success rate is 23% in the training set and 17% in the test set. For OP, the success rate is 8.8% in the training set and 8.9% in the test set. A greater overall success rate results in greater coefficient values, and can lead to a greater number of influential lags. The DO dataset has two influential lags, both of which have greater coefficient values than the most influential lag in the OP dataset. The total overall probability in the DO data set is 67.8% greater than the corresponding probability in the OP dataset. While the difference in the width of the optimal look-back window also contributes to differences in coefficient sizes, a similar pattern holds comparing the values for *k=8* for both datasets. The greatest predicted probability of success that the model can produce using the OP coefficients is 0.5717. That value is produced for a patient who has missed all nine previous appointments. Because patients, as a whole, typically attend medical appointments, the data records show that even a person with a poor recent attendance record still has a substantial probability of showing up for his/her next appointment.

For both DO and OP, the sequence with the greatest frequency is the all failure sequence. For DO, the sequence 00000000, meaning that the contacted person did not donate on any of the eight previous solicitations, contains 38% and 44.2% of the total data for the training and test sets, respectively, with a probability of donation on the ninth occurrence equal to 0.009. For OP, the sequence 000000000, meaning that the patient attended all nine previous appointments, contains 57.2% of the total data, with a 0.045 probability of non-attendance (success) on the tenth appointment, for both training and test. Those characteristics lead to two rich insights. First, repeated failures lead to different outcomes for the datasets. DO has a greater overall probability of past success, but a lower probability of future success, for the all-failure sequence. For this dataset, the data represent a situation in which repeated refusals to donate are a strong signal towards future refusals. OP has a lower overall probability of past success, but a greater probability of future success for the all failure sequence. OP is signaling that repeated shows still could produce a no-show on the next sequence. Such a difference might be due to the nature of medical appointments, where life circumstances could still cause even an excellent attender to no-show on the next appointment.

Second, a model, such as SUMER, induces overall patterns in a data set, and therefore is more likely to be able to continue to produce accurate probability estimates as the records in a data set change from time period to time period. An empirical table is more likely to be influenced by particular idiosyncrasies that are present at the time it is constructed.

2.4.4 Model Comparison

For additional analysis of SUMER's predictive ability, we compared SUMER with MTDg, BG/BB, and two traditional methods used for binary data analysis, Logistic Regression (LR) and Classification and Regression Trees (CART). We again used DeLong's method of comparison, as opposed to a WSSE or the Brier score (Brier 1950), because DeLong's method allows for the analysis of statistical differences. We coded the BG/BB model in Excel as per Fader et al. (2010). We used the MARCH software (andrewberchtold.com) for the MTDg calculations. The LR and CART models were estimated in IBM SPSS Statistics 21. To do a direct comparison of SUMER and BG/BB, we ran the BG/BB model for a history of length *k+1*, and calculated a weighted average of the probabilities generated for the two sequences with the same first *k* appointment orderings. For example, to generate BG/BB predictions for the OP dataset at *k=10*, we ran the BG/BB model for a history of length 11, and calculated a weighted average of the probabilities generated for the two sequences with the same first 10 appointment orderings.

Table 2.7. AUCs for SUMER, MTDg, BG/BB, LR, CART, and Table Probabilities on DO and OP test data

	DO			OP		
	AUC	Std. Error	*p*-value	AUC	Std. Error	*p*-value
SUMER	0.9160	0.00330	-	0.7064	0.00138	-
MTDg	0.9162	0.00330	0.1630	0.7064	0.00138	0.3740
BG/BB	0.9133	0.00333	0.0009	0.7026	0.00137	<.0001
LR	0.9159	0.00330	0.4753	0.7063	0.00138	0.4050
CART	0.9126	0.00341	<.0001	0.6798	0.00137	<.0001
Table	0.9018	0.00395	<.0001	0.7066	0.00138	0.062

Table 2.7 lists the AUCs, standard errors, and the *p*-value of a χ^2 test to determine if SUMER's AUC is statistically greater than the AUC of the other models. SUMER is significantly superior to the BG/BB model on both datasets. Recall that DO is the dataset that was used to tune and to test the BG/BB model. The BG/BB model incorporates the concept of "death" for the failure case. "Death," for DO, connotes that a person has become inactive, and will no longer donate. For OP, "death" implies that a patient has permanently stopped attending his or her appointments. Such a permanent change in behavior might be due to actual death, or to a change in behavior brought about by a change in attitude or health status. While the OP dataset was shown to have characteristics that would be beneficial for the BG/BB model, the assumption that recency and frequency are sufficient to accurately predict outcomes did not provide adequate enough estimates for either dataset. SUMER and MTDg have AUCs that are not statistically different for both models, but both are greater than BG/BB and the Table representation. SUMER has an advantage over MTDg in its ability to handle large datasets. For OP with 4,760,733 records, all models are able to compute estimates in less than 1 minute, except for MTDg which took over 72 hours.

SUMER is significantly superior to CART for both datasets. CART also lacks the interpretability of the SUMER parameters. The output of CART is a tree or association rules that can be followed to associate each lag to the next occurrence. This type of output lacks direct relatability of the weight of each lag, which is available with the SUMER parameter estimates. SUMER and LR are not statistically different for both DO and OP, but both are greater than BG/BB and the Table for DO. The primary advantage of SUMER over LR is also the interpretability of its parameters. Because we model a human behavioral process, we assume that the coefficient values will decrease as the lag value increases, so occurrences that are more recent have a greater effect. As shown in Tables 2.3 and 2.4, even though SUMER is not constrained to produce decreasing parameter estimates, it is able to produce estimates that follow the assumed behavioral trend for DO and OP. Table 2.8 lists the LR coefficients. Each coefficient represents the change in the log-odds of a success at time *t*, all other coefficients held constant. The constant value represents the log-odds of a success at time *t* when there are no successes in the look-back window. For both datasets, the log-odds values are not ordered, and therefore, do not fit the inherent structure of the behavioral process.

Table 2.8. LR Coefficients for DO (k=8) and OP (k=9)

	constant	**lag 1**	**lag 2**	**lag 3**	**lag 4**	**lag 5**	**lag 6**	**lag 7**	**lag 8**	**lag 9**
DO	-3.50	2.04	1.30	0.65	0.48	0.33	0.17	0.27	0.31	
OP	-2.91	1.33	0.49	0.42	0.38	0.35	0.36	0.34	0.34	0.40

As an additional evaluation of model performance, we calculated cumulative gain for each model. To calculate gain, the predictions for each model are rank ordered from the greatest probability of success to the lowest, and the data are split into equally sized groups. Gain is calculated as the percentage of the total successes represented in each group. Gain values are cumulative across groups, such that the gain of the last group is 1.

From a managerial standpoint, it is preferable to reach the greatest number of successes in the fewest number of trials. Therefore, a larger gain value for the top few groups is ideal. DO was split into five groups to calculate the cumulative gain, so each group contains 20% of the 11,104 records. Table 2.9 lists the cumulative gain values for the top three groups.

Table 2.9. Gain Values for SUMER, MTDg, BG/BB, LR, CART, and Table Probabilities on DO test data

Group	SUMER	MTDg	BG/BB	LR	CART	Table
1	**0.7160**	0.7150	0.7047	0.7157	0.7153	0.7015
2	0.9554	0.9548	**0.9556**	0.9544	0.9541	0.9448
3	**0.9785**	**0.9785**	**0.9785**	**0.9785**	0.9745	0.9681

The greatest values in each row are highlighted in bold. SUMER's predictions for Group 1 allow 71.6% of the total donors in the dataset to be targeted by contacting only 20% of DO's test set, that is, 1,386 of the 1,936 donors will be targeted by contacting only 2,221 people. The results for LR and CART are similar, with 1,385 and 1,384 donors being targeted in Group 1. For Group 2, BG/BB is the only model that has a greater cumulative gain than SUMER. For Group 3, SUMER, MTDg, BG/BB, and LR have identical values. The results in Table 2.9 indicate that the subset comprised of the top three groups contains 97.85% of the donors in DO's dataset for all three models.

OP was split into ten groups to perform gain analysis. The values for the top five groups are listed in Table 2.10. For Group 1, the Table probabilities provide the greatest gain, followed by SUMER and MTDg. For the three models, Table, SUMER, and MTDg, targeting 10% of OP's test set – 47,314 patients – allows a clinic to target 31.96%, 31.85%, and 31.83%, respectively, of patients who no-show. Those percentages represent 13,513, 13,465, and 13,458 patients, respectively. For groups two through five, the model with the greatest cumulative gain varies. By subset five, SUMER, MTDg, BG/BB, and LR all contain 74.93% of the total no-shows in the dataset.

Table 2.10. Gain Values for SUMER, MTDg, BG/BB, LR, CART, and Table Probabilities on OP test data

Subset	SUMER	MTDg	BG/BB	LR	CART	Table
1	0.3185	0.3183	0.2987	0.3181	0.3100	**0.3196**
2	0.4849	0.4854	0.4697	0.4851	0.4538	**0.4858**
3	0.5933	0.5933	**0.5933**	0.5926	0.5804	0.5932
4	0.6871	**0.6876**	0.6869	**0.6876**	0.6449	0.6876
5	**0.7493**	**0.7493**	**0.7493**	**0.7493**	0.7041	0.7492

2.5 DISCUSSION AND CONCLUSIONS

In this chapter, we presented a new predictive analytics model to address binary data that evolve from human behavioral processes, by combining regression modeling with sums of exponential functions. We focus on the properties of a human behavioral process, because such data have a random component and a habitual component that is associated with the user. Those components, properly understood, may be used to inform planning and scheduling decisions. We contribute to the literature on predictive analytics for binary data sequences in several ways.

One, we present a parsimonious prediction model that combines regression-like modeling and the use of the sum of exponential functions to produce probability estimates. The use of a regression-like approach allows the model to be easily understood by a practitioner, and to produce parameters that can be used in interpreting human behavior. Those characteristics are valuable, because we seek to both predict future occurrences and to explain those occurrences, based upon past realizations of a process. The coefficients of the induced model provide insight, derived from the data, as to how much of a person's past behavior should be included when predicting how he/she will behave in the immediate future. The coefficients also provide an indication as to the rate at which past behavior becomes increasingly irrelevant. The empirical differences we observed when applying the model to two real data sets, one of charitable donations and one of outpatient attendance, show the flexibility of SUMER to adapt to datasets with different underlying human behavioral patterns.

Two, we established that the process for estimating SUMER's coefficients yields an optimal and unique solution. The estimated parameters minimize the weighted sum of squared errors of the model as outlined in Equations (2.5) through (2.7), and satisfy the Karush-Kuhn-Tucker (KKT) conditions. This result is desirable, given that our model is motivated by the need to obtain accurate no-show predictions that can be used as an input to larger prediction model or a scheduling model.

Three, the computational complexity of SUMER model is a function of the length of the past history considered, not of the number of observations in the data set, so that SUMER is particularly well-suited to Big Data applications. Given the influx of large datasets in analytics, and the desire to process information quickly, a model which can perform well regardless of dataset size is beneficial. SUMER is able to achieve this, while also producing a mechanism to determine when past history can stop being considered.

Four, our model acts as a valuable input to planning and scheduling decisions in a service operations environment, such as an outpatient healthcare clinic. The use of an average no-show probability, or no-show rate, for all patients does not give proper insight into patient heterogeneity, and therefore does not serve to address the uncertainty and volatility that no-shows present in a system. SUMER is a novel predictive analytics model that can be used in conjunction with patient demographics and appointment characteristics to provide a reliable estimate of patient no-show probabilities. Because the model was constructed in a way that does not directly depend on the application domain, it should generalize well to other service operations that would benefit from the reliable identification of customer behavior over time.

Several extensions of the SUMER model might be addressed by future research. One, the current structure of the SUMER model does not treat the time between occurrences as a parameter. Because human behavior may be affected by time lapse as well as the number of past occurrences, incorporating the time between outcomes might enhance the quality of the model's predictions, and may allow for additional insight. Intuitively, an increased lag time should decrease the effect a past incident has on the next outcome. Two, the SUMER model might be modified so that it is able to predict multiple future outcomes, not only the next outcome. Three, the current models are built solely on the basis of successes; failures do not adjust the predicted probabilities either up or down. In situations where prior failures provide information about future successes, the model's parameter space could be expanded to reflect that information.

3.0 ANALYSIS OF ADVANCE CANCELLATIONS

We perform a descriptive analysis of no-shows and cancellations to determine if no-show and cancellation probabilities are similar enough to justify combining the probabilities in appointment management analysis. We then present a model to predict advance cancellations, as well as how far in advance of the appointment such cancellations occur. Key factors in the model include prior cancellation behavior and appointment lead time. Our single-step approach performs similarly to conventional data mining methods. The model is validated using data from VA Healthcare System outpatient clinics.

3.1 BACKGROUND AND PAST RESEARCH

Failure of patients to keep an outpatient clinic appointment may cause clinic inefficiencies due to underutilized resources. Additionally, in a clinic where patient demand exceeds appointment supply, an appointment that is not completed can exacerbate a clinic's access issues. One of the factors that has been found to effect patient access is patient attendance behavior. The failure of a patient to attend an appointment, a no-show, and its effect on a clinic schedule has been well-studied (see Cayirli & Veral (2003) and Gupta & (2008) for reviews). A patient behavior that is less studied is appointment cancellations.

During an appointment's lead time – the time between when an appointment is made and when it is to occur – a patient can call into a clinic to cancel her appointment. If the appointment is cancelled with enough time to allow another patient to book the same appointment slot, there is less disruption to a clinic. We refer to these cancellations as advance cancellations. The length of time before an appointment date that allows it to be designated as an advance cancellation is clinic dependent. Huang and Zuniga (2014) performed a survey of 40 clinics, and found that

45% of the clinics have a 24-hour advance cancellation policy. If appointments are cancelled too close to the appointment time, the appointment slot may go unused. We refer to these cancellations as late cancellations. The literature typically groups late cancellations with no-shows. Gupta and Denton (2008) refer to late cancellations and no-show research as open research challenges in the outpatient appointment management literature. In the remainder of this chapter, unless otherwise specified, cancellation refers to advance cancellations.

Alaeddini et al. (2015) and Norris et al. (2014) developed multinomial logistic regressions to predict probabilities of show, no-show, and cancellation. Alaeddini et al. (2015) used the same predictor variables for all probabilities, and Norris et al. (2014) analyzed the effects of variables on the three outcomes. Norris et al. (2014) found that the most influential factors to nonattendance are appointment lead time, patient past history, age, and financial payer. Reid et al. (2015) also used prior cancellation history in their logistic regression model and found it to be a significant predictor of no-shows. Galluci et al. (2005) also used a logistic regression model, but predicted no-show and cancellations together as a single variable. Chariatte et al. (2008) constructed a Double-Chain Markov model to predict missed appointments. While they did not specifically predict cancellations, they found that adding cancellations as a predictor variable allows for a more precise model of missed appointments.

Liu et al. (2010) and Parizi and Ghate (2016) accounted for no-shows and cancellations in scheduling applications. Liu et al. (2010) developed a cancellation distribution that is based upon the length of a patient's appointment lead time; the longer the lead time the greater the probability of cancellation. Parizi and Ghate (2016) assumed a similar relationship between lead time and cancellation probability, but also accounted for the type of appointment that is being scheduled.

We hypothesize that cancellations – especially advance cancellations – should not be grouped with no-shows when conducting patient attendance analysis or scheduling applications. The behavior and demographics of patients who cancel and those who no-show vary, and must be understood separately. Additionally, because the timing of a cancellation during the lead time has an effect, this must also be studied in addition to if a cancellation will occur. We also hypothesize that if late cancellations and no-shows are the same, then late cancellations either a) act as a substitute for no-shows and should be inversely related with a patient's probability of no-show or b) patients who cancel late also no-show and thus the two probabilities should be directly related.

In this chapter we first conduct a descriptive analysis of cancellations and no-shows to test our hypotheses. The analysis is performed on data from a Veteran's Health Administration (VHA) clinic. The results of our descriptive analysis indicate that no-shows and cancellations are not statistically similar, and do not have equal sample medians. We analyze the sample medians, because analysis of the sample distributions led to the conclusion that the data are not normal. Additionally, patient demographics such as age group, gender, and marital status have varying effects on the probability of no-show and cancellation within our sample. We also find that the time to cancel an appointment is not related to the probability of no-show, and thus should also be modeled separately. Based upon this finding, we then build a model to predict the fraction of the lead time that expires before a patient cancels, or made to cancel ratio (MTCR), in an effort to gain more insight into cancellations.

3.2 HYPOTHESIS DEVELOPMENT

3.2.1 The Correlation between No-show and Cancellation Probabilities

Correlation measures the linear relationship between two variables. Measuring the correlation between no-show and cancellation probabilities allows us to determine the association between them. If it is a valid assumption that cancellations should be grouped with no-shows, the two variables should have a linear relationship. Additionally, this relationship should be a positive relationship. Thus, as one probability increases, the other should follow the same trend. Therefore, knowing the cancellation probability would allow an analyst to infer similar behavior of the no-show probability, and justify only accounting for one of the factors. We assume that cancellations should not be grouped with no-shows, or excluded from analysis. Given these arguments, we propose the following hypothesis:

Alternative Hypothesis 1 (H1). No-show probability does not have a positive correlation with cancellation probability.

3.2.2 Comparison of No-show and Cancellation Sample Distribution Medians

Comparing the medians of the sample distributions of the no-show and cancellation probabilities allows us to infer if the medians of the two population distributions can be assumed to be equal to each other. The sample medians were compared because analysis of the sample distributions led to the conclusion that the data are not normal. If they are equal, then combining the no-show and cancellation probability distributions into a single distribution, or excluding the cancellation distribution is justified. Based upon these points, we propose the following hypothesis:

Alternative Hypothesis 2 (H2). The no-show probability and cancellation probability distributions do not have statistically equal sample medians.

3.2.3 Analysis of Variance of No-show and Cancellation Probabilities

To study a few of the factors that influence no-shows and cancellations, we analyze the effects of three patient demographic variables: age group, gender, and marital status. In recent studies, these variables have been found to have an effect on no-show probabilities (Chariatte et al. (2008), Norris et al. (2014), Bean & Talaga (1995), Daggy et al. (2010)). If cancellations and no-shows are similar, then we can assume that the effect of age group, gender, and marital status will be the same for both probabilities. If not, they will have varying effects, and should be considered separately. Thus, our final hypothesis is as follows:

Alternative Hypothesis 3 (H3). Different demographic groups have different no-show and cancellation rates.

3.3 RESEARCH METHODS

3.3.1 Data Operationalization

To test our hypotheses, we used de-identified, administrative data derived from the VHA corporate data warehouse. Data recorded appointments from multiple types of outpatient clinics from 2011 to 2016. The data included the date an appointment was requested, the date it was scheduled, if it was cancelled, when it was cancelled, the patient's gender, age group, and marital status. We define a cancellation as any appointment that was recorded as cancelled before the appointment's scheduled date and time. Appointments that were not cancelled and not attended are referred to as no-show appointments.

The dataset contained approximately 2.2 million appointment records. Records with missing demographic information, or incorrect information, such as a negative lead time, were removed. Same day appointments were also removed, to allow the analysis to include only appointments for which a patient had sufficient lead time to cancel without it being considered a late cancellation.

In order to have sufficient data to compare the two probabilities, we removed all records for patients who did not have at least ten appointments. Harris, May & Vargas (2016) find that, for a similar dataset, nine historical occurrences were sufficient to determine how a patient's past history will affect her future no-show behavior. Our final analysis was completed on a total of 35,895 patients who scheduled 1,592,923 appointments. Analysis was completed in Statgraphics Centurion XVII. Table 3.1 lists details for each of variables used in the analyses.

Table 3.1. Operationalization of Variables Used in Analyses

Name	Description	Operationalization	Min	Max	Std. Dev.
Probability of Cancellation	Fraction of a patients total no-showed, cancelled, and completed appointments that were cancelled before the appointment was to occur	Continuous	0	1	0.127
Probability of No-show	Fraction of a patients total no-showed and completed appointments that were not attended	Continuous	0	1	0.175
Age	Age of the patient at the time of the last appointment in the data sample	Categorical (Under 65 (45.05%), 65-85 (48.51%), Over 85 (6.45%))	19	99	15.088
Gender	Gender of the patient in the data sample	Categorical (Male (92.92%), Female			

Name	Description	Operationalization	Min	Max	Std. Dev.
		(7.08%)			
Marital Status	Marital Status of the patient in the data sample	Categorical (Married (44.97%), Never Married (15.72%), Uncoupled (39.31%))			
Made to Cancel Ratio (MTCR)	Fraction of the appointment lead time that passes before a patient cancels the appointment; average for each patient	Continuous	0.0009	0.9999	0.183

3.3.2 No-show and Cancellation Probabilities

We calculated a no-show and cancellation probability for each patient in our sample. To calculate these probabilities, we first calculated the appointment total as the sum of cancelled, no-showed, and completed appointments for each patient.

3.3.2.1 Probability of Cancellation

The probability of cancellation for a patient was calculated by dividing the total number of cancellations for each patient by the appointment total. Because patients could have cancelled none or all of the appointments, probabilities of 0 and 1 are permitted. Figure 3.1 displays a histogram of the cancellation probabilities for all 35,895 patients.

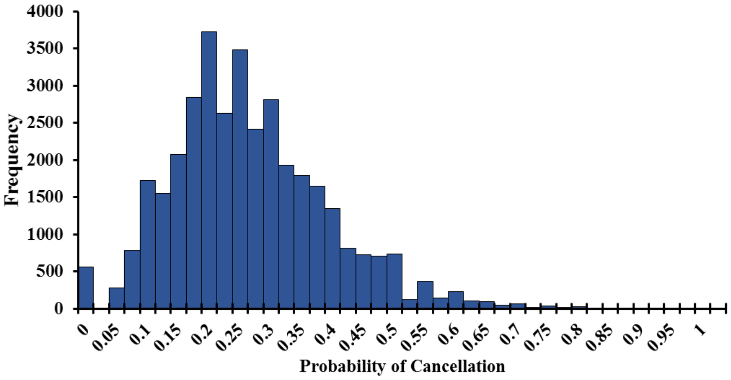


Figure 3.1. Histogram of Cancellation Probabilities

The sample of cancellation probabilities has an average of 0.255 and median of 0.239, so the majority of patients tend to not cancel their appointments. Of the 35,895 patients, 557 cancelled no appointments, and 5 cancelled all of their appointments. The number of appointments per patient ranged from 10 to 1023. The correlation between number of appointments and probability of cancellation is -0.0984 with a *p*-value of zero. The distribution that best fits the sample is a Largest Extreme Value distribution with a mode (α) of 0.1963 and a scale (β) of 0.1068. The pdf of the Largest Extreme Value distribution is given by: $f(x) = \frac{e^{(\alpha - x)/\beta - e^{(\alpha - x)/\beta}}}{\beta}$. The log-likelihood of the fit is 24,529.9; Figure 3.2 displays the Quantile-Quantile plot.

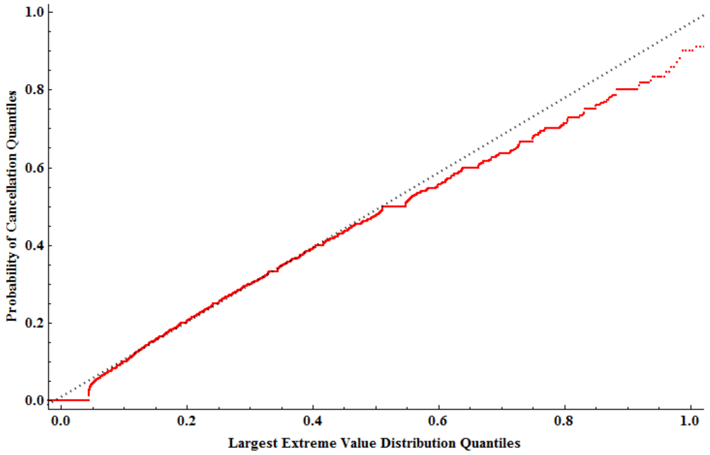


Figure 3.2. Q-Q Plot of Fitted Distribution versus Probability of Cancellation

3.3.2.2 Probability of No-show

To calculate the probability of no-show, we divide the total number of no-showed appointments by the total number of completed and no-showed appointments for each patient. This represents a situation where cancellations are not considered in the analysis, or are grouped with the initial no-show calculation, and not added separately to the appointment total. This calculation allows us to compare no-shows as represented in literature, and cancellations as they would be calculated if they were to be included in the analysis. The drawback to this type of calculation is that the probability of no-show for patients who have cancelled all of their appointments is not defined. Thus, for the 5 patients who cancelled all of their appointments, we set their probability of no-show to zero.

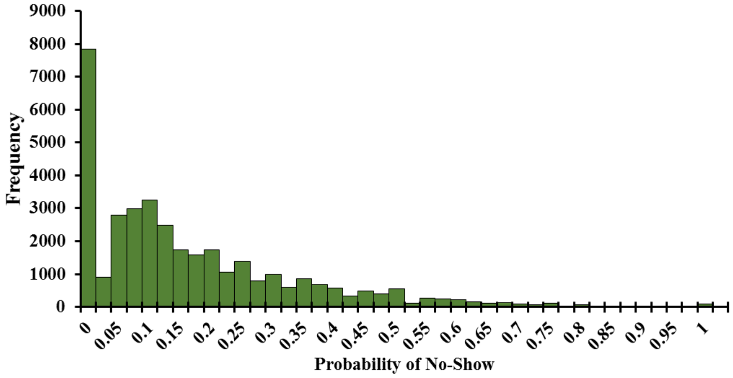


Figure 3.3. Histogram of No-show Probabilities

The mean and median of the no-show probability sample are 0.1607 and 0.1042, respectively. So, on average, patients tend to cancel their appointments more than they no-show. Similar distribution fitting analysis was done for the no-show probability sample. Figure 3.3 displays the histogram of the sample. The best fitting distribution for this sample is an Exponential distribution with $\lambda=0.1607$, and pdf $f(x)=\lambda e^{-\lambda x}$. The log-likelihood of the fit is 29,719.2. Figure 3.4 displays the Quantile-Quantile plot.

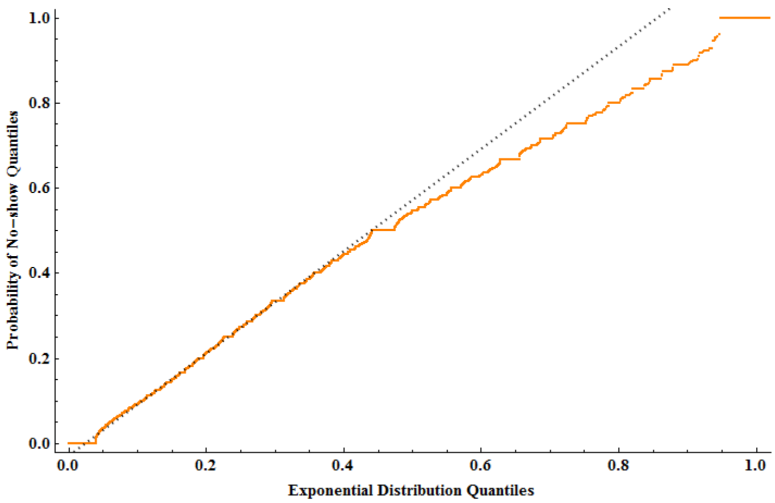


Figure 3.4. Q-Q Plot of Fitted Distribution versus Probability of No-show

Figure 3.5 displays a 3D histogram of the no-show and cancellation probabilities, with probabilities grouped in intervals of width 0.1. Each bin is left-closed and right-open, such that the minimum of the interval is included in the interval and the maximum of the interval is not. For the interval 0.9 to 1, both 0.9 and 1 are included in the interval. The *x*-axis of each histogram is the probability of no-show, and the *z*-axis is the probability of cancellation.

Due to the range of frequencies within each group, Figure 3.5 is split into four separate panels in Figure 3.6. If cancellation and no-show probabilities are interchangeable, then the bins with the greatest frequencies should occur where the no-show and cancellation probabilities intervals are the same. The panel in the upper left-hand corner of Figure 3.6 has the greatest frequencies. This panel represents no-show and cancellation probabilities in the interval [0, 0.5); so, the majority of patients fall in the lower probability groups. The greatest frequency is 5,024, which is the number of patients whose no-show probability falls in the interval [0, 0.1), and cancellation probability falls in the interval [0.2, 0.3). There are only 58 patients whose probability of cancellation is in the interval of [0.8, 1]; all of the frequencies that are 0 occur where a patient's probability of cancellation is falls within this interval.

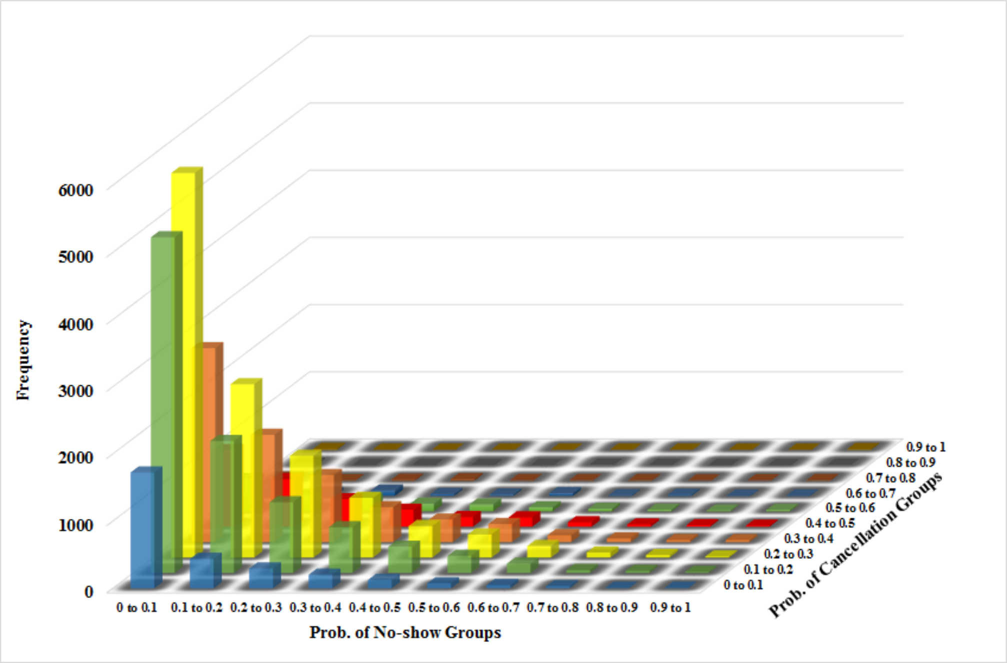


Figure 3.5. 3D Histogram of No-show and Cancellation Probabilities

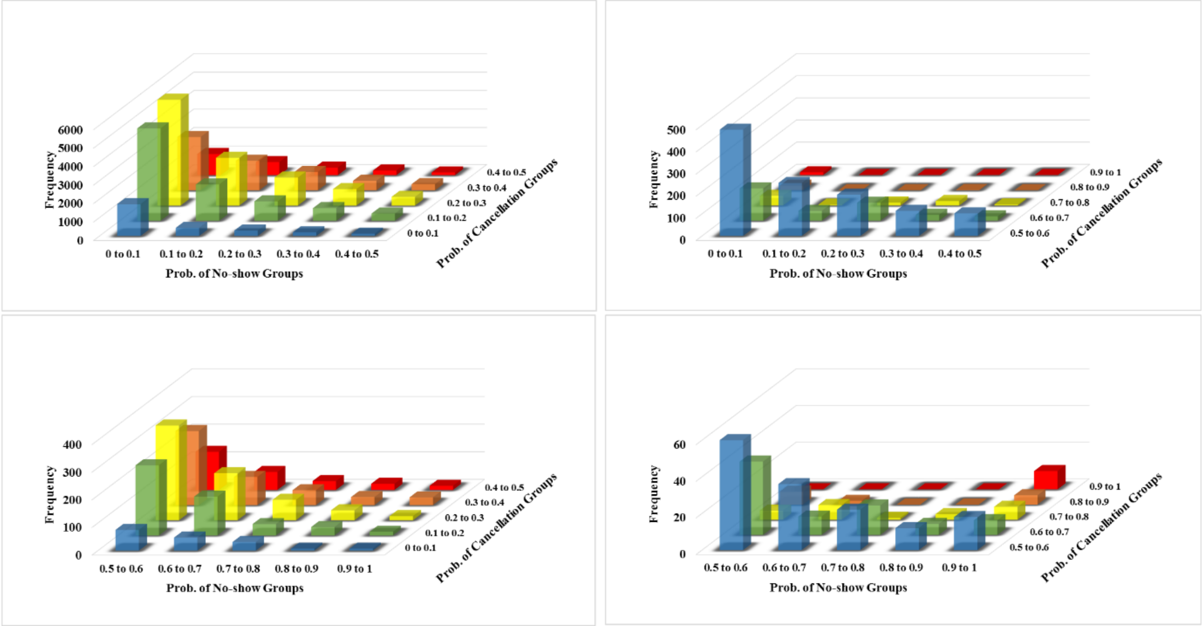


Figure 3.6. 3D Histogram of No-show and Cancellation Probabilities Split into Four Groups

3.3.3 Demographic Variables

3.3.3.1 Age

The age of the patient was collected for each scheduled appointment. Due to data collection restrictions, the age of each patient, in our sample, is her age as of her last made appointment in the sample. Thus, we chose to not include Age as a continuous variable, but create three age categories to perform analyses. Based upon prior analysis of Age (Davies et al. 2016), we chose to separate the data into three buckets: Under Sixty-Five, Sixty-Five to Eighty-Five, and Over Eighty-Five.

3.3.3.2 Marital Status

Marital status – as of the time of the appointment – is recorded in the VHA as Married, Never Married, Separated, Divorced, or Widowed. All records that had missing or unknown marital status were removed from the analysis. Marital status has been found to relate to no-show probability (Daggy et al. 2010), and we believe, speaks to the support system that someone has to encourage them to attend their appointment. Based upon the results in Goffman et al. (2016), we combine the Separated, Divorced, and Widowed marital statuses as Uncoupled. Uncoupled thus refers to a patient who was once married, but is now, for any of the three reasons, not married. This patient may still have access to the support system that was in place when they were married, and thus, may exhibit different tendencies then a Never Married person.

3.3.3.3 Gender

The Gender variable is recorded as Male and Female. The majority of the population we sampled are Male. We have found this is typical of the overall VHA patient demographic, and thus our sample is a representative sample in terms of gender.

3.4 DATA ANALYSES AND RESULTS

3.4.1 Hypotheses Testing

3.4.1.1 Testing of H1

In H1, we hypothesize that there is no correlation between no-show probability and cancellation probability. To test this hypothesis, we analyzed a scatterplot of the data and calculated the Spearman Rank Correlation between the two variables. Spearman Rank Correlation was used, as opposed to Pearson Correlation, because the distributions of the samples are not normal.

Figure 3.7 is a scatter plot of the two probabilities with the probability of cancellation on the *x*-axis and probability of no-show on the *y*-axis. This scatterplot represents the paired no-show and cancellation probabilities for each of the 35,895 patients in the sample. The grey line is a reference line that represents a perfect positive correlation between the two probabilities. In general, the data do not look to follow a linear pattern. There are patients who do not cancel, but no-show at various levels, and vice versa. The majority of the patients fall either above or below the reference line.

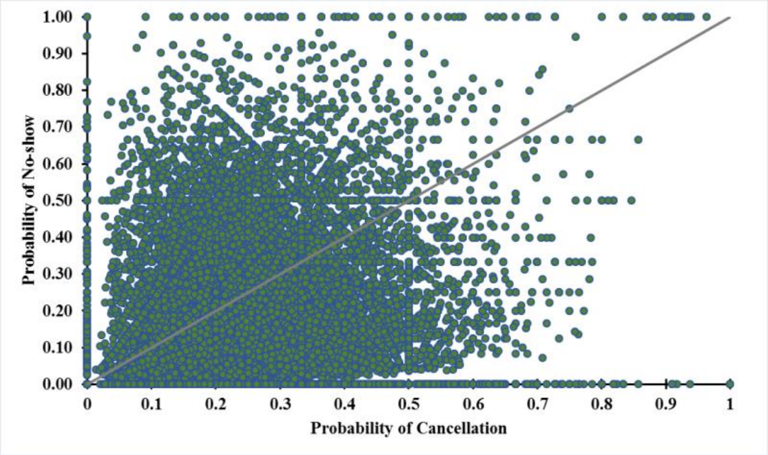


Figure 3.7. Scatterplot of Probability of No-show vs. Probability of Cancellation with a Perfect Positive Correlation Reference Line

Table 3.2 lists the results from the Spearman Rank Correlation analysis. This analysis exhibits similar results to the analysis of the scatterplot. The correlation coefficient of 0.1170 was found to be statistically significant, but the value of the coefficient does not suggest a relationship between the two probabilities. Additionally, due to the sample size, we would expect to find a significant correlation (Berger 1985). Because the coefficient is positive, we expect both probabilities to increase together, so a patient with a higher cancellation probability can be expected to also have a high no-show probability.

Table 3.2. Spearman Rank Correlation between the Probability of Cancellation and No-show

	Probability of Cancellation
Probability of No-show	0.1170 (0.000)

To gain more insight into the correlation between no-show and cancellation probability, we performed correlation analysis on all two-way combinations of the demographic factors, and on discrete groups of the cancellation and no-show probabilities. Figure 3.8 through Figure 3.10 display scatterplots of the data for each pair of demographic factors. The legend on each panel gives the two categories plotted, the count of the number of points in the group, the correlation coefficient, and the *p*-value for the Spearman Rank correlation test.

Females consist of 7.08% of the sample, so, the groups with Female as a factor have fewer points; the maximum number of points analyzed when Female is a factor is 2,110. All but two combinations with Female – 65 to 85 & Female and Uncoupled & Female – have negative correlations, although only one is statistically significant at a 95% confidence level – Never Married & Female. The only other negative correlation occurs in group Never Married & Under 65, but the correlation coefficient is -0.0004 and it is not statistically significant. All other combinations are positively correlated; the only correlation that is not statistically significant is Never Married & Over 85, which has a small sample size of 119. The greatest significant correlation occurs in the Never Married & 65 to 85 combination, with a coefficient of 0.2185; the least is 0.0470 for the combination Under 65 & Male.

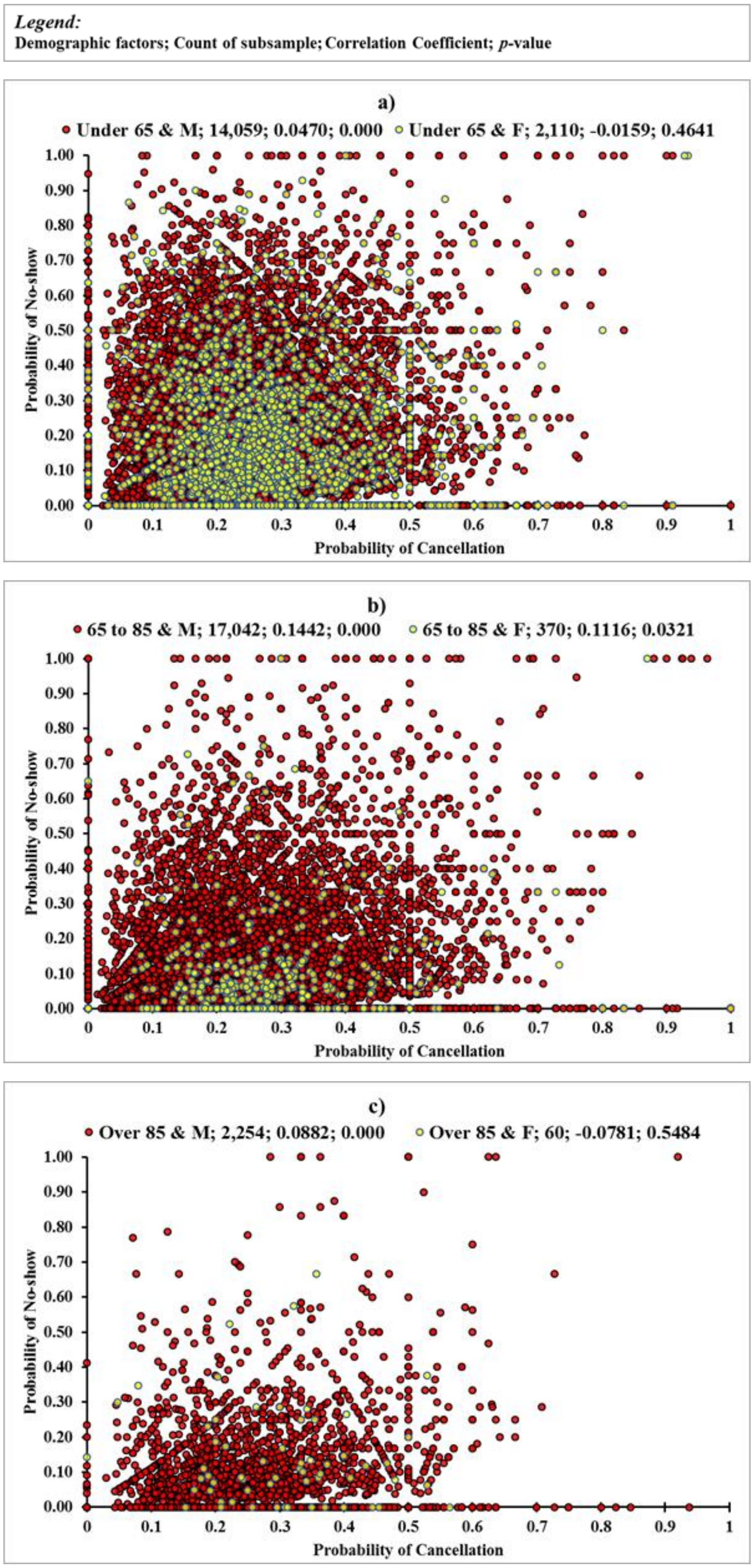


Figure 3.8. Scatterplot of Probability of No-show vs. Probability of Cancellation for Patients **a)** Under 65 and Gender, **b)** 65 to 85 and Gender, and **c)** Over 85 and Gender

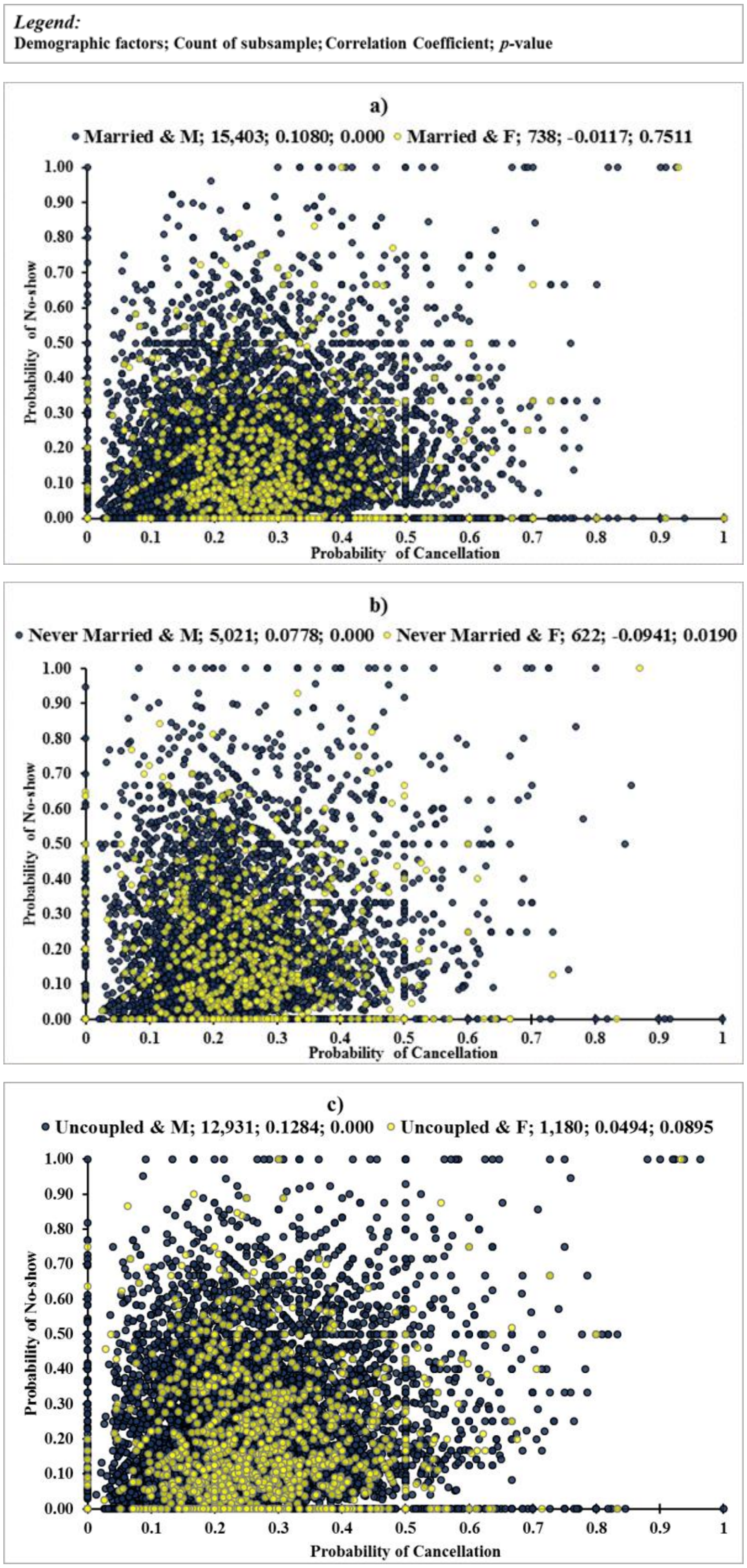
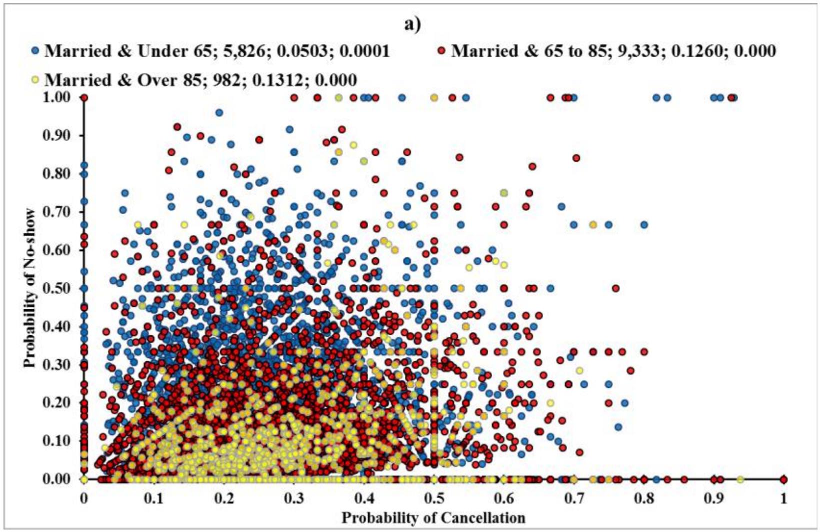
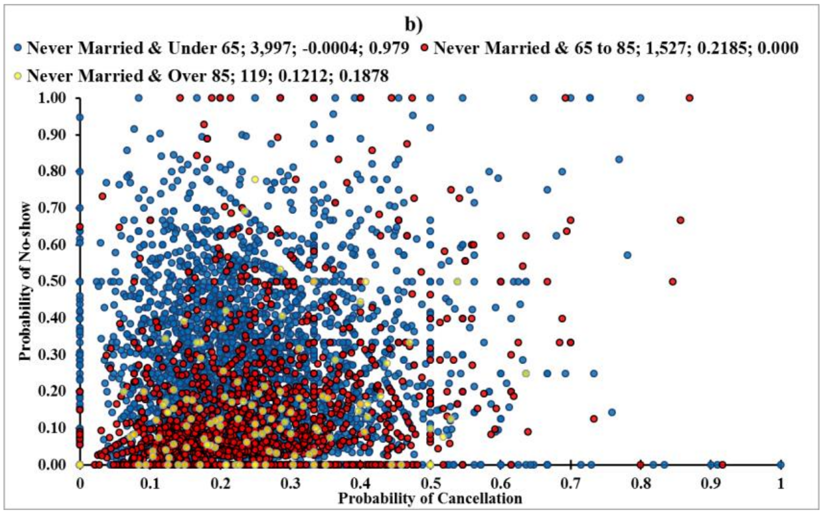


Figure 3.9. Scatterplot of Probability of No-show vs. Probability of Cancellation for Patients **a)** Married Gender, **b)** Never Married and Gender, and **c)** Uncoupled and Gender

Legend:
Demographic factors; Count of subsample; Correlation Coefficient; *p*-value





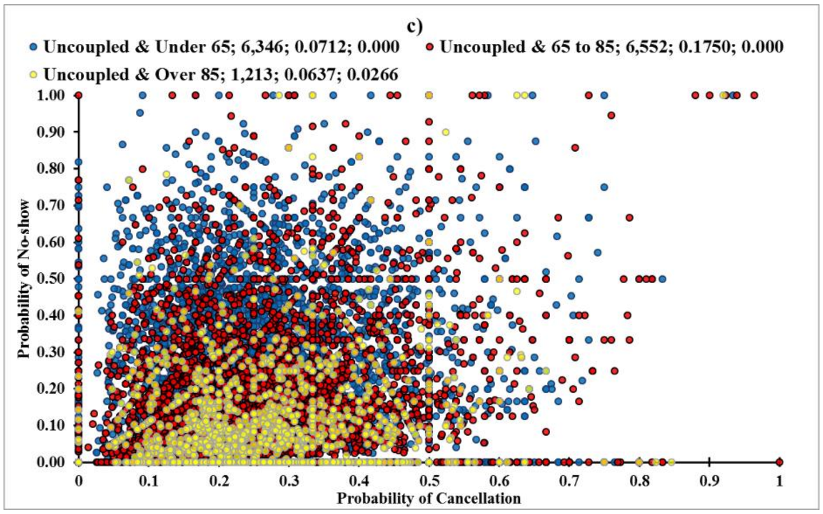


Figure 3.10. Scatterplot of Probability of No-show vs. Probability of Cancellation for Patients **a)** Married and Age Group **b)** Never Married and Age Group, and **c)** Uncoupled and Age Group

Table 3.3 lists the Spearman rank correlations for probabilities grouped in intervals of width 0.25. The intervals are left-closed and right-open, with the final interval also being right-closed. The number of patients in each group is listed in parenthesis next to the correlation coefficient. The greatest significant correlation is 0.5687, when a patient's probability of no-show and cancellation is in the interval [0.75, 1]. Thus, for the nineteen patients in this group a greater no-show probability is also associated with a greater cancellation probability. The group with probability of cancellation interval [0.75, 1] and probability of no-show interval [0, 0.25) has a correlation of -0.3372. The 49 people in this group have an inversely related no-show and cancellation probability. The rest of the significant correlations are less than 0.2 in absolute value.

Table 3.3. Spearman Rank Correlations for Discrete Probability of Cancellation and No-show Groups

		Probability of Cancellation			
		0 to 0.25	**0.25 to 0.5**	**0.5 to 0.75**	**0.75 to 1**
Probability of No-show	**0 to 0.25**	0.0936 (14633)**	0.0192 (11315)*	0.0669 (1098)*	-0.3372 (49)*
	0.25 to 0.5	-0.0235 (3035)	-0.0129 (3043)	0.1861 (399)**	0.2513 (11)
	0.5 to 0.75	-0.0745 (861)*	-0.0728 (873)*	-0.1099 (169)	-0.2092 (14)
	0.75 to 1	0.0313 (133)	0.0579 (169)	-0.0022 (74)	0.5687 (19)*

*p<0.5, **p<0.1

The results of our analysis to test H1 show that the correlation between no-show probability and cancellation varies across demographic groups. The greatest correlation among the demographic groups is 0.2185, as shown in Figure 3.10b, with one negative statistically significant coefficient at a value of -0.0941, as shown in Figure 3.9b. When the data are analyzed in discrete groups, the majority of groups do not have significant correlations, and those that do have varying relationships. When the data are grouped, they have a correlation of 0.1117, so that a linear regression of cancellation rate on no-show rate (or vice versa) would yield an R-squared value of 1.24%. Therefore, we determine that the degree of linear relationship between the two variables is weak, and does not justify combining the probabilities, or eliminating cancellations. Thus, we conclude that these results provide support to H1.

3.4.1.2 Testing of H2

H2 theorizes that the no-show and cancellation probabilities do not have statistically equal medians. To test this hypothesis, we performed a Wilcoxon signed-rank test, where the null hypothesis states that the difference between the medians of the two sample distributions equals zero, or $H_0 : \tilde{x}_{differences} = 0$. This test was performed because the distributions are not independent, as each is associated with a single patient, and because the sample distributions were found to be not normal. Table 3.4 lists the medians for each sample and the results of the signed-rank test.

Table 3.4. Results of the Wilcoxon signed-rank Test of Sample Medians

Name	Median	Wilcoxon signed-rank test Statistic	*p*-value
Probability of Cancellation	0.2391	88.571	0
Probability of No-show	0.1042		

Given that the *p*-value of the signed-rank test is 0, we can reject H_0, and conclude that the differences of the two medians is not zero. This result supports H2, and provides evidence that cancellations and no-shows should be considered separately.

3.4.1.3 Testing of H3

H3 states that the patient demographics, age group, gender, and marital status, have different effects on the two probabilities. We test this hypothesis by performing a multi-factor ANOVA. The dependent variable is a sample that contains both no-show and cancellation probabilities for each patient. A factor labeled as *Type* is included as a differentiator of the types of probabilities in the sample. Age group, gender, and marital status for each patient were also used as factors. To validate our hypothesis, we analyze if the interaction effect of each of the demographic variables with Type is significant. Main effects for each variable were not tested, as these tests give no insight into the differing influences of the factors on type of probability.

Table 3.5 lists the ANOVA table for this test. Main effects are not shown, but all but Gender are significant at the 0.05 level. Each of the patient demographics and its interaction with Type is significant in the model. Therefore, we can conclude that H3 is supported and age group, gender, and marital status have different effects on no-show and cancellation probabilities. Figure 3.11 through Figure 3.13 depict each of the interactions with 95% confidence intervals.

Table 3.5. Results of ANOVA

Source	Sum of Squares	df	Mean Square	F-Ratio	*p*-Value
Interactions					
Gender & Type	36.153	2	18.0765	836.81	0
Age Group & Type	1.79126	1	1.79126	82.92	0
Marital Status & Type	21.7509	2	10.8754	503.46	0
Residual	1550.52	71778	0.0216016		
Total	1843.32	71789			

In our sample, patients in each age group are more likely to cancel than they are to no-show. We achieved similar results as in prior literature (Bean & Talaga (1995), Daggy et al. (2010), Norris et al. (2014)), that find that younger patients are more likely to no-show for their appointments than older patients. For cancellations, the oldest age group, Over 85, are the most likely to cancel. Cancelling an appointment allows the clinic time to potentially reschedule the appointment slot, and patients Over 85 are the most likely to give the clinic this courtesy. The younger age groups' tendency to cancel is not statistically different from each other, but they are both more likely to cancel than to no-show.

The majority of our sample, 92.92%, are Males. This is representative of the VHA patient base. As in prior literature (Galluci et al. 2005), we find that Males are more likely to no-show than Females. The probability of cancellation is greater for both genders, but the tendencies are reversed, with Males having a statistically significantly lesser cancellation probability as opposed to Females.

Patients who are Married are less likely to no-show than patients who are Uncoupled or Never Married. These results follow with the results in Daggy et al. (2010). This could be due to the support structure in the home, encouraging the patient to attend an appointment. Again, the probability of cancellation for all groups is greater than the probability of no-show. Married and Uncoupled patients have the highest probability of cancellation on average, although they are not significantly different from each other. Both are statistically more likely to cancel than patients who were Never Married.

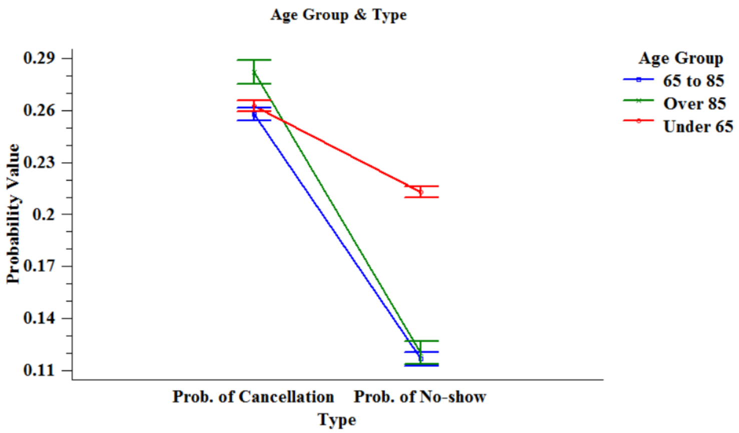


Figure 3.11. Interaction Effects of Age Group on No-show and Cancellation Probabilities

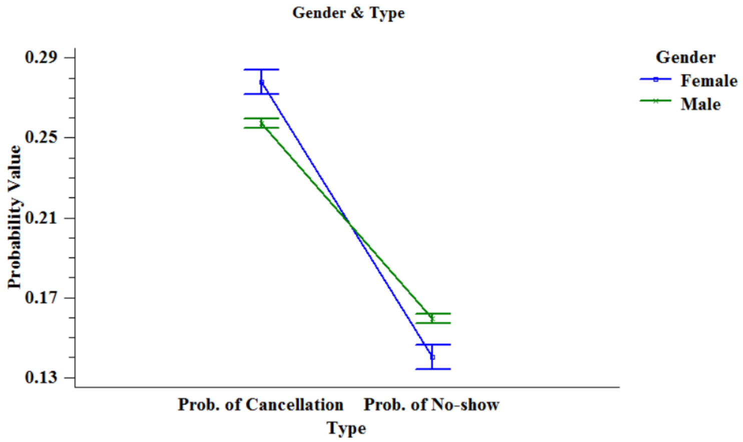


Figure 3.12. Interaction Effects of Gender on No-show and Cancellation Probabilities

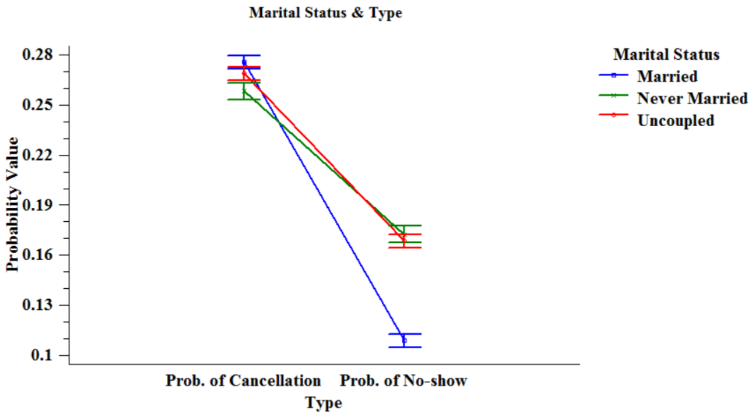


Figure 3.13. Interaction Effects of Marital Status on No-show and Cancellation Probabilities

The results of the ANOVA indicate that a patient’s tendency to no-show or cancel are, at times, reversed across demographic groups. Thus, grouping these probabilities, or excluding cancellations, does not allow an analyst to get as rich an insight from his predictive analysis or scheduling application. Table 3.6 lists each hypothesis with the test results.

Table 3.6. Result of Hypotheses Analyses

Hypothesis	Result
H1. No-show probability does not have a positive correlation with cancellation probability.	Supported
H2. The no-show probability and cancellation probability distributions do not have statistically equal sample medians.	Supported
H3. Different demographic groups have different no-show and cancellation rates.	Supported

3.4.2 Post hoc Analyses: the Made to Cancel Ratio (MTCR)

As we saw with patients over 85, there are patients who no-show, but have a higher tendency to cancel appointments. This could be because cancellations act as a substitute for no-shows – a patient realizes the clinic can reuse the appointment, so instead of not showing up, she notifies the clinic beforehand. Given this assumption, we expect that patients who cancel near the end of their lead time will have fewer no-shows. Alternatively, patients who cancel too late could also be the patients who no-show, because these patients have a tendency to blow-off their appointments by cancelling or no-showing. If either of these assumptions is true, then the made to cancel ratio (MTCR) for each patient should be correlated – negatively for those who substitute cancellations and positive for those who cancel and no-show – with their probability of no-show.

Figure 3.14 is a scatterplot of the probability of no-show and average MTCR for 35,338 of the patients. Patients not represented in the plot did not cancel an appointment in the sample, and therefore do not have a made to cancel ratio. The majority of patients fall below the reference line in Figure 3.14. Thus, patients who have a greater MTCR typically have a no-show rate that is lower than their MTCR. This provides support to the assumption that later cancellations act as substitute for no-shows, but a patient who has a lower MTCR is not more likely to show. If patients who cancel are the patients who also no-show, points would be clustered around the grey line. This assumption does not seem to be supported by the scatterplot, as the majority of points are clustered underneath the line, not around it.

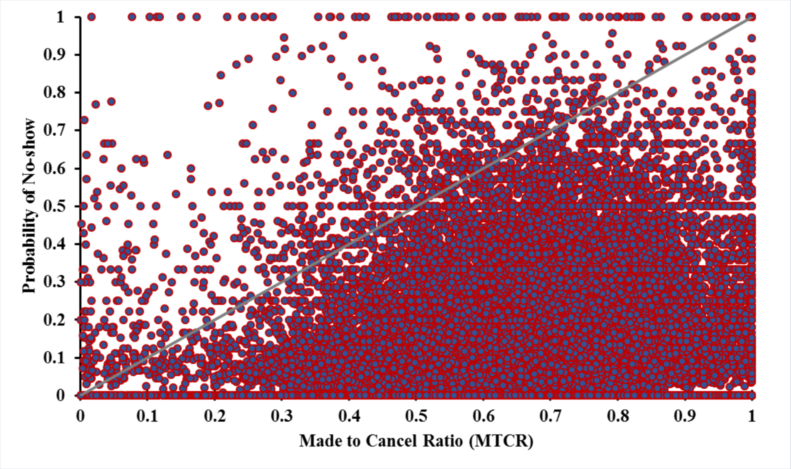


Figure 3.14. Scatterplot of Probability of No-show vs. Made to Cancel Ratio (MTCR) with a Perfect Positive Correlation Reference Line

The Spearman correlation coefficient for no-show probability and MTCR was calculated to be 0.1585 with a *p*-value=0. This correlation is stronger than the no-show probability correlation with cancellation probability, but still does not justify eliminating cancellations from patient attendance analysis. Given the results of the hypotheses test, and the post hoc analyses, in the next section we develop a model to predict a patient's MTCR. This model can be used in conjunction with a no-show predictive model, to inform an analyst about both types of patient attendance behavior. The time to cancel is predicted, not just the probability of cancellation, because the timing of a cancellation is important to a clinic who will seek to rebook appointments that are cancelled.

3.5 MTCR PREDICTIVE MODEL

Our goal is to predict the occurrence of advance appointment cancellations, and when such cancellations occur, as a proportion on the time interval between when the appointment is made and when it is to occur. The model is motivated by a study of outpatient clinics in the Veterans Affairs (VA) Healthcare System, but can be applied to other situations that involve a binary outcome plus a continuous value for one of the binary outcomes, such as airline or hotel cancellations (Zadrozny & Elkan 2001), and charitable solicitations (Ling & Li 1998, Zadrozny & Elkan 2001). Advance cancellations may affect scheduling decisions (Liu et al. 2010), but also can contribute to the estimation of net demand in revenue management settings (Morales & Wang 2010). Given the impact cancellations may have on managerial decision-making, a reliable and easy-to-use model may be an important factor in operational performance.

Sample selection bias arises as an issue when predicting when a patient will cancel, because only the patients who cancelled a have a dependent variable to predict. Zadrozny and Elkan (2001) proposed a two-step Heckman procedure (Heckman 1979), in which the binary class variable is first modeled using a probit linear model, and that value is transformed and added as an input to the linear regression used to predict the continuous variable. Heckman found that such a procedure yields unbiased estimates. Zadrozny and Elkan (2001) compare the two-step procedure with a one-step cost-sensitive decision-making model, and found that the two-step model is able to obtain higher profit estimates.

Our model uses a regression-like approach to predict both the class variable and the continuous variable using a single ordinary least squares (OLS) model. We contribute to the two-class prediction literature by adding a numerical modeling parameter that is assigned to all records in the failure class. This parameter is the enhancement that allows us to overcome sample selection bias, and predict both variables simultaneously. We found that a single regression model performs similarly to established data mining models, such as logistic regression (LR) and C5 (Quinlan 2004) combined with an OLS model, while also providing a one-step method to gaining a consistent measure of cancellation prediction.

3.5.1 Methodology

The data used to train and test the predictive model is a subset of the data sample described in Section 3.3.1. To predict cancellations, we collected all appointments in the Psychiatry medical specialty, and eliminated all patients that did not have at least ten Psychiatry appointments. We focused our model on predicting if and when each patient, with at least ten Psychiatry appointments, cancelled the last appointment in the sample. The model was induced from a training set of 5,041 records and validated using a test set of 2,185 records.

To prepare the data, we calculated the appointment lead time and the made to cancel ratio (MTCR) for cancelled appointments. We used the ratio of lead time, so that the effect of the time-to-cancel may be the same across varying lead times, and because lead time has been found to be significantly related to missed appointments (Chariatte et al. 2008, Norris et al. 2014). The MTCR values range between 0 and 1. We do not permit MTCR values equal to 0 or 1, because this would indicate a patient who cancels at the same time she makes the appointment, or at the same time the appointment is to occur. Figure 3.15a and Figure 3.15b display histograms of the calculated MTCR values for the training and test set, respectively.

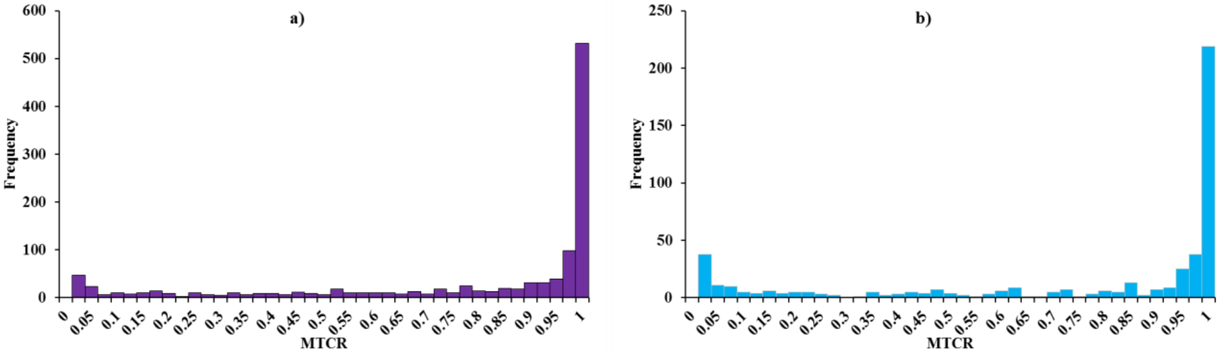
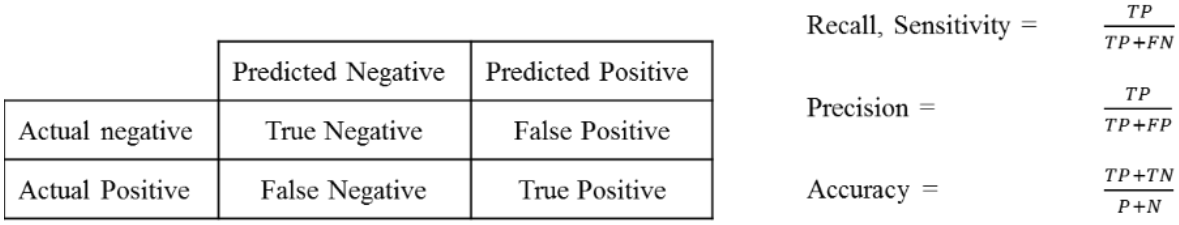


Figure 3.15. a) Histogram of the calculated MTCR values for the training dataset and **b)** Histogram of the calculated MTCR values for the test dataset

To control for selection bias, and to permit a single model to predict both cancellation and time to cancel, we assign a constant, α , as the MTCR value to all patients who did not cancel. The value of α is greater than or equal to 1, and the preferred value is a function of the data set. We assign values greater than or equal to 1 to indicate a “phantom” cancel, where a cancellation occurs after the appointment has occurred, and, therefore was not an advance cancellation. For this application, we vary α between 1 and 1.1. To determine the preferred α model, we choose the model with the greatest F_1 score, $F_1 = 2 \times \frac{precision \times recall}{precision + recall}$, which trades off the values of Recall and Precision. Example calculations for MTCR are in Table 3.7; Recall, Precision, and Accuracy formulas can be seen in Figure 3.16.



	Predicted Negative	Predicted Positive
Actual negative	True Negative	False Positive
Actual Positive	False Negative	True Positive

$$\text{Recall, Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Accuracy} = \frac{TP+TN}{P+N}$$

Figure 3.16. Confusion Matrix and Recall, Precision, Accuracy Equations

Our model uses the calculated MTCR and assigned α values as the dependent variable. To induce predicted MTCR values, we ran an OLS model to predict the dependent variable. We then chose a breakpoint of 1, and assigned all appointments with a predicted MTCR less than 1 as cancelled, and a MTCR greater than or equal to 1 as not cancelled. The predicted MTCR value, for appointments in the cancel class, is used as the prediction for the continuous variable of when the patient will cancel during the appointment lead time. To determine the accuracy of the continuous variable, we calculate the Mean Absolute Error (MAE) of all predictions. The MAE calculations only include records where a cancellation occurred, and records where the model was able to correctly predict the appointment as being in the cancel class.

Table 3.7. Example MTCR Calculation and Assignment

Date Appt Made	Cancel Date	Appt Date	Cancel(0,1)	Lead Time	MTCR
4/1/2011		4/13/2011	0	12	1.07
6/22/2012	6/29/2012	7/3/2012	1	11	.6364
2/13/2016	3/8/2016	3/9/2016	1	24	.9583

We chose two two-phase modeling techniques to compare to the MTCR model performance. The Logistic Regression-OLS (LR-OLS) model is a model where the class assignments are made based upon the results of a Logistic Regression, where the cut-off value is generated from relative misclassification costs. The C5-OLS model class predictions are made from a C5 decision tree induced using relative misclassification costs. The preferred model is chosen as the model with the greatest F_1 score. For the second phase, both models predict the continuous variable, the MTCR, for patients who cancelled. As in the Heckman procedure, we applied the standard logistic transformation on the observed MTCR value, mapping the interval (0,1) into (-∞,∞), and used OLS regression to generate predictions for the MTCR. Both two-phase models use the same second phase OLS approach. Performance of the MTCR predictions are measured using an MAE calculation as in the MTCR model.

3.5.2 Analyses

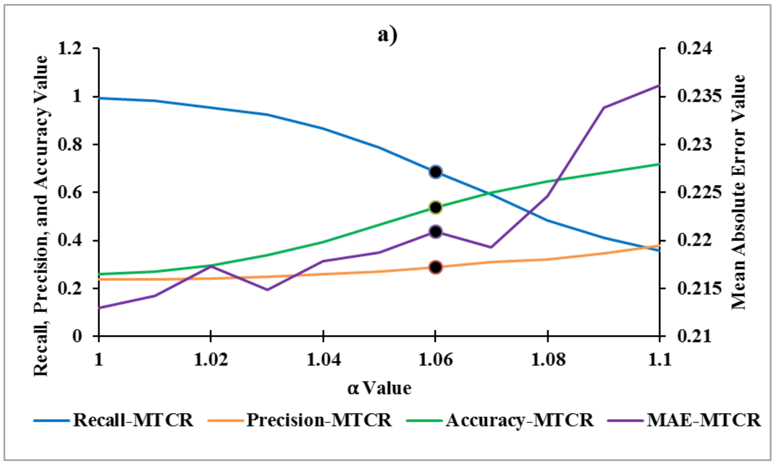
The MTCR, LR, and OLS portions of the models were estimated using IBM SPSS Statistics 21, and the C5 model was induced using IBM Modeler 15. Appointment characteristics, patient age group, gender, marital status, and appointment behavior were used as the independent variables in the models. Table 3.8 lists descriptive statistics for the variables used in the analysis from the training dataset.

Table 3.8. Operationalization of Variables from the Training Dataset

Name	Description	Operationalization	Min	Max	Std. Dev.
Probability of Cancellation	Fraction of a patient's total no-showed, cancelled, and completed appointments that were cancelled before the appointment was to occur, excluding the last appointment	Continuous	0	1	0.134
Probability of No-show	Fraction of a patient's total no-showed and completed appointments that were not attended, excluding the last appointment	Continuous	0	1	0.179
LN of Lead Time	Natural log of lead time	Continuous	0	5.897	1.047
Total Cancelled	Count of the number of cancelled appointments in the sample for each patient	Continuous (capped at 10)	0	10	3.17
Cancel Last	Flag to indicate if the patient cancelled the last appointment	Binary: 0=non cancelled (75.94%), 1=cancelled (24.06%)			
Month	Month of the date of the scheduled appointment	Categorical (Jan (14.68%), Feb (1.86%), Mar (2.64%), Apr (3.11%), May (3.27%), Jun (3.35%), Jul			

Name	Description	Operationalization	Min	Max	Std. Dev.
		(4.60%), Aug (4.44%), Sep (6.41%), Oct (9.30%), Nov (14.24%), Dec (32.08%))			
Age	Age of the patient at the time of the last appointment in the data sample	Categorical (Under 65 (82.52%), 65-85 (16.58%), Over 85 (.9%))	20	94	14.948
Gender	Gender of the patient in the data sample	Categorical (Male (89.41%), Female (10.59%)			
Marital Status	Marital Status of the patient in the data sample	Categorical (Married (37.41%), Never Married (24.09%), Uncoupled (38.50%))			
Made to Cancel Ratio (MTCR)	Fraction of the appointment lead time that passes before a patient cancels the appointment; calculated for last appointment in the sample	Continuous	0.0008	0.9999	0.316

The three comparison modeling techniques are the MTCR model (MTCRM), LR-OLS, and C5-OLS. Several models were induced for each technique: eleven MTCR models with α ranging from 1 to 1.1, twenty-one LR-OLS models with misclassification costs ranging from 1 to 5, and fifteen C5-OLS models with misclassification costs ranging from 1 to 6. Figure 3.17a through Figure 3.17c display the Precision, Recall, Accuracy and MAE curves for all models run on the training dataset for MTCRM, C5-OLS, and LR-OLS, respectively. The preferred α model for the MTCRM and the preferred misclassification cost model for C5-OLS and LR-OLS are marked with a black marker on each curve.



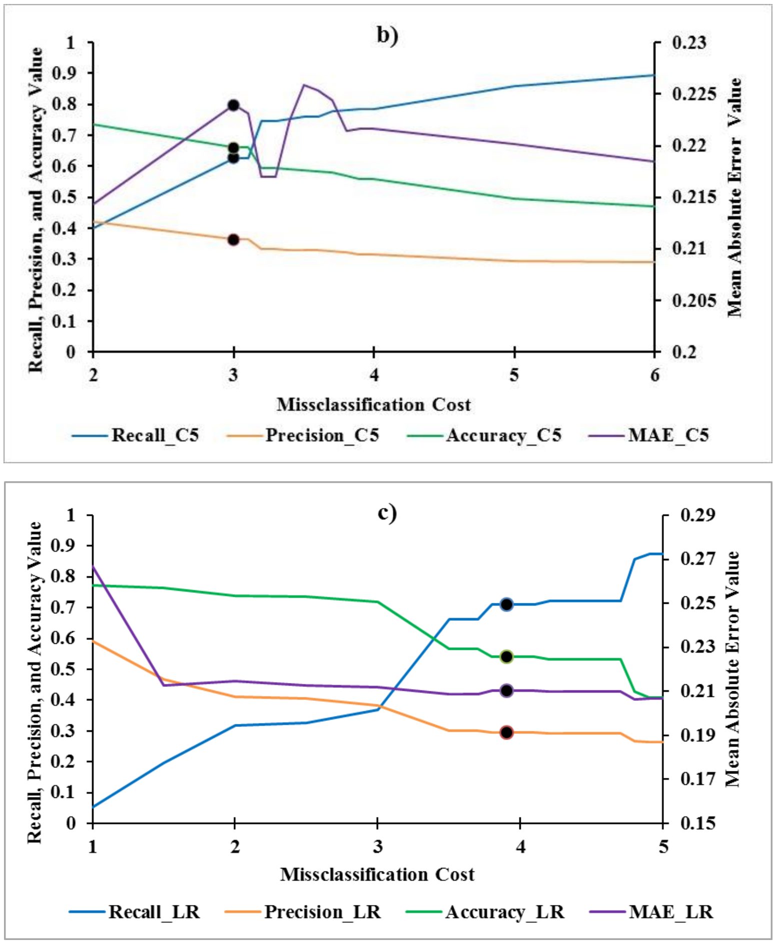


Figure 3.17. Precision, Recall, Accuracy, and MAE curves for **a)** MTCR models, **b)** C5-OLS models, and **c)** LR-OLS models on the Training Dataset

For the LR-OLS and C5-OLS models, as the relative cost of misclassifying a cancellation as a non-cancellation increases, Recall increases and Precision and Accuracy decrease. This is because more false positives are being predicted. The MTCR model has an opposite pattern. As the value for the parameter α increases, Recall decreases and Precision and Accuracy increase. This occurs because all estimated values increase, resulting in more predictions being above 1, and being assigned to the non-cancel class. As 77.17% of our dataset consists of appointments that are not cancelled, there is an upward trend in Accuracy. The MAE curves between MTCRM and the two-phase models also has a reverse pattern. (The C5-OLS models were run at misclassification cost intervals of 1. Between 3 and 4 the interval was decreased to 0.1, which accounts for the pattern seen in Figure 3.17b). Table 3.9 lists the values of each metric from the training dataset along with the calculated F_1 score. The preferred α and misclassification costs were used to generate metrics from the test dataset. The test dataset values are also listed in Table 3.9.

Table 3.9. Metrics for Preferred Models

		α / Misclass. cost	**Cutoff**	**Recall**	**Precision**	**Accuracy**	**MAE**	***F*₁ score**
MTCRM	Training	1.06		0.686	0.289	0.539	0.221	0.407
	Test	1.06		0.667	0.276	0.536	0.281	0.391
LR-OLS	Training	3.9	0.199	0.709	0.295	0.542	0.210	0.417
	Test	3.9	0.199	0.485	0.278	0.605	0.278	0.394
C5-OLS	Training	3		0.628	0.364	0.661	0.224	0.461
	Test	3		0.669	0.280	0.542	0.260	0.353

The C5-OLS model with a misclassification cost of 3 has the greatest F_1 score on the training dataset, followed by LR-OLS with a misclassification cost of 3.9, and MTCRM with α set to 1.06. Thus, on the training dataset, the C5-OLS model is the best model, of the three, to predict the class variable. This ordering does not transfer to the test dataset, the dataset used to test the utility of the tuned model. On the test dataset, LR-OLS is the preferred model, followed by MTCRM and C5-OLS. MTCRM has the minimum change in F_1 score between the training and test set. All models have an increase in MAE between the training and test set. C5-OLS has the minimum MAE on the test set, followed by LR-OLS and MTCRM.

Table 3.10 presents the parameter results of the models. The generated Beta value from MTCRM, LR, and OLS are listed with the standard error in parentheses. The top ten variables with their predictor importance in the decision tree model are listed for the C5 model. For MTCRM, the variables listed were used to predict the continuous variable and the class variable, with a cutoff of 1. A positive coefficient for the MTCRM model decreases the probability that a patient will cancel, given that all predictions less than 1 are assumed to be in the cancel class, and the predicted MTCR value is used as the predicted time to cancel. For LR and C5 the variables listed were used to predict the class variable. The OLS model is the continuous variable (when a patient will cancel) prediction for the LR and C5 models. The calculated MTCR values were transformed from the interval (0,1) to the interval (-∞,∞) before the model was run. The values were transformed back to the interval (0,1) before MAE was calculated.

Table 3.10. Model Results

Variable	MTCRM	LR	C5	OLS
LN Lead Time	-0.02 (0.003)	0.302 (0.037)	0.2	-0.951 (0.133)
Probability of Cancellation	-0.092 (0.02)	2.134 (0.259)	0.14	
Cancel Last		0.169 (0.083)*	0.03	
Married			0.06	
Male			0.04	
Jan	0.039 (0.009)	-0.421 (0.108)		1.195 (0.341)
Mar			0.05	
Apr			0.05	
Jun				-1.222 (0.591)*
Jul			0.05	
Aug			0.07	
Sep		-0.367 (0.149)*		
Oct	0.035 (0.01)	-0.862 (0.136)		
Nov	0.05 (0.009)	-0.894 (0.119)		
Dec	0.036 (0.007)	-0.786 (0.09)	0.03	
Constant	1.053 (0.01)	-2.23 (0.148)		6.858 (.464)

*$p<0.5$. all others $p<0$

The natural log of lead time is significant in all models. As the lead time increases, patients are more likely to be in the cancel class (prediction below 1) for MTCRM. For both the MTCRM and OLS model, a longer lead time decreases the predicted MTCR. Thus, patients with longer lead times will cancel closer to the date they called to make an appointment. The constant of the MTCRM model is above 1 (1.053), thus, when all variables are equal 0, a patient will be predicted to be in the non-cancel class. The two factors that could reduce this prediction to below 1 are the natural log of lead time and probability of cancellation.

If appointments occur in the months of January, October, November, and December, a patient is less likely to be in the cancel class for both the MTCRM and LR models. The same trend applies when a patient cancelled the last appointment for the LR model. Thus, cancellations are typically followed by a non-cancelled appointment, for the LR model. As the probability of cancellation increases, a patient is more likely to be in the cancel class for LR and MTCRM.

For the OLS model, appointments in the month of June decrease a patient's predicted MTCR, making it more likely that she will cancel close to the date she made the appointment. January has the reverse effect. The most important predictor in the C5 model is the natural log of lead time followed by the probability of cancellation of past appointments. The C5 model is the only model where Married and Male were included as significant variables. Although probability of no-show was included in all models as an independent variable, it is not significant in any of the models.

3.6 DISCUSSION AND CONCLUSIONS

In this chapter we perform a descriptive analysis of cancellation and no-show probabilities, in an effort to determine if cancellations can be grouped with no-shows, or not considered in patient attendance analysis. We also propose a novel model for predicting the fraction of a patient's lead time that will pass before he or she cancels, or made to cancel ratio. Predicting an appointment's MTCR, as opposed to the probability it will be cancelled, assists a clinic in identifying not only if a patient will cancel, but *when*. When a patient will cancel becomes important when differentiating between advance cancellations that can be rescheduled with a high probability, and late cancellations which cannot.

Our predictive analysis indicated that a patient's cancellation probability is statistically different than the no-show probability. Additionally, patient demographics such as, age, gender, and marital status, have different effects on each probability. Thus, we conclude that cancellations should be considered as an independent type of patient attendance behavior in any model that uses patient attendance behavior as an input.

A patient's MTCR was also found to be different than no-show probability. Our model to predict MTCR is able to perform similarly to a two-phase LR-OLS and two-phase C5-OLS model, when analyzing the MAE and F_1 score of the predictions, while also providing a consistent measure for predictions. The MTCRM model has the minimum change in MAE and

F_1 score between the training and test datasets among all modeling techniques. Lead time and prior cancellation history were significant variables in all models. As further support of the difference between no-shows and cancellations, prior no-show history was not included as a significant predictor in any of the models.

Extensions of this work include performing more descriptive analyses on the relationship between patient demographics and no-show and cancellation probabilities, and continuing to adjust the model to improve predictions. Given the “bathtub” shape of the distribution of the calculated MTCR, and that the majority of appointments are cancelled close to the appointment date, it is difficult to obtain accurate estimates for patients with a low MTCR. Additionally, a consistent measure of performance for the class designation could be developed. A typical measure used to analyze a class assignment is a Receiver Operator Characteristic (ROC) curve. For our application, an ROC curve is not preferred because we have highly unbalanced data (Davis & Goadrich 2006), and because the MTCRM does not require a change in cutoff value to determine performance; it requires a change in α value. Davies & Goadrich (2006) propose a Precision Recall (PR) curve as an alternative to the ROC curve, but, for our application, the range of Precision and Recall values is model dependent, and thus, it is unclear how to compare the values across models.

A single measure to determine a preferred model based upon the accuracy of the class and continuous variable predictions, would improve model selection. Currently the F_1 score allows for a preferred model to be chosen based upon the Precision and Recall of the class assignment, and the MAE score allows for a preferred model to be chosen based upon predictive accuracy of the continuous variable. An analyst can determine which variable is more relevant in their context, and choose a preferred model, but a single measure that encompasses both predictions would be a valuable addition.

4.0 ONLINE OVERBOOKING MODEL

We develop strategies that a clinic can utilize to determine if and when to overbook patients, over a finite horizon, in an online scheduling environment. We incorporate clinic parameters, including indirect waiting, no-shows, and cancellations to inform the overbooking decisions. We find that the optimal overbooking strategies are a function of both no-shows and cancellations, and that a clinic can, under certain conditions, achieve a greater service reward by overbooking patients than it can by not utilizing overbooking. Our work is motivated, in part, by our observations of scheduling decision-making at a Veterans Health Administration (VHA) specialty clinic.

4.1 BACKGROUND

Timely patient access to healthcare systems is an on-going problem that is yet to be resolved (IOM 2015). Lengthy patient scheduling queues, and wait times at a clinic, may reduce patient satisfaction, and, perhaps, lead to "poorer health outcomes" (IOM 2015, p. 11). Patient behavior, such as no-shows and cancellations, can lead to schedule inefficiencies, such as underutilization of clinic resources or overtime. Cancellations may be grouped into two categories: advance and late cancellations. The two types differ in their effects on the clinic schedule. Advance cancellations are appointments that are cancelled far enough in advance that the clinic may assume, with a high probability, that the appointment slot freed up by the cancellation may be reassigned to another patient. Late cancellations have a lesser probability of being reassigned, and are, at times, grouped with no-shows, because patients who cancel late may free up a time slot that cannot be rescheduled by the clinic (Gupta and Denton 2008).

No-shows and cancellations need to be considered in clinics where the demand for service exceeds the number of available appointment slots. Examples of strategies used to mitigate the negative effect of appointment no-shows and cancellations include overbooking and the use of overtime slots. Overbooking can be implemented in a naïve or informed manner. Kim and Giachetti (2006) define naïve overbooking as the practice of providers to "overbook based on either intuition of what they think the no-show rate is or…on just the average no-show rate." We define informed overbooking as the practice of providers to overbook based on the results of a prescribed analytical model that uses clinic parameters and patient behavior as inputs to direct decision-making. Overbooking should be applied in an informed manner to prevent additional issues, such as excessive provider overtime and in-clinic wait times. In the remainder of this chapter, unless specified otherwise, overbooking always refers to informed overbooking. In this chapter we show that both types of patient behavior, i.e., no-shows and cancellations, must be considered when prescribing overbooking strategies.

We develop an overbooking model that incorporates no-shows and cancellations. We limit the model's decision space to determine if and when a patient should be overbooked. The model is restricted to making overbooking decisions, because all other decisions are exogenous to the model. We assume the number of appointment slots is fixed, and that the length of each appointment is constant. In addition, demand for appointments exceeds appointment supply, and all available slots are already filled with patients, based on their preferences. Overbooking decisions are made over a multi-day horizon of predetermined length. This modeling structure is directly motivated by our observations of scheduling in a specialty health clinic at a Veterans Health Administration (VHA) hospital.

Overbooking, in the clinics we studied, may be necessary due to scheduling time constraints. In the VHA, patients are referred to specialty clinics by their primary care physicians, and need to be seen within a specific time window. Patients may need a single visit to the specialty clinic, or multiple visits scheduled periodically. For example, if a patient must be seen every two weeks for an oncology appointment, the subsequent appointments must be scheduled at that fixed interval. If no appointment slots are available within this interval, overbooking may become necessary.

In the clinic we observed, clinic schedulers do not differentiate among patients based upon their unique probabilities of no-show and cancellation, so, in our model, we assume homogeneous no-show and cancellation probabilities. Patients request appointments for a specific day, and we assume that patients prefer to be seen as soon as possible. That assumption is based on the finding that, in Mental Health, an example of a specialty clinic, patients respond best to care when they first realize there is a problem (Kenter et al. 2013). Additionally, the assumption corresponds with an Open Access (OA) policy, where patients are asked to come immediately, or to call on the day on which they need an appointment (Liu, Ziya & Kulkarni 2010). Scheduling is done on an online basis. That is, patients must be offered an appointment slot when they contact the clinic, and, once scheduled, cannot be moved by the clinic to a different day or time slot.

Additional assumptions are as follows. All requested appointments are assigned to a particular day and time slot in the scheduling horizon. If patients show, they show punctually. We assume the clinic assigns a cost to both direct waiting (the time a patient waits for service in the clinic) and to indirect waiting (the time between when a patient requests an appointment and the appointment day). In addition, the no-show and cancellation probabilities are increasing functions of the indirect wait time. The objective of the overbooking model is to obtain the maximum clinic service reward. Thus, a clinic service reward is a function of the number of patients who complete their appointments, direct and indirect waiting time, and overtime. This corresponds to a clinic that seeks to maximize the number of patients it sees within a scheduling horizon, while seeking to limit patient waiting times and clinic overtime.

The primary contributions of this chapter are as follows. We show that the optimal overbooking strategy is a function of both the no-show and the cancellation probabilities. These probabilities affect both the day on which an overbooking may occur, and the appointment slot in which the patient is overbooked. The overbooking strategy is a nonlinear function of these two probabilities, as well as the other parameters that describe the clinic operations. We provide a model that yields generalized rules for overbooking over a multi-day horizon, in the presence of no-shows and cancellations. We limit the discussion to overbooking up to two patients per day, because that is the typical number of patients that are overbooked in the clinic we observed. We consider both direct and indirect waiting times, which Gupta and Denton (2008) identify as a gap in the current literature. Finally, we show how our proposed overbooking strategies can be used to motivate managerial decision making in a clinic.

The rest of the chapter is organized as follows. Section 4.2 includes a review of the related literature. In Section 4.3, we present our model. Section 4.4 outlines the model solution technique, Section 4.5 describes the model properties, Section 4.6 provides empirical results, and Section 4.7 summarizes our findings and conclusions.

4.2 LITERATURE

Addressing scheduling models with overbooking has been studied for some time, beginning with Bailey (1952). Cayirli and Veral (2003) presented a review of developments since then; Gupta and Denton (2008) reviewed general methodologies and outlined possible open challenges in healthcare scheduling. The literature most relevant to our work are models that address no-shows, cancellations, and/or multi-day horizons, with the goal of proposing overbooking strategies. Kim and Giachetti (2006), LaGanga and Lawrence (2007), LaGanga and Lawrence (2012), Huang and Zuniga (2012), and Zacharias and Pinedo (2014) focused on overbooking strategies to mitigate no-shows in single server and single day models. These papers did not address the effect of cancellations or indirect waiting. Kim and Giachetti (2006) formulated a stochastic mathematical overbooking model to assign an optimal overbooking level within a day. Their model accounts for direct waiting and clinic overtime, but does not address the slot assignments for the overbooked patients. They found that clinic profit can be increased if overbooking occurs when no-show rates are high or variable. Huang and Zuniga (2012) also sought to find an optimal overbooking level, but accounted for slot assignments by solving for a no-show probability that accommodates overbooking in a single slot.

LaGanga and Lawrence (2007) developed a simulation model that overbooks patients using slot compression, where the number of slots is increased by setting the time between scheduled appointments, at a value less than the service time, as opposed to scheduling several patients into a single slot. They considered homogeneous no-show rates, and balanced clinic overtime, patient waiting time, and the clinic benefit for seeing a patient. They found that overbooking provides utility when no-show rates and the number of patients requesting appointments are high, and service variability is low. LaGanga and Lawrence (2012) extended that work to include convex waiting and overtime functions, and no-show probabilities that vary by time of day. The results of both papers found that overbooking levels are a function of clinic size and cost parameters, and therefore, generalized rules should not be stated without considering these variables. Zacharias and Pinedo (2014) developed a static model – where all patient requests are assumed to be known a priori – that assumes heterogeneous patients, and that uses the results of their static model as a basis for a dynamic model. They found that no-show rates and patient heterogeneity impact overbooking decisions.

Muthuraman and Lawley (2008) and Zeng et al. (2010) both assumed heterogeneous no-show probabilities in sequential scheduling models – where patient requests for appointment are not known a priori – that incorporate direct waiting. Muthuraman and Lawley (2008) developed a myopic sequential scheduling model and algorithm, where future call-ins are not taken into account, and patients are overbooked until adding another patient to the schedule decreases the objective function. They found that the order of patient appointment requests and the clinic cost parameters affects the overbooking decision. Zeng et al. (2010) extended this work to include additional algorithms that allow for individualized patient no-show probabilities.

Patrick (2012) and Samorani and LaGanga (2015) developed multi-day scheduling models that account for indirect waiting and day-specific no-show probabilities, but do not consider cancellations. Patrick (2012) did not account for patient slot assignment within a day, only day assignments. He concluded that after a clinic day capacity reaches a “certain threshold” – which is a function of clinic service benefit, idle time, overtime, and lead time cost – it becomes optimal to begin deferring patients to a future day in the scheduling horizon. Samorani and LaGanga (2015) assigned patients to a day using slot compression, and concluded that accurate prediction of patient no-show probability is integral to the success of overbooking.

Liu et al. (2010) and Parizi and Ghate (2016) accounted for no-shows and cancellations in a multi-day scheduling model. Both papers developed a dynamic scheduling model that assumes time dependent no-show and cancellation probabilities, with the objective of choosing which day is optimal for a patient. Slot scheduling is not addressed in either paper. Additionally, while overbooking is discussed, specific overbooking strategies are not articulated. Liu et al. (2010) focused on discussing optimal scheduling heuristics, and Parizi and Ghate (2016) focused on the structure and performance of their specified Markov Decision Process (MDP).

We extend this literature by addressing patient no-shows and cancellations, clinic cost parameters, multi-day scheduling, and slot placement. The inclusion of these parameters in one model allows us to discuss how overbooking is affected by no-show rates and patient request levels, how clinic parameters affect overbooking, and where and how a patient should be booked in a scheduling horizon, with a single model. This allows us to build overbooking rules that incorporate more of a clinic's priorities, and thus increase the usability and applicability to an actual specialty clinic.

4.3 MODEL DESCRIPTION

We model a clinic that services patients on an appointment basis over a scheduling horizon of *h* days. There are *N* appointment slots each day the clinic is open. Each appointment lasts a fixed and constant duration. Fixed service times are assumed in order to create a base cost estimate for a model with patients who no-show or cancel (LaGanga and Lawrence 2007). Patients request an appointment for day *i*, $i = 1,...,h$ in the clinic scheduling horizon, and M_i denotes the number of patients who request an appointment for day *i*. All patients who request appointments must be scheduled. Patients may be scheduled on the day they request an appointment, or for any future day in the scheduling horizon. We seek to present overbooking solutions for clinics with access issues, so we assume that the total number of patients who must be scheduled across all days in the horizon is greater than the total number of slots available across the entire horizon, or $\sum_{i=1}^{h} M_i \geq (h \times \mathrm{N})$. The number of patient appointment requests for any day is at least as great as the number of unassigned slots available for scheduling on that day. Therefore, on at least one day, at least one patient must be overbooked or scheduled on a future day, in order to accommodate all patients. The clinic is permitted to use overtime, and the daily overtime is not bounded, so that all patients scheduled on a day are seen that day. Scheduled appointments are not removed from the schedule unless requested by the patient. Patients are scheduled to arrive at the beginning of their assigned time slot, and are assumed to arrive punctually, if at all.

We assume that the no-show and cancellation probabilities for a given day are equal for all patients with the same lead time. Patients are assumed to show for an appointment, given they have not cancelled, based upon the time between their appointment request and the day upon which their appointment is scheduled to occur. It has been shown that patients are more likely to show for appointments closer to their request date (Davies et al. 2016, Gallucci et al. 2005), so we assume that the probability of show decreases with an increase in the indirect waiting time. Let *p* denote the probability of show for a requested-day appointment. We assume that the probability of showing decreases by a factor of α, $0<\alpha<1$, for each day of indirect waiting. Thus, the probability of showing with *d-i* days of indirect waiting, is given by $p[i,d]=p\alpha^{d-i}$.

We also assume that patients do not cancel request-day appointments, although they may no-show; advance cancellations are defined only for patients whose indirect waiting time is at least one day. The probability of cancellation is assumed to increase with indirect waiting time, and hence, the probability of non-cancellation, or retention probability, is assumed to decrease. Let θ represent the retention probability for next-day appointments, and assume that the retention probability decreases by a factor of β, $0<\beta<1$, for each day of indirect waiting. Then, the retention probability for patients who incur *d-i* days of indirect waiting is given by $\theta[i,d]=\theta\beta^{d-i-1}$.

Appointments for which the patient retains and shows are referred to as *completed appointments*. The probability of a completed appointment is given by $p[i,d]\times\theta[i,d]=p\alpha^{d-i}\theta\beta^{d-i-1}$. Figure 4.1 displays the possible outcomes for a patient whose indirect waiting time equals zero, and Figure 4.2 displays the possible outcomes for each day a patient is in the system, when the indirect waiting time is greater than zero.

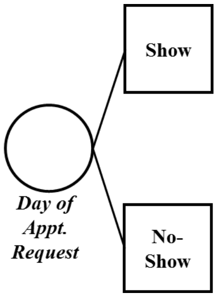


Figure 4.1. Possible Outcomes for Patients with No Indirect Waiting

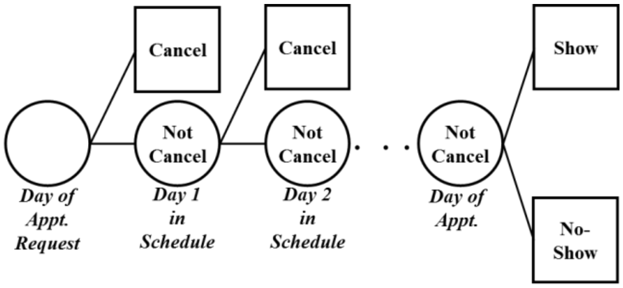


Figure 4.2. Possible Outcomes for Patients with Indirect Waiting

Additional assumptions of our model are as follows. Only requested appointments are considered; walk-ins are not considered. We assume that the clinic has a single server. When all patients scheduled for a single slot are homogeneous, as in the current model, they are serviced using a FCFS discipline. No patients are booked, a priori, to overtime slots. Overtime is used to accommodate service that runs over the predetermined service window.

4.3.1 Model

The goal of the clinic scheduler is to find the optimal schedule, $\boldsymbol{S}$, that maximizes the clinic expected net reward across the scheduling horizon. The clinic's expected net reward, $R(\boldsymbol{S})$, is equal to the service benefit from all completed appointments, minus total indirect waiting costs, total direct waiting costs, and the cost of clinic overtime. These components are influenced by the schedule, the number of people who are expected to no-show or to cancel, and the patient backlog.

The patient backlog is the number of patients who experience direct waiting at the end of a time-slot. Let $B[k,d,j]$ denote the probability of *k* patients in backlog at the end of slot *j*, $1 \le j \le N_+$, on day *d*, $1 \le d \le h$, where N_+ represents the latest possible appointment slot in a clinic, including overtime. Additionally, let $s(i,d,j)$ denote the number of patients scheduled to arrive at the beginning of slot *j* from day $i, 1 \le i \le d$ appointment requests, on day *d*. All patients scheduled on day 1 are considered same-day appointments, and no cancellations are considered. The backlog at the end of any slot depends on the number of people in backlog at the end of the prior slot, *l*, and the number of people assigned to the current slot, $s(1,1,j)$. Let

$b\left[n,p,k\right]=\binom{n}{k}p^{k}\left(1-p\right)^{n-k}$ be the probability mass function of a binomial random variable with parameters *n* and *p*. Thus, the backlog probability for the first day in the horizon can be expressed as follows:

$$
\begin{aligned}
&B\left[k,1,j\right]=\begin{cases}
\begin{array}{l}B\left[0,1,j-1\right]\times\left(b\left[s\left(1,1,j\right),p\left[1,1\right],0\right]+b\left[s\left(1,1,j\right),p\left[1,1\right],1\right]\right)+\\ B\left[1,1,j-1\right]\times b\left[s\left(1,1,j\right),p\left[1,1\right],0\right]\end{array} & \text{for } k=0\\
\displaystyle\sum_{l=0}^{K\left[1,j-1\right]} B\left[l,1,j-1\right]\times b\left[s\left(1,1,j\right),p\left[1,1\right],k-l+1\right] & \text{for } 1\le k\le K\left[d,j\right]
\end{cases}\\
&B[0,d,0]=1 \qquad\qquad (4.1)\\
&B[a,d,0]=0,\ a\in\mathbb{Z}^{+}
\end{aligned}
$$

where $K\left[d,j\right]=(K\left[d,j-1\right]+\sum_{i=1}^{d}s\left[i,d,j\right]-1)$ is the maximum backlog at end of slot *j* on day *d*.

The day 1 backlog equation is the backlog equation on page 6 of Zacharias and Pinedo (2015). To achieve a backlog of zero at the end of slot *j*, either there were no patients in backlog at the end of slot *j-1* and at most one patient shows in slot *j*, or there was one person in backlog at the end of slot *j-1*, and no patients show up in slot *j*. To achieve *k* people in backlog at the end of slot *j*, there must have been *l* people in backlog at the end of slot *j-1*, and *k-l+1* patients show in slot *j*.

To extend Equation (4.1), we develop the equation for all subsequent days in the scheduling horizon. Figure 4.3 depicts the inflow of patients into slot *j*, for day $d,\ d\ge 1$, in order to realize *k* people in backlog at the end of slot *j*. In general, the backlog at the end of slot *j* is affected by *l* (the number of patients in backlog at the end of slot *j-1*), *g* (the number of patients who are assigned from previous day's requests), the number of patients assigned from the current day's requests, and the person serviced in slot *j*.

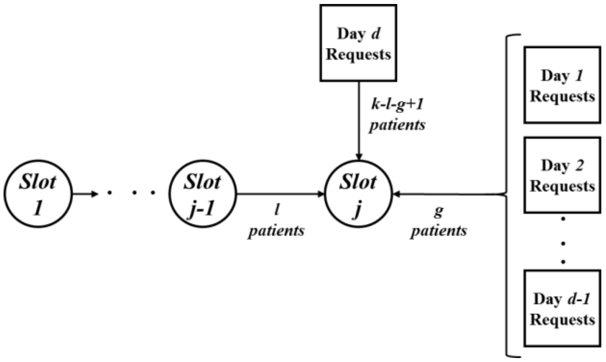


Figure 4.3. Inflow of Patients into Slot j to Reach k Patients in Backlog

The probability of *l* patients in backlog at the end of slot *j-1* can be represented recursively through a backlog equation. The probability of *g* patients showing in slot *j* on day *d* from the day *i*'s assignments is given by

$$\begin{aligned}
&\gamma[g,i,d,j]=\sum_{a=0}^{g}\gamma[a,i-1,d,j]\times\sum_{z_i=0}^{s(\mathrm{i},d,j)}b\left[s(\mathrm{i},d,j),\theta[i,d],z_i\right]\times b\left[z_i,p(i,d),g-a\right] \qquad i=1,..,d-1\\
&\gamma[0,0,d,j]=1 \qquad\qquad (4.2)\\
&\gamma[a,0,d,j]=0,\ a\in\mathbb{Z}^{+}
\end{aligned}$$

where $\theta[i,d]=\beta^{d-i-1}\theta$ denotes the probability that a patient does not cancel, and z_i is the number of people who do not cancel from day *i* requests. The number of people from prior day's requests assigned to slot *j* on day *d* is given by $L[d,j]=\sum_{i=1}^{d-1}s(i,d,j)$. Then, the backlog probability equation for day *d, d>1* is given by

$$\underset{d>1}{B[k,d,j]}=\begin{cases}
B[0,d,j-1]\times\left\{\gamma[0,d-1,d,j]\times\left(b\left[s(d,d,j),p(d,d),0\right]+b\left[s(d,d,j),p(d,d),1\right]\right)\right\}+ & \\
B[0,d,j-1]\times\gamma[1,d-1,d,j]\times b\left[s(d,d,j),p(d,d),0\right]+ & \text{for } k=0\\
B[1,d,j-1]\times\gamma[0,d-1,d,j]\times b\left[s(d,d,j),p(d,d),0\right] & \\
\sum_{l=0}^{K[d,j-1]}B[l,d,j-1]\times\sum_{g=0}^{L[d,j]}\gamma[g,d-1,d,j]\times b\left[s(d,d,j),p(d,d),k-l-g+1\right] & \text{for } 1\le k\le K[d,j]
\end{cases} \qquad (4.3)$$

Equation (4.3) for day *d, d>1* follows similar logic to that of Equation (4.1) for day 1, but also accounts for cancellations and for previous day's assignments.

4.3.2 Clinic Service Benefit

The clinic is assumed to receive a benefit for every patient serviced in the scheduling horizon. The benefit corresponds to the financial profit, or goodwill received from attending to a patient, and thus is applicable to both not-for-profit and for-profit organizations. The clinic receives a benefit, *π*, for seeing a patient. Assuming that the total benefit to the clinic is linear in the number of patients who show, the expected service benefit function for schedule ***S*** is:

$$\Pi(\boldsymbol{S}) = \sum_{d=1}^{h} \pi S[d] \tag{4.4}$$

where $S[d] = \sum_{i=1}^{d}\sum_{j=1}^{N} s(i,d,j) \times p[i,d] \times \theta[i,d]$ denotes the expected number of completed appointments on day *d*.

4.3.3 Clinic Indirect Waiting Time Cost

The clinic is penalized for each day a patient is delayed service. We assume that patients prefer to be seen as soon as possible to their request day. A penalty, *δ*, is incurred by the clinic for each day a patient's appointment is scheduled later than their request day, even if the patient later cancels or no-shows for that appointment. Our formulation is similar to the indirect waiting calculation in Samorani and LaGanga (2015). The total indirect waiting cost for schedule ***S*** is:

$$I(\boldsymbol{S}) = \sum_{d=1}^{h} \delta A[d] \tag{4.5}$$

where $A[i] = \sum_{d=i}^{h}\sum_{j=1}^{N} s(i,d,j) \times (d-i)$ denotes the total patient delay for patients who request an appointment for day *i*.

4.3.4 Patient Direct Waiting Time Cost

A potential consequence of overbooking appointment slots is the need for patients to wait for service. The cost of patient waiting quantifies patient dissatisfaction, loss of patient goodwill, and potential loss of business. We assume the patient's attendance behavior is sensitive to waiting time. Let *w*(*k*) represent the waiting cost function, where *k* denotes the number of appointment slots a patient must wait for service. We assume that *w*(*k*) is a convex function for $k \geq 0$, and *w*(*0*)*=0*. Thus, the expected waiting time cost across all appointment slots and possible levels of backlog is:

$$W(\boldsymbol{S}) = \sum_{d=1}^{h} \sum_{j=1}^{N_+} \sum_{k=0}^{K[d,\mathrm{j}]} B[k,d,j] \times w(k) \tag{4.6}$$

where N_+ represents the latest possible appointment slot in a clinic, including overtime.

4.3.5 Clinic Overtime Cost

The need to service patients during overtime, in order to service patients waiting at the end of the last scheduled slot of the day, is another consequence of overbooking. The cost of overtime is realized by the clinic, in costs such as provider time, wages paid to clinic staff, and loss of goodwill with the patient. Let *y*(*k*) represent the overtime cost incurred per slot of overtime used, given schedule ***S***. The function is assumed to be convex and to behave similarly to the waiting cost function. The expected clinic overtime costs for the scheduling horizon are:

$$O(\boldsymbol{S}) = \sum_{d=1}^{h} \sum_{k=0}^{K[d,N]} B[k,d,N] \times y(k) \tag{4.7}$$

where each term of the summation represents the expected overtime penalty for *k* patients waiting for service at the end of the last scheduled appointment slot.

4.3.6 Clinic Net Reward

The total expected clinic net reward is equal to the service benefit, minus the waiting time costs and the overtime cost. For this application, we assume that *w*(*k*) and *y*(*k*) are linear functions of *k*. Let $w(k) = \omega \times k$, where *ω* is the cost incurred for each slot a patient waits for service, and $y(k) = \sigma \times k$, where *σ* is the cost incurred for each slot of overtime used by the clinic. Given the prior definitions, the clinic net reward is given by Equation (4.8).

$$R(\boldsymbol{S}) = \sum_{d=1}^{h}\left(\pi S[d] - \delta A[d] - \sum_{j=1}^{N_+}\sum_{k=0}^{K[d,j]} B[k,d,j]\times\omega\times k - \sum_{k=0}^{K[d,N]} B[k,d,N]\times\sigma\times k\right) \tag{4.8}$$

4.4 THE OVERBOOKING MODEL

The goal of the clinic scheduler is to assign the patient appointment requests to appointment slots across the horizon, so that the expected net reward is maximized. For our application, we assume the clinic scheduler will need to decide between overbooking patients on their request day, or making them incur indirect waiting. We formulate the problem as the nonlinear integer program shown below in Equations (4.9)-(4.11). The decision variables for the problem are the *s*(*i*,*d*,*j*) values, or the number of people booked in slot *j* on day *d* from day *i* appointment requests.

$$\max \sum_{d=1}^{h}\left(\pi S[d] - \delta A[d] - \sum_{j=1}^{N_+}\sum_{k=0}^{K[d,j]} B[k,d,j]\times\omega\times k - \sum_{k=0}^{K[d,N]} B[k,d,N]\times\sigma\times k\right) \tag{4.9}$$

$s.t.$

$$\sum_{d=i}^{h}\sum_{j=1}^{N} s(i,d,j) = M_i, \quad \forall i \tag{4.10}$$

$$s(i,d,j) \in \{0,1,\ldots,M_i\}, \quad \forall (d,i,j) \tag{4.11}$$

The formulation in (4.9)-(4.11) maximizes the expected net reward over all days in the horizon. Constraint (4.10) ensures that each patient is assigned to a slot, and that assignments do not exceed the number of requests. Constraint (4.11) constrains the decision variables to be integers and between zero and the maximum number of requests. Equation (4.8) is an extension of the models in LaGanga and Lawrence (2012) and in Zacharias and Pinedo (2015), and so inherits similar properties within a multi-day framework. In the next section, we define properties of the model that assist in determining how to overbook up to two patients per day during the scheduling horizon.

4.5 MODEL PROPERTIES

To determine if a patient should be overbooked, we calculate the change in the objective function of Equation (4.8) when adding an additional patient. When that value is non-negative, overbooking a patient increases the objective function. The following propositions characterize when and where up to two patients per day should be overbooked. Proposition 1 provides a general overbooking rule. Propositions 2 through 6 apply to the first patient to be overbooked; and Propositions 7 through 10 apply for the second patient to be overbooked on a given day.

Propositions 1, 2, and 3 correspond to Propositions 3, 4, and 5 in LaGanga and Lawrence (2012), adapted for our notation. Note that those authors only consider overbooking a single patient, with a one day scheduling time horizon.

PROPOSITION 1. *A clinic schedule that fills all available appointment slots in a day before overbooking, has greater reward than one that overbooks when an open slot is available.*

PROPOSITION 2. *A clinic day with N+1 appointment requests and N appointment slots achieves a maximal reward when the additional patient is overbooked in slot j*, according to the following rules:*

$$\begin{array}{llll} \textbf{(i)}\ \ \mathit{if} & \dfrac{\sigma}{\omega} > \dfrac{p}{(1-p)} & \mathit{then} & j^* = 1 \\ \textbf{(ii)}\ \ \mathit{if} & \dfrac{\sigma}{\omega} < \dfrac{p}{(1-p)} & \mathit{then} & j^* = N \\ \textbf{(iii)}\ \mathit{if} & \dfrac{\sigma}{\omega} = \dfrac{p}{(1-p)} & \mathit{then} & j^* = any\ j \in \{1,\ldots,N\} \end{array} \tag{4.12}$$

PROPOSITION 3. *In a clinic day with N+1 appointment requests and N appointment slots, overbooking the additional patient in slot j* results in increased net reward, according to the following rules:*

$$\begin{array}{llll} \textbf{(i)} & j^* = 1, & \mathit{if} & \pi \ge p\left(\omega\left(\dfrac{1-p^N}{1-p}\right) + \sigma p^{N-1}\right) \\ \textbf{(ii)} & j^* = N, & \mathit{if} & \pi \ge p(\omega + \sigma) \\ \textbf{(iii)} & j^* = any\ j \in \{1,\ldots,N\}, & \mathit{if} & \pi \ge p\left(\dfrac{\omega}{1-p}\right) \end{array} \tag{4.13}$$

4.5.1 Scheduling the First Overbooking Request of a Clinic Day

Proposition 2 details where a single patient should be overbooked within a clinic day. The placement is dependent on the value of the overtime cost, σ, in relation to the waiting cost, ω, and the odds of a patient showing, $\frac{p}{1-p}$. For an overtime to waiting cost ratio greater than the odds ratio of a patient showing, the patient should be overbooked in the first slot of the day. If the overtime to waiting cost ratio is less than $\frac{p}{1-p}$, then the patient should be scheduled in the last slot of the day. Otherwise, the patient may be scheduled in any slot.

Proposition 3 specifies when it is optimal to overbook the additional patient. Optimality is determined based upon when in the day the additional patient is overbooked. The left-hand side (LHS) of the optimality rules is the service benefit for seeing the additional patient, and the right-hand side (RHS) is the expected service cost if the patient is overbooked. When the service benefit is greater than or equal to the service cost, it is optimal to overbook the additional patient.

PROPOSITION 4. *Let i denote a day with N booked appointments, for which an additional patient requests an appointment. Let d represent a day in the future of the scheduling horizon with appointment availability, and let* $r = \alpha^{d-i}\beta^{d-i-1}\theta$. *The clinic achieves a maximal reward by overbooking the additional patient in slot j* on day i, according to the following rules:*

$$\begin{aligned}
&\textbf{(i)} \quad j^* = 1, && \textit{if} \quad \pi \ge \frac{p\left(\omega\left(\dfrac{1-p^N}{1-p}\right) + \sigma p^{N-1}\right) - \dfrac{\delta}{p}(d-i)}{(1-r)} \\
&\textbf{(ii)} \quad j^* = N, && \textit{if} \quad \pi \ge \frac{p(\omega+\sigma) - \dfrac{\delta}{p}(d-i)}{(1-r)} \qquad (4.14) \\
&\textbf{(iii)} \quad j^* = any\ j \in \{1,\ldots,N\}, && \textit{if} \quad \pi \ge \frac{p\left(\dfrac{\omega}{1-p}\right) - \dfrac{\delta}{p}(d-i)}{(1-r)}
\end{aligned}$$

Otherwise, the patient should be booked into any open slot on day d.

Because our model allows for scheduling over a given horizon, we are able to evaluate when it is optimal to defer a patient, as opposed to overbooking the patient on her request day. In the context of a specialty clinic, this proposition arises when a patient requests an appointment for a day in the scheduling horizon where all *N* slots are already booked, but there are subsequent days in the horizon with slot availability. The results of the proposition allow a clinic scheduler to determine when it is optimal to overbook a patient on her request day, and when to schedule her on a later day.

The LHS of the rules in (4.14) is the service benefit received from booking the additional patient on day *i*. The parameter *r* is the reduction in the probability that an appointment will be completed, if the patient incurs indirect waiting. Thus, (*1-r*) is the difference in the service benefit that will be received if the patient is scheduled on day *d* and not on day *i*. The RHS of the deferment rules represents the change in expected service costs of booking the additional patient on day *i* instead of on day *d*, divided by (*1-r*). Given that *α*, *β*, and *θ* are between 0 and 1, and (*d-i*) is always non-negative, *r* is always between 0 and 1, and (*1-r*) is always positive. As (*1-r*) increases, the patient is more likely to be overbooked on day *i* and to incur no direct waiting, because of the greater reduction in the probability of the appointment being completed. Because day *d* is not fully booked, if the patient is deferred to day *d*, the patient should be booked in any open slot.

When using this method to make informed overbooking decisions, it becomes important to consider the prospect of a patient cancelling the appointment. If cancellations are not considered, the denominator of the RHS, (*1-r*), where $r = \alpha^{d-i}\beta^{d-i-1}\theta$, only changes with *α*. When *θ* is assumed to equal 1, the denominator is smaller, thus, it would appear optimal to make a patient incur indirect waiting, when in fact, the patient should be overbooked on day *i*. Proposition 5 formally describes the way in which *θ* influences the day on which the patient is booked.

PROPOSITION 5. *Let i denote a day with N booked appointments, for which an additional patient requests an appointment. Let d represent a day in the future of the scheduling horizon with appointment availability. As a function of θ, the patient should be booked into slot j*, according to the following rules:*

$$\begin{array}{llll} \textbf{(i)} & j^* = 1, & if & \theta \le \dfrac{\pi - p\left(\omega\left(\dfrac{1-p^N}{1-p}\right) + \sigma p^{N-1}\right) + \dfrac{\delta(d-i)}{p}}{\pi\alpha^{d-i}\beta^{d-i-1}} \\ \textbf{(ii)} & j^* = N, & if & \theta \le \dfrac{\pi - p(\omega+\sigma) + \dfrac{\delta(d-i)}{p}}{\pi\alpha^{d-i}\beta^{d-i-1}} \\ \textbf{(iii)} & j^* = any\ j \in \{1,...,N\}, & if & \theta \le \dfrac{\pi - p\left(\dfrac{\omega}{1-p}\right) + \dfrac{\delta(d-i)}{p}}{\pi\alpha^{d-i}\beta^{d-i-1}} \end{array} \tag{4.15}$$

Otherwise, the patient should be booked into any open slot on day d.

We assert that the probability of retention for a patient affects the optimal clinic schedule. When overbooking one patient, cancellations can have an effect when making the decision to overbook the patient on day *i*, or to book in an available slot on day *d*, as in Proposition 4. Proposition 5 outlines when the retention probability, *θ*, will affect this decision.

Because day *d* is empty, Proposition 5 may also affect the slot placement of the patient. If overbooked on day *i*, the patient will be overbooked in slot *j**, but if booked on day *d*, the patient will be booked in any available slot, which might differ from the value of *j**.

PROPOSITION 6. *Let i denote a day with N booked appointments, for which an additional patient requests an appointment. Let d represent a day in the future of the scheduling horizon with N booked appointments, and let* $r = \alpha^{d-i}\beta^{d-i-1}\theta$. *A clinic schedule achieves a maximal reward by overbooking the additional patient in slot j* on day i, according to the following rules:*

$$\begin{array}{llll} \textbf{(i)}\ if & j^* = 1 & and & \pi \ge p\left(\omega\left(\dfrac{1-p^N}{1-p}\right) + \sigma p^{N-1}\right) - \dfrac{\delta(d-i)}{p(1-r)} \\ \textbf{(ii)}\ if & j^* = N & and & \pi \ge p(\omega+\sigma) - \dfrac{\delta(d-i)}{p(1-r)} \\ \textbf{(iii)}\ if & j^* = any\ j \in \{1,...,N\} & and & \pi \ge p\left(\dfrac{\omega}{1-p}\right) - \dfrac{\delta(d-i)}{p(1-r)} \end{array} \tag{4.16}$$

Otherwise, the patient should be overbooked in slot j on day d.*

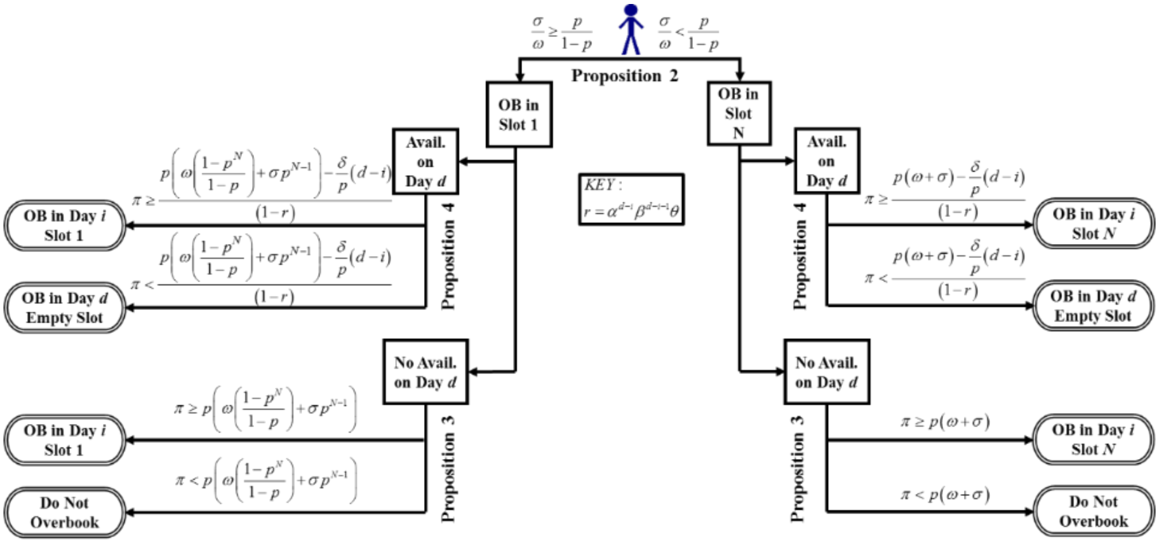
The scenario of Proposition 6 arises in a clinic when a patient requests an appointment in the scheduling horizon, but all days in the scheduling horizon are fully booked. Proposition 6 allows a clinic scheduler to determine on which day the patient should be overbooked. The results from Proposition 2, to determine where the overbooked patient should be scheduled in a clinic day, hold for all days in the clinic schedule. Thus, the patient should be overbooked in slot *j**, regardless of the day assignment. Given the relationship between Equations (4.16) and (4.13), (4.16) is always less than (4.13) for all cases. Thus, if it is not optimal to overbook on day *i*, it is not optimal to overbook the additional patient on day *d*.

Propositions 2 through 6 outline the steps for overbooking a single patient as follows:

1) Determine the optimal slot placement for overbooking the patient on her request day, using Proposition 2.

2) If another day in the scheduling horizon, day *d*, has available appointment slots, determine if it is optimal to book the patient on day *d* as opposed to overbooking on day *i* using Proposition 4. Book the patient on the optimal day.

3) If all days in the scheduling horizon are booked, determine if it is optimal to overbook the patient on day *i*, using Proposition 3. If it is optimal, overbook the patient on her request-day, day *i*. It is never optimal to overbook a single patient at a later day in the scheduling horizon, if all days are full. If it is not optimal, the schedule cannot accommodate overbooked patients, given the clinic parameters and patient attendance characteristics. The steps are also outlined in Figure 4.4.



$\frac{\sigma}{\omega} \geq \frac{p}{1-p}$
$\frac{\sigma}{\omega} < \frac{p}{1-p}$
Proposition 2
OB in Slot 1
OB in Slot N
KEY: $r = \alpha^{d-i}\beta^{d-i-1}\theta$
Avail. on Day *d*
Proposition 4
$\pi \geq \frac{p\left(\omega\left(\frac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right)-\frac{\delta}{p}(d-i)}{(1-r)}$
OB in Day *i* Slot 1
$\pi < \frac{p\left(\omega\left(\frac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right)-\frac{\delta}{p}(d-i)}{(1-r)}$
OB in Day *d* Empty Slot
No Avail. on Day *d*
Proposition 3
$\pi \geq p\left(\omega\left(\frac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right)$
OB in Day *i* Slot 1
$\pi < p\left(\omega\left(\frac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right)$
Do Not Overbook
Avail. on Day *d*
Proposition 4
$\pi \geq \frac{p(\omega+\sigma)-\frac{\delta}{p}(d-i)}{(1-r)}$
OB in Day *i* Slot *N*
$\pi < \frac{p(\omega+\sigma)-\frac{\delta}{p}(d-i)}{(1-r)}$
OB in Day *d* Empty Slot
No Avail. on Day *d*
Proposition 3
$\pi \geq p(\omega+\sigma)$
OB in Day *i* Slot *N*
$\pi < p(\omega+\sigma)$
Do Not Overbook

Figure 4.4. Flowchart for Overbooking a Single Patient in the Scheduling Horizon

4.5.2 Examples for Scheduling the First Overbooking Request for a Clinic Day

Figures 4.5 through 4.7 and Table 4.1 demonstrate the results of Propositions 2 through 6, given sample parameters. The shaded regions in the figures represent where it is optimal to overbook (OB) the additional patient in slot 1, per Proposition 2, for $p=0.5$ through 0.9, $\omega=0.1$ through 1, and $\sigma=0.5$, 1, and 1.5. Each cell displays results for a ω and p combination. The shaded areas for each value of σ represent where $\frac{\sigma}{\omega} \geq \frac{p}{(1-p)}$; values where $\frac{\sigma}{\omega} = \frac{p}{(1-p)}$ have been grouped with where a patient should be overbooked in slot 1. The shaded areas are cumulative as σ increases. For example, when $p=0.5$and $\omega=0.6$, the additional patient should be overbooked in slot 1 if $\sigma=1$ and $\sigma=1.5$ and slot N if $\sigma=0.5$.

Similar to the previous scheduling literature (e.g. LaGanga and Lawrence (2012)), the overbooking strategy for the first patient is dependent on clinic cost parameters and on the show rate. As the clinic overtime cost parameter, σ, increases, the shaded areas become larger, and it is more likely a patient will be overbooked in slot 1. As the show rate, p, increases, the shaded areas become smaller, and the area where it is optimal to overbook in slot N, increases. These results agree with the scheduling practice of overbooking more in the beginning of the day, especially if the population of patients has a low estimated probability of show, to "pad" the schedule, and to ensure that the provider is not idle for the first slot of the day (Bailey 1952, LaGanga and Lawrence 2012).

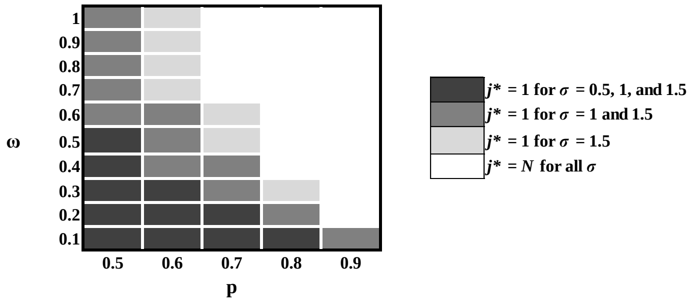


Figure 4.5. Per Proposition 2: Values of the Optimal Slot Placement for the First Overbooked Patient for Varying values of p and ω, and $\sigma=0.5$, 1, and 1.5

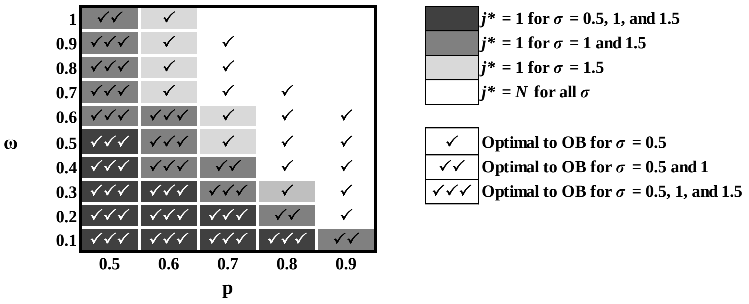


Figure 4.6. Per Proposition 3: Values of the Optimal Slot Placement for the First Overbooked Patient, and When it is Optimal to Overbook on day *i*, for Varying values of *p* and *ω*, *N=5*, and *σ =0.5*, *1*, and *1.5*

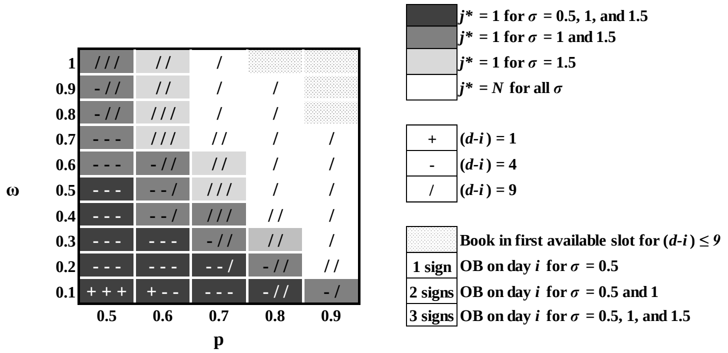


Figure 4.7. Per Proposition 4: Values of the Optimal Slot Placement for the First Overbooked Patient, and When it is Optimal to Overbook on day *i* versus day *d*, for Varying values of *p*, *ω*, (*d-i*), *N=5*, *π=1*, *α=β=0.95*, *δ=0.05* and *σ =0.5*, *1*, and *1.5* (assume day *d* is empty)

Table 4.1. Per Proposition 5: Upper Bound Values of the Probability of Retention When the Optimal Day to Overbook a Patient is Affected

		σ = 0.5		*σ* = 1		*σ* = 1.5	
	p	*ω* = 0.1	*ω* = 0.5	*ω* = 0.1	*ω* = 0.5	*ω* = 0.1	*ω* = 0.5
(*d-i*) = 1	0.9	0.543*	0.164*	0.102	-0.310*	-0.209	-0.784*
	0.8	0.663	0.276*	0.490	-0.145*	0.318	-0.566*
	0.7	0.835	0.391*	0.747	0.023*	0.658	-0.159
	0.6	0.954	0.509*	0.913	0.330	0.872	0.289
	0.5	1.039	0.632	1.023	0.615	1.007	0.599
(*d-i*) = 4	0.9	0.977*	0.461*	0.377	-0.183*	-0.046	-0.827*
	0.8	1.170	0.644*	0.936	0.072*	0.701	-0.501*
	0.7	1.443	0.839*	1.322	0.338*	1.202	0.090
	0.6	1.656	1.050*	1.600	0.807	1.544	0.752
	0.5	1.844	1.289	1.821	1.266	1.799	1.244
(*d-i*) = 9	0.9	2.296	1.435	1.294	0.359*	0.588	-0.717*
	0.8	2.702	1.824	2.310	0.867*	1.918	-0.090*
	0.7	3.264	2.255	3.063	1.418	2.862	1.005
	0.6	3.762	2.750	3.669	2.345	3.576	2.252
	0.5	4.275	3.348	4.238	3.311	4.200	3.274

* indicates where the patient is overbooked on day *i* slot *N*

Figure 4.6 depicts the same information as Figure 4.5, updated with whether it is optimal to overbook the additional patient, in the preferred slot. For this example, *N=5;* all other parameters remain the same. One check mark, "✓", represents where it is optimal to overbook when *σ=0.5*, two checks, "✓✓", represents where it is optimal to overbook when *σ=0.5* and *1,* and three checks, "✓✓✓", represents where it is optimal to overbook for all three values of *σ*. For example, the two checks when *p=0.8* and *ω=0.2* in Figure 4.6 denote that it is optimal to overbook the additional patient in slot *N* for *σ=0.5*, and it is optimal to overbook the additional patient in slot 1 for *σ=1*.

For large values of *ω* and *p*, it is never optimal to overbook a patient. This indicates that, when a patient base has a high probability of completing a request-day appointment, and patient waiting is costly to a clinic, the clinic should never overbook an additional patient. This corresponds with the results in the existing literature (e.g. Huang and Zuniga (2012) and LaGanga and Lawrence (2007)), that overbooking is most beneficial when no-show rates are high (show rates are low).

Figure 4.7 is a modification of Figure 4.5 that incorporates the results of Proposition 4. Figure 4.7 depicts when it is optimal to overbook a patient on day *i*, as opposed to booking the patient in an available slot on day *d*, where (*d-i*)*=1,4,* and *9*. These values represent booking one day in advance, 5 days in advance, and 10 days in advance. Combinations of *ω* and *p* for which it is optimal to overbook on day *d* for all *σ* are shaded with grey dots; all other shaded areas have the same interpretation as in Figures 4.5 and 4.6. The values of (*d-i*) represent the number of days of indirect waiting a patient will incur if the patient is booked on day *d*. A plus sign, "+", is used to indicate when it is optimal to book on day *d* when (*d-i*)*=1*, a minus sign, "-", is used when (*d-i*)*=4*, and a front slash sign, "/", when (*d-i*)*=9* . The number of symbols in each cell represents the values of *σ* for which booking on day *d* is optimal, one symbol for *σ=0.5,* two symbols for *σ=0.5* and *1,* and three symbols for *σ=0.5, 1,* and *1.5*. The optimality results are cumulative; if it is optimal to overbook on day *i* for (*d-i*)=*x*, it is optimal for (*d-i*)≥*x*. For example, the one plus sign and two minus signs when *p=0.6* and *ω=0.1* denote that it is optimal to overbook on day *i* slot 1 when (*d-i*)*=1* only if *σ=0.5,* and optimal to overbook on day *i* slot 1when (*d-i*)*=4* for all values of *σ.*

As (*d-i*) increases and *σ* decreases, it is more likely that a patient will be overbooked on day *i* and to incur no indirect waiting. For all cases where it is suboptimal to overbook a patient on day i slot N, per Proposition 3, it is optimal to overbook the patient on day d. Thus, analyzing the optimal slot placements over a scheduling horizon, as opposed to a single day, allows the model to accommodate more patients, which can assist in helping a clinic alleviate access issues.

Additionally, there are instances when it was optimal to overbook the additional patient on day i, per Proposition 3, indicated with a check mark in Figure 4.6, but, after evaluating Proposition 4, it is optimal to overbook the patient on day d. For example, from Proposition 3, when $p=0.7$ and $\omega=0.3$, it is optimal to overbook the additional patient in slot 1 for $\sigma=1$ and $\sigma=1.5$, and in slot N for $\sigma=0.5$. Given the alternative to book the patient in an empty slot when $(d\text{-}i)=1$, it is always optimal to do so, for the values used in this example, for all σ. When $(d\text{-}i)=9$, it remains optimal to overbook the patient on day i, for all σ. These results allow a clinic to optimally evaluate where a patient should be placed in the schedule, to allow for the least amount of patient backlog and clinic overtime. This can assist in improving patient satisfaction, as the patients will incur less waiting when they are in the clinic.

Table 4.1 lists when the probability of retention, θ, affects the day on which the additional patient should be overbooked. Values in the table are the calculations of the expressions in Equation (4.15). When the clinic's expected probability of retention is less than or equal to the corresponding value in Table 4.1, then the clinic should overbook the patient on day i; otherwise, the clinic should book the patient in any available slot on day d. The values in Table 4.1 are decreasing as p, σ, and ω increase, and increasing as $(d\text{-}i)$ increases. So, as the patient's probability of show increases, the more likely he is to be booked on day d and to incur indirect waiting.

4.5.3 Scheduling the Second Overbooking Request for a Clinic Day

Proposition 7. *Let denote a day with $N+1$ booked appointments, for which an additional patient requests an appointment. If the additional patient is to be overbooked on day i, then a clinic schedule achieves a maximal reward by overbooking the additional patient in slot j^{**}, according to the following rules:*

$$\begin{aligned}
&(\mathbf{i}) \quad \textit{if} \quad j^*=1, && \textit{then} \quad j^{**}=\left\lceil \frac{Ln\left[\left(-A+\sqrt{A^2+4A}\right)/2\right]}{Ln[p]} \right\rceil \quad \textit{where} \quad A=p^N\left(\sigma/_{\omega}(1-p)-p\right)\\
&(\mathbf{iia})\ \textit{if} \quad j^*=N \quad \textit{and} \quad \frac{\sigma}{\omega}\geq\frac{(2p-1)}{(1-p)(2-p)}, && \textit{then} \quad j^{**}=1\\
&(\mathbf{iib})\ \textit{if} \quad j^*=N \quad \textit{and} \quad \frac{\sigma}{\omega}<\frac{(2p-1)}{(1-p)(2-p)}, && \textit{then} \quad j^{**}=N
\end{aligned} \tag{4.17}$$

Proposition 7 follows from Proposition 2, and identifies the optimal slot placement in a clinic day for a second overbooked patient. The conditions of Proposition 7 can arise in a specialty clinic when all days in the horizon are full, day *i* is overbooked with a single patient, and an additional patient requests service on day *i*. The optimal slot placement of the second overbooked patient is dependent on the slot placement of the first overbooked patient. For succinctness, we refer to the first overbooked patient as OB1 and to the second overbooked patient as OB2. We assume that the overbooking occurs on day *i*, and all patients have requested an appointment for day *i*. Note that *j*** denotes the optimal slot placement of OB2, if OB2 is overbooked on the same day as OB1. In addition, $\lceil x \rceil$ denotes the integer ceiling of *x*.

When OB1 is overbooked in slot *j*=1*, the value of *j*** is a function of the clinic parameters, the length of the clinic day, and the patient's probability of show, *p*. When $A=p^N\left(\sigma/_{\omega}(1-p)-p\right)=0$, i.e., $p=\frac{\sigma}{\sigma+\omega}$, *j*** is not defined, and it is not optimal to overbook OB2. Additionally, when *j*** is calculated to be a value greater than the value of *N*, or it is calculated that OB2 should be booked outside the length of the clinic day, it is not optimal to overbook OB2. When $A=p^N\left(\sigma/_{\omega}(1-p)-p\right)<0$, *j*** is not defined. This occurs when $\frac{\sigma}{\omega}<\frac{p}{1-p}$, and hence, *j*=N*, which would contradict the assumption that *j*=1*. When OB1 is overbooked in slot *j*=N*, the value of *j*** is dependent on the relationship of *σ* with *ω* and *p*, as in Proposition 2. For Proposition 7, we grouped the case when it is optimal to overbook in any slot with the case when it is optimal to overbook OB2 in slot 1.

Proposition 7 outlines the importance of assigning both a day and slot when overbooking. Assigning a patient to a day because it decreases the objective function is not sufficient. Without knowledge of the current schedule layout, it becomes difficult to determine how the overbooked patients should be placed in the schedule to cause the least amount of clinic disruption.

Proposition 8. *Let i denote a day with N+1 booked appointments, for which an additional patient requests an appointment. Let d represent a day in the future of the scheduling horizon with N booked appointments, and let $r=\alpha^{d-i}\beta^{d-i-1}\theta$. A clinic schedule achieves a maximal reward by overbooking the additional patient in slot j** on day i, according to the following rules:*

$$\begin{array}{llll}
\textbf{(i)} & j^{**}\text{ from Prop.7,} & \textit{if}\quad j^*=1 \quad \textit{and} & \pi \ge p\left(\omega\left(\dfrac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right)+\dfrac{p\left(\omega\left(\dfrac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)-\sigma p^N\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma\left(1-p\right)-\omega p\right)\right)-\dfrac{\delta}{p}\left(d-i\right)}{1-r} \\
\textbf{(ii)} & j^{**}=1, & \textit{if}\quad j^*=N \quad \textit{and} & \pi \ge p\left(\omega+\sigma\right)+\dfrac{p\left(\omega\left(\dfrac{1-p^{N-1}\left(2p-1\right)}{1-p}\right)+\sigma p^{N-1}\left(2-p\right)-\left(\omega+\sigma\right)\right)-\dfrac{\delta}{p}\left(d-i\right)}{1-r} \\
\textbf{(iii)} & j^{**}=N, & \textit{if}\quad j^*=N \quad \textit{and} & \pi \ge p\left(\omega+\sigma\right)+\dfrac{p\left(\omega+\sigma\left(1-p\right)\right)-\dfrac{\delta}{p}\left(d-i\right)}{1-r}
\end{array}$$

(4.18)

Otherwise, the patient should be overbooked in slot j on day d.*

From Proposition 6, we know that OB1 is always overbooked on day *i*, if it is optimal to overbook one patient. Thus, when OB2 requests an appointment for day *i*, the options are to overbook OB2 with OB1 on day *i* in the *j*** slot designated in Proposition 6, or overbook OB2 on day *d*. If OB2 is overbooked on day *d*, he is the first patient overbooked on that day, and his slot placement is equal to *j** as given in Proposition 2.

When *j*=1*, the service costs for overbooking OB2 on day *i* are dependent on the value of *j***. As in Proposition 4, as $(1-r)$ increases, the patient is more likely to be overbooked on day *i* and to incur no direct waiting.

Proposition 9. *Let i denote a day with N+1 booked appointments, for which an additional patient requests an appointment. Let d represent a day in the future of the scheduling horizon with N booked appointments. As a function of θ, the patient should be booked into slot j** on day i, according to the following rules:*

$$\textbf{(i)}\quad j^{**}\text{ from Prop.7,}\quad \textit{if}\quad j^{*}=1\quad \textit{and}\quad \theta \le \frac{\pi - p\left(\omega\left(\frac{1-p^{N}}{1-p}\right)+\omega\left(\frac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)+\sigma p^{N-1}-\sigma p^{N}\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma\left(1-p\right)-\omega p\right)\right)+\frac{\delta\left(d-i\right)}{p}}{\left(\pi - p\left(\omega\left(\frac{1-p^{N}}{1-p}\right)+\sigma p^{N-1}\right)\right)\alpha^{d-i}\beta^{d-i-1}}$$

$$\textbf{(ii)}\quad j^{**}=1,\quad \textit{if}\quad j^{*}=N\quad \textit{and}\quad \theta \le \frac{\pi - p\left(\omega\left(\frac{1-p^{N-1}\left(2p-1\right)}{1-p}\right)+\sigma p^{N-1}\left(2-p\right)-\left(\omega+\sigma\right)\right)+\frac{\delta\left(d-i\right)}{p}}{\left(\pi - p\left(\omega+\sigma\right)\right)\alpha^{d-i}\beta^{d-i-1}}$$

$$\textbf{(iii)}\quad j^{**}=N,\quad \textit{if}\quad j^{*}=N\quad \textit{and}\quad \theta \le \frac{\pi - p\left(2\left(\omega+\sigma\right)-\sigma p\right)+\frac{\delta\left(d-i\right)}{p}}{\left(\pi - p\left(\omega+\sigma\right)\right)\alpha^{d-i}\beta^{d-i-1}}$$

(4.19)

Otherwise, the patient should be booked into slot j on day d.*

Proposition 9 outlines how θ affects the clinic schedule, when deciding to overbook two patients on one day, or making a patient incur indirect waiting. The formulation of Proposition 9 is similar to Proposition 5, but for an additional patient.

PROPOSITION 10. *Let i denote a day with N+1 booked appointments, for which an additional patient requests an appointment. Let d represent a day in the future of the scheduling horizon with N booked appointments. It is optimal to overbook the additional patient, according to the following rules:*

$$\textbf{(i)}\quad j^{**}\text{ from Prop.7,}\quad \textit{if}\quad j^{*}=1,\quad \text{Prop7}\rightarrow \textit{Day } i,\quad \textit{and}\quad \pi \ge p\left(\omega\left(\frac{1-p^{N}}{1-p}\right)+\omega\left(\frac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)+\sigma p^{N-1}-\sigma p^{N}\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma\left(1-p\right)-\omega p\right)\right)$$

$$\textbf{(ii)}\quad j_d^{*}=1,\quad \textit{if}\quad j^{*}=1,\quad \text{Prop7}\rightarrow \textit{Day } d,\quad \textit{and}\quad \pi \ge p\left(\omega\left(\frac{1-p^{N}}{1-p}\right)+\sigma p^{N-1}\right)+\frac{\delta\left(d-i\right)}{pr}$$

$$\textbf{(iii)}\quad j^{**}=1,\quad \textit{if}\quad j^{*}=N,\quad \text{Prop7}\rightarrow \textit{Day } i,\quad \textit{and}\quad \pi \ge p\left(\omega\left(\frac{1-p^{N-1}\left(2p-1\right)}{1-p}\right)+\sigma p^{N-1}\left(2-p\right)\right)$$

$$\textbf{(iv)}\quad j_d^{*}=N,\quad \textit{if}\quad j^{*}=N,\quad \text{Prop7}\rightarrow \textit{Day } d,\quad \textit{and}\quad \pi \ge p\left(\omega+\sigma\right)+\frac{\delta\left(d-i\right)}{pr}$$

$$\textbf{(v)}\quad j^{**}=N,\quad \textit{if}\quad j^{*}=N,\quad \text{Prop7}\rightarrow \textit{Day } i,\quad \textit{and}\quad \pi \ge p\left(2\left(\omega+\sigma\right)-\sigma p\right)$$

$$\textbf{(vi)}\quad j_d^{*}=N,\quad \textit{if}\quad j^{*}=N,\quad \text{Prop7}\rightarrow \textit{Day } d,\quad \textit{and}\quad \pi \ge p\left(\omega+\sigma\right)+\frac{\delta\left(d-i\right)}{pr}$$

(4.20)

Where j_d^{*} denotes the slot in which OB2 is overbooked on day d.

Proposition 10 outlines when it is optimal to overbook OB2 after the day and slot placements have been determined. The LHS of the optimality rules is the benefit derived from overbooking OB2, and the RHS are the service costs for overbooking. When Proposition 8 leads to overbooking on day *i*, the RHS is the expected waiting and overtime from overbooking two patients on the same day. When the optimal day is day *d*, OB2 is the first overbooked patient on that day, and the optimality rules reflect this result. A flowchart for the steps to overbook OB2 can be found in Appendix B.

4.5.4 Examples for Scheduling the Second Overbooking Request for a Clinic Day

Figures 4.8 and 4.9 illustrate overbooking OB2 as per Proposition 7. The example is a continuation of the example in Section 5.2. In Figure 4.8, the numbers in each cell represent *j*** for *σ=0.5, 1,* and *1.5*, when *j*=1*, as per Proposition 7. For example, when *p=0.7* and *ω=0.1*, *j***=4, 3, 2, for *σ=0.5, 1,* and *1.5*, respectively. Cells with one number list *j*** when *σ=1.5*, cells with two stars list *j*** when *σ=1* and *1.5*, and cells with three numbers list *j*** for all 3 values of *σ*. Cells that are shaded for *j*=1* with no number indicate where it is not optimal to overbook OB2. When *j**=5*, OB2 is overbooked in slot *N*.

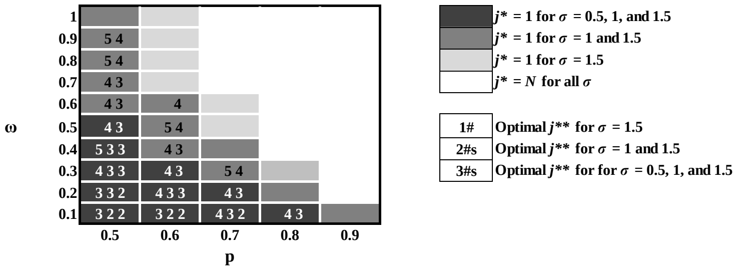


Figure 4.8. Per Proposition 6: Values of the Optimal Slot Placement for Second Overbooked Patient, when *j*=1*, for Varying values of *p* and *ω*, *N=5*, and *σ =0.5*, *1*, and *1.5*

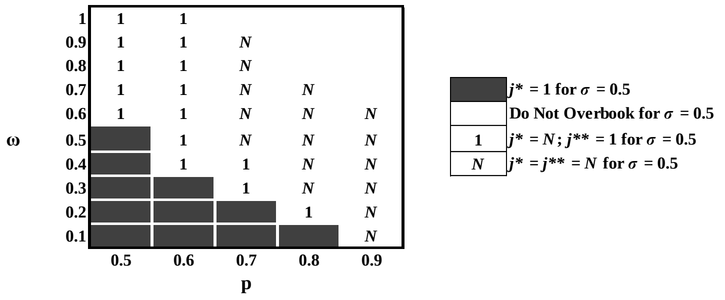


Figure 4.9. Per Proposition 6: Values of the Optimal Slot Placement for Second Overbooked Patient, when *j*=N*, for Varying values of *p* and *ω*, *N=5*, and *σ =0.5*, *1*, and *1.5*

In Figure 4.8, as *σ* increases, *j*** decreases, and OB2 is booked closer to the beginning of the clinic day. Overbooking towards the beginning of the day decreases the expected overtime, which is beneficial for larger values of σ. As p and ω increase, j^{**} increases, and OB2 is more likely to be overbooked in a later appointment slot. As in Proposition 2, that corresponds with the scheduling practice of "padding" the front end of a schedule with patients who are less likely to show, and scheduling patients more likely to show at the end of the day, to decrease the effects on waiting time accumulating throughout the day. Additionally, for the values represented in the figure, j^{**} never equals 1; so it is never optimal to overbook OB1 and OB2 in the same slot.

Figure 4.9 shows when OB2 should be overbooked in slot 1 or N. The shaded cells represent where $j^{*}=1$, and this case does not apply; cells with no shading and no j^{**} value represent where it is not optimal to overbook OB2. Similar to the case when $j^{*}=1$, as p and ω increase, j^{**} increases, and OB2 is more likely to be overbooked in slot N.

The optimal placement of two overbooked patients when they are overbooked sequentially is depicted in Figure 4.10. The results reflect the calculations from Propositions 8 and 10. The results are shown for a two day scheduling horizon, so $(d\text{-}i)=1$. The values in each cell are listed as OB1 day/slot; OB2 day/slot. When it is not optimal to overbook an additional patient, the cell lists DNOB, or Do Not Overbook, for the ω and p combination.

When $\sigma=1$ and 1.5, if it is optimal to overbook, the typical overbooking strategy is to overbook the first slot of day 1 and of day 2 with OB1 and OB2, respectively. The cells showing DNOB for OB2 are when j^{**} is not defined for the given values of ω and p. When $\sigma=0.5$, there is more variability in the slot placement of OB2. The most common overbooking strategy is to overbook both patients in slot N, on days 1 and 2. There are no instances where it is optimal to overbook both patients in the same slot.

σ = 0.5

ω	0.5	0.6	0.7	0.8	0.9
1	i/N; d/N	i/N; d/N	DNOB	DNOB	DNOB
0.9	i/N; d/N	i/N; d/N	i/N; DNOB	DNOB	DNOB
0.8	i/N; d/N	i/N; d/N	i/N; d/N	DNOB	DNOB
0.7	i/N; i/1	i/N; d/N	i/N; d/N	i/N; DNOB	DNOB
0.6	i/N; i/1	i/N; d/N	i/N; d/N	i/N; d/N	i/N; DNOB
0.5	i/1; DNOB	i/N; d/N	i/N; d/N	i/N; d/N	i/N; d/N
0.4	i/1; i/5	i/N; i/1	i/N; d/N	i/N; d/N	i/N; d/N
0.3	i/1; i/4	i/1; DNOB	i/N; d/N	i/N; d/N	i/N; d/N
0.2	i/1; i/3	i/1; i/4	i/1; DNOB	i/N; d/N	i/N; d/N
0.1	i/1; i/3	i/1; i/3	i/1; d/1	i/1; DNOB	i/N; d/N

p

σ = 1

ω	0.5	0.6	0.7	0.8	0.9
1	i/1; DNOB	DNOB	DNOB	DNOB	DNOB
0.9	i/1; i/5	DNOB	DNOB	DNOB	DNOB
0.8	i/1; d/1	DNOB	DNOB	DNOB	DNOB
0.7	i/1; d/1	DNOB	DNOB	DNOB	DNOB
0.6	i/1; d/1	i/1; DNOB	DNOB	DNOB	DNOB
0.5	i/1; d/1	i/1; d/1	DNOB	DNOB	DNOB
0.4	i/1; d/1	i/1; d/1	i/1; DNOB	DNOB	DNOB
0.3	i/1; d/1	i/1; d/1	i/1; d/1	DNOB	DNOB
0.2	i/1; i/3	i/1; d/1	i/1; d/1	i/1; DNOB	DNOB
0.1	i/1; i/2	i/1; d/1	i/1; d/1	i/1; d/1	i/1; DNOB

p

σ = 1.5

ω	0.5	0.6	0.7	0.8	0.9
1	DNOB	DNOB	DNOB	DNOB	DNOB
0.9	i/1; DNOB	DNOB	DNOB	DNOB	DNOB
0.8	i/1; d/1	DNOB	DNOB	DNOB	DNOB
0.7	i/1; d/1	DNOB	DNOB	DNOB	DNOB
0.6	i/1; d/1	i/1; DNOB	DNOB	DNOB	DNOB
0.5	i/1; d/1	i/1; d/1	DNOB	DNOB	DNOB
0.4	i/1; d/1	i/1; d/1	DNOB	DNOB	DNOB
0.3	i/1; d/1	i/1; d/1	i/1; d/1	DNOB	DNOB
0.2	i/1; d/1	i/1; d/1	i/1; d/1	DNOB	DNOB
0.1	i/1; i/2	i/1; d/1	i/1; d/1	i/1; d/1	DNOB

p

DNOB	**Do Not Overbook**
OB1 day/slot; OB2 day/slot	

Figure 4.10. Optimal Sequential Overbooking Strategies when Overbooking Two Patients for Varying Values of *p*, *ω*, *σ*, (*d*-*i*)=*1*, *π=1*, *N=5*, *α=β=0.95*, and *δ=0.05*

For each day in which there is only one overbooked patient, our sample specialty clinic could still be willing to overbook an additional patient. If we assume that scheduling occurs over a finite horizon (i.e., there are two days of availability in a week and all patients who need appointments in that week must be scheduled before the end of the horizon), and day *i* only has one overbooked patient, the clinic can evaluate overbooking another patient on day *i* as opposed to on day *d**, using the results from Propositions 7 through 9. In such conditions, *d** represents a day in the scheduling horizon that does not currently have an overbooked patient. As in Proposition 6, if a clinic is evaluating overbooking a patient, and the two days under consideration have the same number of patients, if it is not optimal to overbook on day *i*, it is not optimal to defer to the later day. If only one day is available for overbooking a second patient, the clinic can evaluate if the costs of overbooking that patient outweigh the benefits, similar to Proposition 7, before overbooking the patient.

4.6 EMPIRICAL RESULTS

To test the model formulation, and to compare the analytical results with the results as obtained from solving the model in Equations (4.9) through (4.11), we ran a series of optimizations for sample outpatient specialty clinics. Table 4.2 lists the parameters of the models that were executed.

Table 4.2. Parameters of Optimization for Sample Clinics

Parameter	***Value(s)***	***Parameter***	***Value(s)***
$\boldsymbol{h}$	2	$\boldsymbol{\sigma}$	0.5, 1, and 1.5
$\boldsymbol{N}$	5	$\boldsymbol{\alpha}$	0.95
$\boldsymbol{M}$	{(5,6,7,8,9,10),5}	$\boldsymbol{\beta}$	0.95
$\boldsymbol{\delta}$	0.05	$\boldsymbol{p}$	0.5 through 0.9
$\boldsymbol{\pi}$	1	$\boldsymbol{\theta}$	0.7 through 1
$\boldsymbol{\omega}$	0.1, 0.3, 0.5		

The examples are representative of a specialty clinic that is open for two (h) days per week, with five (N) appointment slots per day. We assume there are five to ten requests on the first day, and five appointment requests for the second day. We assume that all requests for additional appointments occur on the first day, though some patients may have to incur indirect waiting, and be seen on the second day of the scheduling horizon. The clinic receives an award of one (π) for seeing a patient. Indirect waiting costs the clinic 0.05 (δ) per day, direct waiting time in the clinic costs 0.1, 0.3, or 0.5 (ω), and clinic overtime costs either negate the benefit of seeing a patient, or negate the benefit, plus a penalty of 0.5 (σ). It is assumed that the probabilities of show and retention decrease by 5% for each day of indirect waiting (α, β), the probability of show (p) for the patient population ranges from 0.5 and 0.9, and that the probability of retention (θ) for the patient population ranges from 0.7 to 1. A retention rate of 1 indicates a scheduling model that does not account for patient cancellation, or that assumes cancellations are negligible.

The results from our empirical model computations matched the analytical results detailed in Section 4.5 for overbooking up to two patients. Below we discuss the results of the models that involved overbooking more than two patients, to provide insight into how a clinic could handle more advanced overbooking situations.

4.6.1 Overbooking with One to Five Additional Patients

The sample clinic optimizations were performed with five to ten requests on day 1 and five requests on day 2. Because *N=5*, that represents overbooking levels ranging from zero to five patients. Figure 4.11a indicates the optimal number of day 1 requests to accept for *p=0.5* through *0.9* and *ω=0.1; σ=1*, *ω=0.1; σ=1.5*, *ω=0.5; σ=1*, and *ω=0.5; σ=1.5.* Because the greatest service reward is achieved when *θ=1*, that is the optimal value of *θ* for all sample clinics. Figure 4.11b is the percentage change from the baseline service reward for each probability of show. The baseline service rewards are 5, 6, 7, 8, and 9, for *p=0.5* through *0.9*, respectively. As *ω* and *p* increase, the optimal overbooking levels decrease. When the overbooking levels are the same for a value of *ω*, the change in service reward is greater for the smaller value, *ω=0.1*. Similar results are seen with an increase in *σ*.

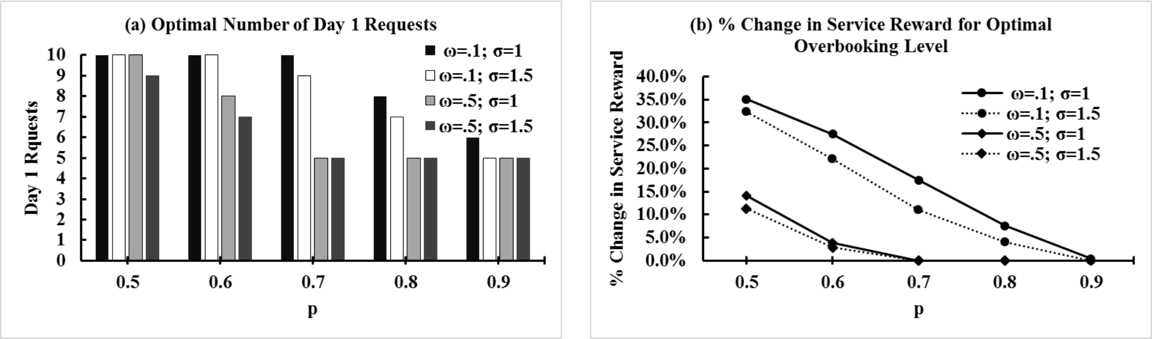


Figure 4.11. (a) Optimal Number of Requests Accepted on Day 1 and (b) Percentage Change in Service Benefit for *p=0.5* through *0.9* and *ω=0.1; σ=1*, *ω=0.1; σ=1.5*, *ω=0.5; σ=1*, and *ω=0.5; σ=1.5*

4.6.2 Patient Access Levels

Figure 4.12 shows the expected number of completed appointments, based upon the optimal number of day 1 requests from Section 4.6.1. The value is a function of the probability of a completed appointment, which decreases if a patient incurs indirect waiting, and the number of people who requested appointments. Greater values of *p* and the number of day 1 requests do not always correspond with more completed appointments. This is due to the decrease in clinic benefit if a patient is deferred, and the propensity to defer patients when their probability of show is greater. For the sample clinics represented in Figure 4.12, the greatest number of expected completed appointments occurs when $\omega=0.1$, $\sigma=1$, and $p=0.7$. The greater the expected number of patients to complete appointments, the more productive the clinic, and more patients can be seen in a shorter timeframe. Given the access issues faced in outpatient specialty clinics, our analysis enables a clinic to assess booking strategies to accommodate the maximum number of patients.

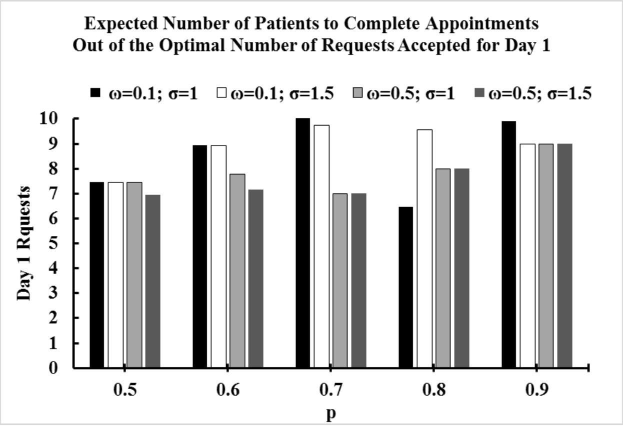


Figure 4.12. Expected Number of Patients to Complete Appointments for $p=0.5$ through 0.9 and $\omega=0.1; \sigma=1$, $\omega=0.1; \sigma=1.5$, $\omega=0.5; \sigma=1$, and $\omega=0.5; \sigma=1.5$

4.7 DISCUSSION AND CONCLUSIONS

No-shows and cancellations can lead to scheduling inefficiencies and to provider underutilization/overutilization of resources. Overbooking is a potential solution to mitigate these negative effects. Overbooking strategies that are informed by an analytical model are preferred to naïve strategies that are based solely on intuition. In this chapter, we presented propositions that allow a clinic to determine overbooking strategies for up to two patients per day in an online scheduling environment. The propositions are derived from a model that incorporates clinic parameters and patient no-show and cancellation probabilities over a multi-day horizon. Because we design strategies for patients who are overbooked sequentially, our strategies can be utilized in an online context.

The research presented in this chapter contributes to the literature in a number of ways. First, we show that the optimal overbooking strategy is a function of both the no-show and the cancellation probabilities. The probability that a patient will cancel her appointment affects the expected service benefit from seeing patients and the probability of a future slot going unused. Failing to incorporate cancellations in an overbooking model may cause a clinic to develop suboptimal overbooking decisions. Second, we show the threshold probability of retention at which the schedule will change. When overbooking up to two patients per day, the probability that a patient will cancel can affect on which day the patient will be scheduled. We show, analytically, that the probability of retention that induces that change is a nonlinear function of the clinic parameters and of the patient probability of show. Third, we consider both direct and indirect waiting times in our model. Typical models in the literature consider only the time a patient waits in a clinic for service. We show that the scheduling decisions change based upon the number of indirect waiting days a patient incurs. In the absence of an open-access scheduling practice, both indirect and direct waiting must be considered.

Finally, we show how our proposed overbooking strategies may be used to motivate managerial decision making in a clinic. The placement of the first overbooked patient in a clinic day is dependent upon the patient's probability of show, and the direct waiting and overtime costs. As overtime costs increase, the patient should be booked at the beginning of the day. For high show probability and high direct waiting time cost, the patient should be overbooked at the end of the day, until it becomes sub-optimal to overbook. The greater a patient's show probability, it is more likely that she will still show for an appointment, even if she incurs indirect waiting. If it is optimal to overbook a second patient in a clinic day, the optimal placement is dependent on the placement of the first overbooked patient. The most common strategy, over the range of values discussed in this chapter, is to overbook the first slot of the day for each day in the scheduling horizon, before overbooking two patients in one day. We found no instances where it is optimal to overbook both patients in the same slot. In a clinic with access issues, these rules and the propositions outlined in this chapter can help a clinic determine how it should adjust clinic parameters or perform mitigation strategies to allow access to additional patients.

The results of our work motivate several extensions. While we model with homogeneous no-show and cancellation probabilities, due to the scheduling practice of the clinic we observed, a possible extension could be to include heterogeneous probabilities, and investigate how the optimal decisions change. The assumptions that patients show punctually, and that service time is fixed, might also be relaxed. Larger clinics could be amenable to overbooking more than two patients in one day; our work can be extended to include these analytical results. Adding more overbooked patients would allow us to analyze the relationship between multiple overbooked slots in one day. Currently, with two overbooked patients, it is never optimal to book more than two patients in a single slot. With additional overbooked patients, it could become optimal to group overbooked patients together, as opposed to spreading them throughout the day or the scheduling horizon. Lastly, we assume linear cost structures for direct waiting and overtime. Those assumptions can be modified, to include different functional forms for direct waiting and overtime costs.

5.0 SUMMARY AND FUTURE WORK

No-shows and cancellations can lead to scheduling inefficiencies and to provider underutilization/overutilization of resources. Overbooking is a potential solution to mitigate these negative effects. In this dissertation, I presented models to predict no-show and cancellation probabilities, and overbooking strategies to overbook up to two patients in a clinic day based upon patient no-show and cancellation probabilities. These models are novel approaches to studying patient behavior and appointment scheduling, and give insight into how patient behavior should be used in addressing patient access issues.

The general overbooking rules generated from the overbooking model can be used to inform managerial decision making in a clinic. The placement of the first overbooked patient in a clinic day is dependent upon the patient's probability of show, and the direct waiting and overtime costs. As overtime costs increase, the patient should be booked at the beginning of the day. For high show probability and high direct waiting time cost, the patient should be overbooked at the end of the day, until it becomes sub-optimal to overbook. The greater a patient's show probability, it is more likely that she will still show for an appointment, even if she incurs indirect waiting. If it is optimal to overbook a second patient in a clinic day, the optimal placement is dependent on the placement of the first overbooked patient. The most common strategy, over the range of values discussed in this chapter, is to overbook the first slot of the day for each day in the scheduling horizon, before overbooking two patients in one day. We found no instances where it is optimal to overbook both patients in the same slot.

The results presented in this dissertation motivate several extensions to the models. The no-show prediction model does not address the time between appointments. Given our findings concerning the importance of the sequence of no-shows, the time between those no-shows may also play a role. Evaluation of the success of the cancellation model involves assessing the accuracy of the class assignment and the time to cancel prediction. To our knowledge, there is no current measure to evaluate the accuracy of both predictions simultaneously. The development of this metric would be a valuable contribution to literature. The overbooking model utilizes homogeneous no-show and cancellation probabilities for each patient. An extension could be to include heterogeneous probabilities, and investigate how the optimal decisions change. The assumptions that patients show punctually, and that service time is fixed, might also be relaxed. Larger clinics could be amenable to overbooking more than two patients in one day; our work can be extended to include these analytical results. Lastly, we assume linear cost structures for direct waiting and overtime. Those assumptions can be modified, to include different functional forms for direct waiting and overtime costs.

APPENDIX A

NO-SHOW HISTORY PREDICTIVE MODEL APPENDICES

Notation

- All vectors are assumed to be column vectors
- $A_k = \left[a_{ij} \right]_k$ and represents the Hessian of the function F_k, whose dimensions are $(k+1)\times(k+1)$

A.1 DERIVATIVES OF OBJECTIVE FUNCTION

$$\frac{\partial F_k}{\partial z_{0k}} = \sum_{i=1}^{2^k} v_{ik} \left(p_{ik} - z_{0k} - \sum_{j=1}^{k} e^{-\lambda_{jk}} x_{ijk} \right) = 0 \tag{A.1}$$

$$\frac{\partial F_k}{\partial e^{-\lambda_{jk}}} = \sum_{i=1}^{2^k} v_{ik} x_{ijk} \left(p_{ik} - z_{0k} - \sum_{j=1}^{k} e^{-\lambda_{jk}} x_{ijk} \right) = 0, \tag{A.2}$$

Equations (A.1) and (A.2) may be simplified, because the x_{ijk} values are binary and known. Assume that the 2^k possible historical sequences are indexed by their binary values, so that the sequence indexed as $i=1$ is the sequence in which all prior outcomes were failures, and the sequence indexed $i=2^k$ is the sequence in which all prior outcomes were successes. Then a table of x_{ij} values for any k has a straightforward structure which can be used to simplify Equations (A.1) and (A.2). Table A1 shows an example of such a table for $k=2$.

Table A1. Table of x_{ij} values for k=2

$$\begin{bmatrix} & j=1 & j=2 \\ i=1 & 0 & 0 \\ i=2 & 0 & 1 \\ i=3 & 1 & 0 \\ i=4 & 1 & 1 \end{bmatrix}$$

Equation (A.3) displays Equation (A.1), for z_{0k}, when $k=2$.

$$z_{02} = \begin{pmatrix} v_{12} \\ v_{22} \\ v_{32} \\ v_{42} \end{pmatrix}^T \left(\begin{pmatrix} p_{12} \\ p_{22} \\ p_{32} \\ p_{42} \end{pmatrix} - \begin{pmatrix} x_{112} & x_{122} \\ x_{212} & x_{222} \\ x_{312} & x_{322} \\ x_{412} & x_{422} \end{pmatrix} \begin{pmatrix} e^{-\lambda_{12}} \\ e^{-\lambda_{22}} \end{pmatrix} \right) \left(\sum_{i=1}^{4} v_{i2} \right)^{-1} \tag{A.3}$$

Using Table A1, and substituting in the x_{ijk} values, $x_{112} = x_{122} = x_{212} = x_{322} = 0, x_{222} = x_{312} = x_{412} = x_{422} = 1$, allows for (A.3) to be simplified to (A.4).

$$z_{02} = \frac{v_{12}\left(p_{12}\right) + v_{22}\left(p_{22} - e^{-\lambda_{22}}\right) + v_{32}\left(p_{32} - e^{-\lambda_{12}}\right) + v_{42}\left(p_{42} - e^{-\lambda_{12}} - e^{-\lambda_{22}}\right)}{v_{12} + v_{22} + v_{32} + v_{42}} \tag{A.4}$$

A.2 EXAMPLE OF CRAMER'S RULE

$$A_k = \begin{bmatrix} \sum\limits_{i=1}^{2^k} v_{ik} & \sum\limits_{\substack{i=1 \\ x_{i1k}=1}}^{2^k} v_{ik} & \cdots & \sum\limits_{\substack{i=1 \\ x_{ikk}=1}}^{2^k} v_{ik} \\ \sum\limits_{\substack{i=1 \\ x_{i1k}=1}}^{2^k} v_{ik} & \sum\limits_{\substack{i=1 \\ x_{i1k}=x_{i1k}=1}}^{2^k} v_{ik} & \cdots & \sum\limits_{\substack{i=1 \\ x_{i1k}=x_{ikk}=1}}^{2^k} v_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum\limits_{\substack{i=1 \\ x_{ikk}=1}}^{2^k} v_{ik} & \sum\limits_{\substack{i=1 \\ x_{i1k}=x_{ikk}=1}}^{2^k} v_{ik} & \cdots & \sum\limits_{\substack{i=1 \\ x_{ikk}=x_{ikk}=1}}^{2^k} v_{ik} \end{bmatrix}, s_{\mathrm{k}} = \begin{bmatrix} z_{0k} \\ e^{-\lambda_{1k}} \\ \vdots \\ e^{-\lambda_{kk}} \end{bmatrix}, b_{\mathrm{k}} = \begin{bmatrix} \sum\limits_{i=1}^{2^k} v_{ik} p_{ik} \\ \sum\limits_{\substack{i=1 \\ x_{i1k}=1}}^{2^k} v_{ik} p_{ik} \\ \vdots \\ \sum\limits_{\substack{i=1 \\ x_{ikk}=1}}^{2^k} v_{ik} p_{ik} \end{bmatrix} \tag{A.5}$$

For $k=2$, we have $A_2 = \begin{bmatrix} v_{12}+v_{22}+v_{32}+v_{42} & v_{32}+v_{42} & v_{22}+v_{42} \\ v_{32}+v_{42} & v_{32}+v_{42} & v_{42} \\ v_{22}+v_{42} & v_{42} & v_{22}+v_{42} \end{bmatrix}$, $s_2 = \begin{bmatrix} z_{02} \\ e^{-\lambda_{12}} \\ e^{-\lambda_{22}} \end{bmatrix}$, and

$$b_2 = \begin{bmatrix} v_{12}p_{12}+v_{22}p_{22}+v_{32}p_{32}+v_{42}p_{42} \\ v_{32}p_{32}+v_{42}p_{42} \\ v_{22}p_{32}+v_{22}p_{42} \end{bmatrix}$$

Which when solved yields

$$s_2 = \frac{1}{\det(A_2)} \begin{bmatrix} p_{12}v_{12}\left(v_{22}v_{32}+v_{22}v_{42}+v_{32}v_{42}\right)+ \\ p_{22}v_{22}\left(v_{32}v_{42}\right)+p_{32}v_{32}\left(v_{22}v_{42}\right)-p_{42}v_{42}\left(v_{22}v_{32}\right) \\ p_{12}v_{12}\left(-v_{22}v_{32}-v_{32}v_{42}\right)+p_{22}v_{22}\left(-v_{12}v_{42}-v_{32}v_{42}\right) \\ +p_{32}v_{32}\left(v_{12}v_{22}+v_{12}v_{42}\right)+p_{42}v_{42}\left(v_{12}v_{22}+v_{22}v_{32}\right) \\ p_{12}v_{12}\left(-v_{22}v_{32}-v_{22}v_{42}\right)+p_{22}v_{22}\left(v_{12}v_{32}+v_{12}v_{42}\right) \\ +p_{32}v_{32}\left(-v_{12}v_{42}-v_{22}v_{42}\right)+p_{42}v_{42}\left(v_{12}v_{32}+v_{22}v_{32}\right) \end{bmatrix},$$

where $\det(A_2) = v_{12}v_{22}v_{32}+v_{12}v_{22}v_{42}+v_{12}v_{32}v_{42}+v_{22}v_{32}v_{42}$.

A.3 FULL LIST OF PARAMETERS GENERATED BY SUMER ON OP AND DO

Table A2. Rate Parameters Generated from SUMER for k=1-9 for DO

k	*k'*	*lag1*	*lag2*	*lag3*	*lag4*	*lag5*	*lag6*	*lag7*	*lag8*	*lag9*
2	**1**	0.9825	0.9825							
2	**2**	**0.8293**	**1.1570**							
3	**1**	1.3527	1.3527	1.3527						
3	**2**	0.8973	1.6717	1.6717						
3	**3**	**0.9208**	**1.3886**	**2.0055**						
4	**1**	1.6127	1.6127	1.6127	1.6127					
4	**2**	0.9094	2.0007	2.0007	2.0007					
4	**3**	**0.9591**	**1.4769**	**2.3349**	**2.3349**					
4	**4**	0.9600	1.4807	2.2708	2.3930					
5	**1**	1.8336	1.8336	1.8336	1.8336	1.8336				
5	**2**	0.8886	2.2873	2.2873	2.2873	2.2873				
5	**3**	0.9695	1.4903	2.6583	2.6583	2.6583				
5	**4**	0.9741	1.5081	2.3829	2.7883	2.7883				
5	**5**	**0.9746**	**1.5103**	**2.3868**	**2.6324**	**2.9524**				
6	**1**	2.0115	2.0115	2.0115	2.0115	2.0115	2.0115			
6	**2**	0.8564	2.5218	2.5218	2.5218	2.5218	2.5218			
6	**3**	0.9675	1.4781	2.9221	2.9221	2.9221	2.9221			
6	**4**	0.9779	1.5149	2.3970	3.0974	3.0974	3.0974			

k	k'	lag1	lag2	lag3	lag4	lag5	lag6	lag7	lag8	lag9
6	5	**0.9799**	**1.5216**	**2.4175**	**2.7245**	**3.3004**	**3.3004**			
6	6	0.9800	1.5219	2.4197	2.7280	3.2308	3.3652			
7	1	2.1765	2.1765	2.1765	2.1765	2.1765	2.1765	2.1765		
7	2	0.8223	2.7239	2.7239	2.7239	2.7239	2.7239	2.7239		
7	3	0.9615	1.4611	3.1408	3.1408	3.1408	3.1408	3.1408		
7	4	0.9781	1.5162	2.3912	3.3353	3.3353	3.3353	3.3353		
7	5	**0.9824**	**1.5282**	**2.4323**	**2.7623**	**3.5426**	**3.5426**	**3.5426**		
7	6	0.9828	1.5290	2.4370	2.7720	3.4037	3.6014	3.6014		
7	7	0.9826	1.5287	2.4358	2.7670	3.3896	3.8173	3.4489		
8	1	2.3317	2.3317	2.3317	2.3317	2.3317	2.3317	2.3317	2.3317	
8	2	0.7920	2.9061	2.9061	2.9061	2.9061	2.9061	2.9061	2.9061	
8	3	0.9556	1.4415	3.3230	3.3230	3.3230	3.3230	3.3230	3.3230	
8	4	0.9778	1.5129	2.3741	3.5289	3.5289	3.5289	3.5289	3.5289	
8	5	**0.9840**	**1.5303**	**2.4345**	**2.7700**	**3.7400**	**3.7400**	**3.7400**	**3.7400**	
8	6	0.9849	1.5321	2.4446	2.7923	3.4581	3.8187	3.8187	3.8187	
8	7	0.9847	1.5318	2.4434	2.7875	3.4408	4.0068	3.7549	3.7549	
8	8	0.9847	1.5318	2.4435	2.7877	3.4426	4.0132	3.7214	3.7833	
9	1	2.4834	2.4834	2.4834	2.4834	2.4834	2.4834	2.4834	2.4834	2.4834
9	2	0.7523	3.1097	3.1097	3.1097	3.1097	3.1097	3.1097	3.1097	3.1097
9	3	0.9420	1.4124	3.5266	3.5266	3.5266	3.5266	3.5266	3.5266	3.5266
9	4	0.9712	1.5024	2.3309	3.7432	3.7432	3.7432	3.7432	3.7432	3.7432
9	5	0.9808	1.5273	2.4173	2.7175	3.9942	3.9942	3.9942	3.9942	3.9942
9	6	**0.9830**	**1.5314**	**2.4397**	**2.7679**	**3.3881**	**4.1380**	**4.1380**	**4.1380**	**4.1380**
9	7	0.9834	1.5319	2.4417	2.7755	3.4141	3.8783	4.2097	4.2097	4.2097
9	8	0.9837	1.5324	2.4433	2.7810	3.4421	3.9847	3.6761	4.5066	4.5066
9	9	0.9847	1.5320	2.4438	2.7878	3.4449	4.0191	3.7294	3.8043	6.5804

Table A3. Coefficients Generated from SUMER for k=1-9 for DO

k	k'	Constant	lag1	lag2	lag3	lag4	lag5	lag6	lag7	lag8	lag9
2	1	0.0415	0.3744	0.3744							
2	2	**0.0423**	**0.4363**	**0.3144**							
3	1	0.0280	0.2585	0.2585	0.2585						
3	2	0.0301	0.4077	0.1879	0.1879						
3	3	**0.0311**	**0.3982**	**0.2494**	**0.1346**						
4	1	0.0178	0.1993	0.1993	0.1993	0.1993					
4	2	0.0215	0.4028	0.1352	0.1352	0.1352					
4	3	**0.0233**	**0.3833**	**0.2283**	**0.0968**	**0.0968**					
4	4	0.0234	0.3829	0.2275	0.1032	0.0914					
5	1	0.0090	0.1598	0.1598	0.1598	0.1598	0.1598				
5	2	0.0149	0.4112	0.1015	0.1015	0.1015	0.1015				
5	3	0.0178	0.3793	0.2253	0.0701	0.0701	0.0701				
5	4	0.0182	0.3775	0.2213	0.0923	0.0615	0.0615				
5	5	**0.0185**	**0.3773**	**0.2208**	**0.0919**	**0.0719**	**0.0522**				
6	1	0.0026	0.1338	0.1338	0.1338	0.1338	0.1338	0.1338			
6	2	0.0103	0.4247	0.0803	0.0803	0.0803	0.0803	0.0803			
6	3	0.0140	0.3800	0.2281	0.0538	0.0538	0.0538	0.0538			
6	4	0.0147	0.3761	0.2198	0.0910	0.0452	0.0452	0.0452			
6	5	**0.0153**	**0.3753**	**0.2184**	**0.0891**	**0.0656**	**0.0369**	**0.0369**			
6	6	0.0154	0.3753	0.2183	0.0889	0.0653	0.0395	0.0346			
7	1	0.0000	0.1134	0.1134	0.1134	0.1134	0.1134	0.1134	0.1134		
7	2	0.0046	0.4394	0.0656	0.0656	0.0656	0.0656	0.0656	0.0656		
7	3	0.0095	0.3823	0.2320	0.0432	0.0432	0.0432	0.0432	0.0432		
7	4	0.0107	0.3760	0.2195	0.0915	0.0356	0.0356	0.0356	0.0356		

k	*k'*	*Constant*	*lag1*	*lag2*	*lag3*	*lag4*	*lag5*	*lag6*	*lag7*	*lag8*	*lag9*
7	**5**	**0.0116**	**0.3744**	**0.2169**	**0.0878**	**0.0631**	**0.0289**	**0.0289**	**0.0289**		
7	**6**	0.0117	0.3743	0.2168	0.0874	0.0625	0.0333	0.0273	0.0273		
7	**7**	0.0115	0.3744	0.2168	0.0875	0.0629	0.0337	0.0220	0.0318		
8	**1**	0.0000	0.0971	0.0971	0.0971	0.0971	0.0971	0.0971	0.0971	0.0971	
8	**2**	0.0000	0.4529	0.0547	0.0547	0.0547	0.0547	0.0547	0.0547	0.0547	
8	**3**	0.0039	0.3846	0.2366	0.0360	0.0360	0.0360	0.0360	0.0360	0.0360	
8	**4**	0.0057	0.3761	0.2203	0.0931	0.0293	0.0293	0.0293	0.0293	0.0293	
8	**5**	**0.0073**	**0.3738**	**0.2165**	**0.0876**	**0.0627**	**0.0238**	**0.0238**	**0.0238**	**0.0238**	
8	**6**	0.0076	0.3735	0.2161	0.0868	0.0613	0.0315	0.0220	0.0220	0.0220	
8	**7**	0.0073	0.3735	0.2162	0.0869	0.0616	0.0320	0.0182	0.0234	0.0234	
8	**8**	0.0074	0.3736	0.2161	0.0869	0.0616	0.0320	0.0181	0.0242	0.0227	
9	**1**	0.0000	0.0835	0.0835	0.0835	0.0835	0.0835	0.0835	0.0835	0.0835	0.0835
9	**2**	0.0000	0.4713	0.0446	0.0446	0.0446	0.0446	0.0446	0.0446	0.0446	0.0446
9	**3**	0.0000	0.3898	0.2436	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294
9	**4**	0.0009	0.3786	0.2226	0.0972	0.0237	0.0237	0.0237	0.0237	0.0237	0.0237
9	**5**	0.0036	0.3750	0.2171	0.0892	0.0660	0.0184	0.0184	0.0184	0.0184	0.0184
9	**6**	**0.0046**	**0.3742**	**0.2162**	**0.0872**	**0.0628**	**0.0338**	**0.0160**	**0.0160**	**0.0160**	**0.0160**
9	**7**	0.0050	0.3740	0.2161	0.0870	0.0623	0.0329	0.0207	0.0149	0.0149	0.0149
9	**8**	0.0059	0.3739	0.2160	0.0869	0.0620	0.0320	0.0186	0.0253	0.0110	0.0110
9	**9**	0.0070	0.3736	0.2161	0.0868	0.0616	0.0319	0.0180	0.0240	0.0223	0.0014

Table A4. Rate Parameters generated from SUMER for k=1-14 for OP

k	*k'*	*lag1*	*lag2*	*lag3*	*lag4*	*lag5*	*lag6*	*lag7*	*lag8*	*lag9*	*lag10*	*lag11*	*lag12*	*lag13*	*lag14*
2	**1**	1.922	1.922												
2	**2**	**1.609**	**2.381**												
3	**1**	2.165	2.165	2.165											
3	**2**	1.642	2.572	2.572											
3	**3**	**1.643**	**2.552**	**2.592**											
4	**1**	2.339	2.339	2.339	2.339										
4	**2**	1.664	2.709	2.709	2.709										
4	**3**	**1.667**	**2.616**	**2.756**	**2.756**										
4	**4**	1.667	2.615	2.765	2.747										
5	**1**	2.476	2.476	2.476	2.476	2.476									
5	**2**	1.680	2.819	2.819	2.819	2.819									
5	**3**	1.686	2.661	2.873	2.873	2.873									
5	**4**	1.687	2.664	2.832	2.893	2.893									
5	**5**	**1.687**	**2.664**	**2.830**	**2.920**	**2.867**									
6	**1**	2.586	2.586	2.586	2.586	2.586	2.586								
6	**2**	1.695	2.907	2.907	2.907	2.907	2.907								
6	**3**	1.703	2.699	2.960	2.960	2.960	2.960								
6	**4**	**1.703**	**2.706**	**2.885**	**2.985**	**2.985**	**2.985**								
6	**5**	1.703	2.706	2.884	2.995	2.980	2.980								
6	**6**	1.703	2.706	2.884	2.989	3.048	2.917								
7	**1**	2.682	2.682	2.682	2.682	2.682	2.682	2.682							
7	**2**	1.707	2.986	2.986	2.986	2.986	2.986	2.986							
7	**3**	1.717	2.731	3.040	3.040	3.040	3.040	3.040							
7	**4**	1.717	2.740	2.925	3.068	3.068	3.068	3.068							
7	**5**	1.717	2.741	2.927	3.047	3.075	3.075	3.075							
7	**6**	1.717	2.740	2.927	3.043	3.117	3.055	3.055							
7	**7**	**1.717**	**2.741**	**2.927**	**3.043**	**3.115**	**3.084**	**3.028**							
8	**1**	2.768	2.768	2.768	2.768	2.768	2.768	2.768	2.768						
8	**2**	1.718	3.058	3.058	3.058	3.058	3.058	3.058	3.058						
8	**3**	1.729	2.758	3.112	3.112	3.112	3.112	3.112	3.112						
8	**4**	1.729	2.771	2.961	3.141	3.141	3.141	3.141	3.141						
8	**5**	**1.729**	**2.772**	**2.965**	**3.089**	**3.154**	**3.154**	**3.154**	**3.154**						
8	**6**	1.729	2.772	2.965	3.088	3.168	3.149	3.149	3.149						
8	**7**	1.729	2.772	2.965	3.088	3.168	3.150	3.149	3.149						
8	**8**	1.729	2.772	2.965	3.088	3.168	3.146	3.198	3.104						
9	**1**	2.843	2.843	2.843	2.843	2.843	2.843	2.843	2.843	2.843					

k	***k’***	***lag1***	***lag2***	***lag3***	***lag4***	***lag5***	***lag6***	***lag7***	***lag8***	***lag9***	***lag10***	***lag11***	***lag12***	***lag13***	***lag14***
9	**2**	1.727	3.123	3.123	3.123	3.123	3.123	3.123	3.123	3.123					
9	**3**	1.739	2.784	3.176	3.176	3.176	3.176	3.176	3.176	3.176					
9	**4**	1.740	2.799	2.993	3.206	3.206	3.206	3.206	3.206	3.206					
9	**5**	**1.740**	**2.800**	**3.000**	**3.127**	**3.220**	**3.220**	**3.220**	**3.220**	**3.220**					
9	**6**	1.740	2.800	3.000	3.128	3.213	3.222	3.222	3.222	3.222					
9	**7**	1.740	2.800	3.000	3.128	3.215	3.200	3.229	3.229	3.229					
9	**8**	1.740	2.800	3.000	3.128	3.214	3.197	3.268	3.211	3.211					
9	**9**	1.740	2.800	3.001	3.128	3.215	3.197	3.262	3.276	3.152					
10	**1**	2.914	2.914	2.914	2.914	2.914	2.914	2.914	2.914	2.914	2.914				
10	**2**	1.734	3.186	3.186	3.186	3.186	3.186	3.186	3.186	3.186	3.186				
10	**3**	1.748	2.804	3.239	3.239	3.239	3.239	3.239	3.239	3.239	3.239				
10	**4**	1.749	2.822	3.018	3.271	3.271	3.271	3.271	3.271	3.271	3.271				
10	**5**	1.749	2.824	3.029	3.158	3.288	3.288	3.288	3.288	3.288	3.288				
10	**6**	1.749	2.824	3.029	3.162	3.246	3.296	3.296	3.296	3.296	3.296				
10	**7**	**1.749**	**2.824**	**3.029**	**3.163**	**3.252**	**3.236**	**3.310**	**3.310**	**3.310**	**3.310**				
10	**8**	1.749	2.824	3.029	3.163	3.252	3.236	3.313	3.309	3.309	3.309				
10	**9**	1.749	2.824	3.029	3.163	3.252	3.236	3.310	3.334	3.298	3.298				
10	**10**	1.749	2.824	3.029	3.163	3.252	3.236	3.310	3.334	3.305	3.291				
11	**1**	2.980	2.980	2.980	2.980	2.980	2.980	2.980	2.980	2.980	2.980	2.980			
11	**2**	1.740	3.246	3.246	3.246	3.246	3.246	3.246	3.246	3.246	3.246	3.246			
11	**3**	1.755	2.820	3.299	3.299	3.299	3.299	3.299	3.299	3.299	3.299	3.299			
11	**4**	1.756	2.842	3.040	3.332	3.332	3.332	3.332	3.332	3.332	3.332	3.332			
11	**5**	1.756	2.844	3.053	3.183	3.353	3.353	3.353	3.353	3.353	3.353	3.353			
11	**6**	1.756	2.844	3.054	3.191	3.275	3.364	3.364	3.364	3.364	3.364	3.364			
11	**7**	1.756	2.844	3.055	3.192	3.285	3.266	3.383	3.383	3.383	3.383	3.383			
11	**8**	**1.756**	**2.844**	**3.055**	**3.192**	**3.286**	**3.270**	**3.347**	**3.392**	**3.392**	**3.392**	**3.392**			
11	**9**	1.756	2.844	3.055	3.192	3.286	3.270	3.348	3.376	3.396	3.396	3.396			
11	**10**	1.756	2.844	3.055	3.192	3.286	3.270	3.348	3.379	3.361	3.414	3.414			
11	**11**	1.756	2.844	3.055	3.192	3.286	3.270	3.348	3.379	3.358	3.451	3.378			
12	**1**	3.040	3.040	3.040	3.040	3.040	3.040	3.040	3.040	3.040	3.040	3.040	3.040		
12	**2**	1.745	3.300	3.300	3.300	3.300	3.300	3.300	3.300	3.300	3.300	3.300	3.300		
12	**3**	1.761	2.836	3.353	3.353	3.353	3.353	3.353	3.353	3.353	3.353	3.353	3.353		
12	**4**	1.762	2.860	3.059	3.387	3.387	3.387	3.387	3.387	3.387	3.387	3.387	3.387		
12	**5**	1.762	2.862	3.076	3.207	3.409	3.409	3.409	3.409	3.409	3.409	3.409	3.409		
12	**6**	1.763	2.863	3.077	3.217	3.303	3.423	3.423	3.423	3.423	3.423	3.423	3.423		
12	**7**	1.763	2.863	3.078	3.219	3.317	3.295	3.443	3.443	3.443	3.443	3.443	3.443		
12	**8**	1.763	2.863	3.078	3.219	3.317	3.301	3.378	3.455	3.455	3.455	3.455	3.455		
12	**9**	1.763	2.863	3.078	3.219	3.318	3.302	3.383	3.412	3.465	3.465	3.465	3.465		
12	**10**	1.763	2.863	3.078	3.219	3.318	3.302	3.383	3.418	3.406	3.484	3.484	3.484		
12	**11**	1.763	2.863	3.078	3.219	3.318	3.302	3.383	3.418	3.403	3.517	3.468	3.468		
12	**12**	**1.763**	**2.863**	**3.078**	**3.219**	**3.318**	**3.302**	**3.383**	**3.418**	**3.403**	**3.511**	**3.549**	**3.395**		
13	**1**	3.096	3.096	3.096	3.096	3.096	3.096	3.096	3.096	3.096	3.096	3.096	3.096	3.096	
13	**2**	1.749	3.352	3.352	3.352	3.352	3.352	3.352	3.352	3.352	3.352	3.352	3.352	3.352	
13	**3**	1.766	2.846	3.406	3.406	3.406	3.406	3.406	3.406	3.406	3.406	3.406	3.406	3.406	
13	**4**	1.767	2.874	3.075	3.440	3.440	3.440	3.440	3.440	3.440	3.440	3.440	3.440	3.440	
13	**5**	1.768	2.876	3.095	3.226	3.464	3.464	3.464	3.464	3.464	3.464	3.464	3.464	3.464	
13	**6**	1.768	2.877	3.097	3.239	3.325	3.480	3.480	3.480	3.480	3.480	3.480	3.480	3.480	
13	**7**	1.768	2.877	3.098	3.241	3.342	3.318	3.503	3.503	3.503	3.503	3.503	3.503	3.503	
13	**8**	1.768	2.878	3.098	3.242	3.343	3.327	3.405	3.518	3.518	3.518	3.518	3.518	3.518	
13	**9**	1.768	2.878	3.098	3.242	3.343	3.328	3.413	3.441	3.532	3.532	3.532	3.532	3.532	
13	**10**	**1.768**	**2.878**	**3.098**	**3.242**	**3.343**	**3.328**	**3.414**	**3.450**	**3.438**	**3.555**	**3.555**	**3.555**	**3.555**	
13	**11**	1.768	2.878	3.098	3.242	3.343	3.328	3.414	3.450	3.438	3.560	3.553	3.553	3.553	
13	**12**	1.768	2.878	3.098	3.242	3.343	3.328	3.414	3.450	3.437	3.555	3.608	3.529	3.529	
13	**13**	1.768	2.878	3.098	3.242	3.343	3.328	3.414	3.450	3.437	3.555	3.606	3.546	3.512	
14	**1**	3.150	3.150	3.150	3.150	3.150	3.150	3.150	3.150	3.150	3.150	3.150	3.150	3.150	3.150
14	**2**	1.752	3.402	3.402	3.402	3.402	3.402	3.402	3.402	3.402	3.402	3.402	3.402	3.402	3.402
14	**3**	1.770	2.856	3.456	3.456	3.456	3.456	3.456	3.456	3.456	3.456	3.456	3.456	3.456	3.456
14	**4**	1.772	2.886	3.087	3.492	3.492	3.492	3.492	3.492	3.492	3.492	3.492	3.492	3.492	3.492
14	**5**	1.772	2.889	3.110	3.241	3.517	3.517	3.517	3.517	3.517	3.517	3.517	3.517	3.517	3.517
14	**6**	1.773	2.890	3.112	3.258	3.342	3.536	3.536	3.536	3.536	3.536	3.536	3.536	3.536	3.536
14	**7**	1.773	2.890	3.113	3.260	3.363	3.336	3.561	3.561	3.561	3.561	3.561	3.561	3.561	3.561
14	**8**	1.773	2.891	3.114	3.261	3.364	3.349	3.426	3.579	3.579	3.579	3.579	3.579	3.579	3.579
14	**9**	1.773	2.891	3.114	3.262	3.365	3.350	3.438	3.464	3.597	3.597	3.597	3.597	3.597	3.597
14	**10**	1.773	2.891	3.114	3.262	3.365	3.351	3.439	3.478	3.463	3.623	3.623	3.623	3.623	3.623
14	**11**	**1.773**	**2.891**	**3.114**	**3.262**	**3.365**	**3.351**	**3.439**	**3.478**	**3.466**	**3.591**	**3.631**	**3.631**	**3.631**	**3.631**

k	*k'*	*lag1*	*lag2*	*lag3*	*lag4*	*lag5*	*lag6*	*lag7*	*lag8*	*lag9*	*lag10*	*lag11*	*lag12*	*lag13*	*lag14*
14	**12**	1.773	2.891	3.114	3.262	3.365	3.351	3.439	3.478	3.466	3.589	3.647	3.626	3.626	3.626
14	**13**	1.773	2.891	3.114	3.262	3.365	3.351	3.439	3.478	3.466	3.589	3.650	3.600	3.638	3.638
14	**14**	1.773	2.891	3.114	3.262	3.365	3.351	3.439	3.478	3.466	3.589	3.650	3.598	3.666	3.611

Table A5. Coefficients Generated from SUMER for k=1-14 for OP

k	*k'*	*Constant*	*lag1*	*lag2*	*lag3*	*lag4*	*lag5*	*lag6*	*lag7*	*lag8*	*lag9*	*lag10*	*lag11*	*lag12*	*lag13*	*lag14*
2	**1**	0.063	0.146	0.146												
2	**2**	**0.063**	**0.200**	**0.092**												
3	**1**	0.058	0.115	0.115	0.115											
3	**2**	0.058	0.194	0.076	0.076											
3	**3**	**0.058**	**0.193**	**0.078**	**0.075**											
4	**1**	0.055	0.096	0.096	0.096	0.096										
4	**2**	0.054	0.189	0.067	0.067	0.067										
4	**3**	**0.054**	**0.189**	**0.073**	**0.064**	**0.064**										
4	**4**	0.054	0.189	0.073	0.063	0.064										
5	**1**	0.052	0.084	0.084	0.084	0.084	0.084									
5	**2**	0.051	0.186	0.060	0.060	0.060	0.060									
5	**3**	0.051	0.185	0.070	0.057	0.057	0.057									
5	**4**	0.051	0.185	0.070	0.059	0.055	0.055									
5	**5**	**0.051**	**0.185**	**0.070**	**0.059**	**0.054**	**0.057**									
6	**1**	0.049	0.075	0.075	0.075	0.075	0.075	0.075								
6	**2**	0.048	0.184	0.055	0.055	0.055	0.055	0.055								
6	**3**	0.048	0.182	0.067	0.052	0.052	0.052	0.052								
6	**4**	**0.048**	**0.182**	**0.067**	**0.056**	**0.051**	**0.051**	**0.051**								
6	**5**	0.048	0.182	0.067	0.056	0.050	0.051	0.051								
6	**6**	0.048	0.182	0.067	0.056	0.050	0.047	0.054								
7	**1**	0.047	0.068	0.068	0.068	0.068	0.068	0.068	0.068							
7	**2**	0.046	0.181	0.050	0.050	0.050	0.050	0.050	0.050							
7	**3**	0.046	0.180	0.065	0.048	0.048	0.048	0.048	0.048							
7	**4**	0.046	0.180	0.065	0.054	0.047	0.047	0.047	0.047							
7	**5**	0.046	0.180	0.065	0.054	0.048	0.046	0.046	0.046							
7	**6**	0.046	0.180	0.065	0.054	0.048	0.044	0.047	0.047							
7	**7**	**0.046**	**0.180**	**0.065**	**0.054**	**0.048**	**0.044**	**0.046**	**0.048**							
8	**1**	0.044	0.063	0.063	0.063	0.063	0.063	0.063	0.063	0.063						
8	**2**	0.044	0.179	0.047	0.047	0.047	0.047	0.047	0.047	0.047						
8	**3**	0.044	0.178	0.063	0.045	0.045	0.045	0.045	0.045	0.045						
8	**4**	0.044	0.177	0.063	0.052	0.043	0.043	0.043	0.043	0.043						
8	**5**	**0.044**	**0.177**	**0.063**	**0.052**	**0.046**	**0.043**	**0.043**	**0.043**	**0.043**						
8	**6**	0.044	0.177	0.063	0.052	0.046	0.042	0.043	0.043	0.043						
8	**7**	0.044	0.177	0.063	0.052	0.046	0.042	0.043	0.043	0.043						
8	**8**	0.044	0.177	0.063	0.052	0.046	0.042	0.043	0.041	0.045						
9	**1**	0.043	0.058	0.058	0.058	0.058	0.058	0.058	0.058	0.058	0.058					
9	**2**	0.042	0.178	0.044	0.044	0.044	0.044	0.044	0.044	0.044	0.044					
9	**3**	0.042	0.176	0.062	0.042	0.042	0.042	0.042	0.042	0.042	0.042					
9	**4**	0.042	0.176	0.061	0.050	0.041	0.041	0.041	0.041	0.041	0.041					
9	**5**	**0.042**	**0.176**	**0.061**	**0.050**	**0.044**	**0.040**	**0.040**	**0.040**	**0.040**	**0.040**					
9	**6**	0.042	0.176	0.061	0.050	0.044	0.040	0.040	0.040	0.040	0.040					
9	**7**	0.042	0.176	0.061	0.050	0.044	0.040	0.041	0.040	0.040	0.040					
9	**8**	0.042	0.176	0.061	0.050	0.044	0.040	0.041	0.038	0.040	0.040					
9	**9**	0.042	0.176	0.061	0.050	0.044	0.040	0.041	0.038	0.038	0.043					
10	**1**	0.041	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.054	0.054				
10	**2**	0.040	0.177	0.041	0.041	0.041	0.041	0.041	0.041	0.041	0.041	0.041				
10	**3**	0.040	0.174	0.061	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039				
10	**4**	0.040	0.174	0.059	0.049	0.038	0.038	0.038	0.038	0.038	0.038	0.038				
10	**5**	0.040	0.174	0.059	0.048	0.043	0.037	0.037	0.037	0.037	0.037	0.037				
10	**6**	0.040	0.174	0.059	0.048	0.042	0.039	0.037	0.037	0.037	0.037	0.037				
10	**7**	**0.040**	**0.174**	**0.059**	**0.048**	**0.042**	**0.039**	**0.039**	**0.037**	**0.037**	**0.037**	**0.037**				
10	**8**	0.040	0.174	0.059	0.048	0.042	0.039	0.039	0.036	0.037	0.037	0.037				
10	**9**	0.040	0.174	0.059	0.048	0.042	0.039	0.039	0.037	0.036	0.037	0.037				
10	**10**	0.040	0.174	0.059	0.048	0.042	0.039	0.039	0.037	0.036	0.037	0.037				
11	**1**	0.040	0.051	0.051	0.051	0.051	0.051	0.051	0.051	0.051	0.051	0.051	0.051			
11	**2**	0.039	0.176	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039	0.039			
11	**3**	0.039	0.173	0.060	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037			
11	**4**	0.039	0.173	0.058	0.048	0.036	0.036	0.036	0.036	0.036	0.036	0.036	0.036			
11	**5**	0.039	0.173	0.058	0.047	0.041	0.035	0.035	0.035	0.035	0.035	0.035	0.035			
11	**6**	0.039	0.173	0.058	0.047	0.041	0.038	0.035	0.035	0.035	0.035	0.035	0.035			
11	**7**	0.039	0.173	0.058	0.047	0.041	0.037	0.038	0.034	0.034	0.034	0.034	0.034			
11	**8**	**0.039**	**0.173**	**0.058**	**0.047**	**0.041**	**0.037**	**0.038**	**0.035**	**0.034**	**0.034**	**0.034**	**0.034**			
11	**9**	0.039	0.173	0.058	0.047	0.041	0.037	0.038	0.035	0.034	0.033	0.033	0.033			
11	**10**	0.039	0.173	0.058	0.047	0.041	0.037	0.038	0.035	0.034	0.035	0.033	0.033			
11	**11**	0.039	0.173	0.058	0.047	0.041	0.037	0.038	0.035	0.034	0.035	0.032	0.034			
12	**1**	0.038	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048	0.048		
12	**2**	0.038	0.175	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037	0.037		
12	**3**	0.038	0.172	0.059	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035		
12	**4**	0.038	0.172	0.057	0.047	0.034	0.034	0.034	0.034	0.034	0.034	0.034	0.034	0.034		
12	**5**	0.038	0.172	0.057	0.046	0.040	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033		
12	**6**	0.038	0.172	0.057	0.046	0.040	0.037	0.033	0.033	0.033	0.033	0.033	0.033	0.033		
12	**7**	0.038	0.172	0.057	0.046	0.040	0.036	0.037	0.032	0.032	0.032	0.032	0.032	0.032		

k	***k'***	***Constant***	***lag1***	***lag2***	***lag3***	***lag4***	***lag5***	***lag6***	***lag7***	***lag8***	***lag9***	***lag10***	***lag11***	***lag12***	***lag13***	***lag14***
12	**8**	0.038	0.172	0.057	0.046	0.040	0.036	0.037	0.034	0.032	0.032	0.032	0.032	0.032		
12	**9**	0.038	0.172	0.057	0.046	0.040	0.036	0.037	0.034	0.033	0.031	0.031	0.031	0.031		
12	**10**	0.038	0.172	0.057	0.046	0.040	0.036	0.037	0.034	0.033	0.033	0.031	0.031	0.031		
12	**11**	0.038	0.172	0.057	0.046	0.040	0.036	0.037	0.034	0.033	0.033	0.030	0.031	0.031		
12	**12**	**0.038**	**0.172**	**0.057**	**0.046**	**0.040**	**0.036**	**0.037**	**0.034**	**0.033**	**0.033**	**0.030**	**0.029**	**0.034**		
13	**1**	0.037	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	0.045	
13	**2**	0.036	0.174	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	
13	**3**	0.036	0.171	0.058	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	
13	**4**	0.036	0.171	0.056	0.046	0.032	0.032	0.032	0.032	0.032	0.032	0.032	0.032	0.032	0.032	
13	**5**	0.036	0.171	0.056	0.045	0.040	0.031	0.031	0.031	0.031	0.031	0.031	0.031	0.031	0.031	
13	**6**	0.037	0.171	0.056	0.045	0.039	0.036	0.031	0.031	0.031	0.031	0.031	0.031	0.031	0.031	
13	**7**	0.037	0.171	0.056	0.045	0.039	0.035	0.036	0.030	0.030	0.030	0.030	0.030	0.030	0.030	
13	**8**	0.037	0.171	0.056	0.045	0.039	0.035	0.036	0.033	0.030	0.030	0.030	0.030	0.030	0.030	
13	**9**	0.037	0.171	0.056	0.045	0.039	0.035	0.036	0.033	0.032	0.029	0.029	0.029	0.029	0.029	
13	**10**	**0.037**	**0.171**	**0.056**	**0.045**	**0.039**	**0.035**	**0.036**	**0.033**	**0.032**	**0.032**	**0.029**	**0.029**	**0.029**	**0.029**	
13	**11**	0.037	0.171	0.056	0.045	0.039	0.035	0.036	0.033	0.032	0.032	0.028	0.029	0.029	0.029	
13	**12**	0.037	0.171	0.056	0.045	0.039	0.035	0.036	0.033	0.032	0.032	0.029	0.027	0.029	0.029	
13	**13**	0.037	0.171	0.056	0.045	0.039	0.035	0.036	0.033	0.032	0.032	0.029	0.027	0.029	0.030	
14	**1**	0.036	0.043	0.043	0.043	0.043	0.043	0.043	0.043	0.043	0.043	0.043	0.043	0.043	0.043	0.043
14	**2**	0.035	0.173	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033
14	**3**	0.035	0.170	0.058	0.032	0.032	0.032	0.032	0.032	0.032	0.032	0.032	0.032	0.032	0.032	0.032
14	**4**	0.035	0.170	0.056	0.046	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030
14	**5**	0.035	0.170	0.056	0.045	0.039	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030	0.030
14	**6**	0.035	0.170	0.056	0.045	0.038	0.035	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029
14	**7**	0.035	0.170	0.056	0.044	0.038	0.035	0.036	0.028	0.028	0.028	0.028	0.028	0.028	0.028	0.028
14	**8**	0.035	0.170	0.056	0.044	0.038	0.035	0.035	0.033	0.028	0.028	0.028	0.028	0.028	0.028	0.028
14	**9**	0.035	0.170	0.056	0.044	0.038	0.035	0.035	0.032	0.031	0.027	0.027	0.027	0.027	0.027	0.027
14	**10**	0.036	0.170	0.056	0.044	0.038	0.035	0.035	0.032	0.031	0.031	0.027	0.027	0.027	0.027	0.027
14	**11**	**0.036**	**0.170**	**0.056**	**0.044**	**0.038**	**0.035**	**0.035**	**0.032**	**0.031**	**0.031**	**0.028**	**0.026**	**0.026**	**0.026**	**0.026**
14	**12**	0.036	0.170	0.056	0.044	0.038	0.035	0.035	0.032	0.031	0.031	0.028	0.026	0.027	0.027	0.027
14	**13**	0.036	0.170	0.056	0.044	0.038	0.035	0.035	0.032	0.031	0.031	0.028	0.026	0.027	0.026	0.026
14	**14**	0.036	0.170	0.056	0.044	0.038	0.035	0.035	0.032	0.031	0.031	0.028	0.026	0.027	0.026	0.027

APPENDIX B

OVERBOOKING MODEL APPENDIX

B.1 PROOFS

PROPOSITION 1. *A clinic schedule that fills all available appointment slots in a day before overbooking, has greater schedule reward than one that overbooks when an open slot is available.*

PROOF. See LaGanga and Lawrence (2012) Appendix 3 Page 1.

PROPOSITION 2. *A clinic day with N+1 appointment requests and N appointment slots achieves a maximal reward when the additional patient is overbooked in slot j*according to the following rules:*

$$\begin{array}{llll} (\mathbf{i}) & if \quad \dfrac{\sigma}{\omega} > \dfrac{p}{(1-p)} & then & j^* = 1 \\ (\mathbf{ii}) & if \quad \dfrac{\sigma}{\omega} < \dfrac{p}{(1-p)} & then & j^* = N \\ (\mathbf{iii}) & if \quad \dfrac{\sigma}{\omega} = \dfrac{p}{(1-p)} & then & j^* = any\ j \in \{1, \ldots, N\} \end{array} \tag{B.1}$$

PROOF. See LaGanga and Lawrence (2012) Appendix 3 Page 4.

PROPOSITION 3. *In a clinic day with N+1 appointment requests and N appointment slots, overbooking the additional patient in slot j* results in increased net reward, according to the following rules:*

$$
\begin{array}{llll}
\textbf{(i)} & j^* = 1, & if & \pi \ge p\left(\omega\left(\dfrac{1-p^N}{1-p}\right) + \sigma p^{N-1}\right) \\
\textbf{(ii)} & j^* = N, & if & \pi \ge p(\omega + \sigma) \\
\textbf{(iii)} & j^* = any\ j \in \{1,\ldots,N\}, & if & \pi \ge p\left(\dfrac{\omega}{1-p}\right)
\end{array}
\tag{B.2}
$$

PROOF. See LaGanga and Lawrence (2012) Appendix 3 Page 6.

PROPOSITION 4. *Let i denote a day with N booked appointments, for which an additional patient requests an appointment. Let d represent a day in the future of the scheduling horizon with appointment availability, and let $r = \alpha^{d-i}\beta^{d-i-1}\theta$. The clinic achieves a maximal reward by overbooking the additional patient in slot j* on day i, according to the following rules:*

$$
\begin{array}{llll}
\textbf{(i)} & j^* = 1, & if & \pi \ge \dfrac{p\left(\omega\left(\dfrac{1-p^N}{1-p}\right) + \sigma p^{N-1}\right) - \dfrac{\delta}{p}(d-i)}{(1-r)} \\
\textbf{(ii)} & j^* = N, & if & \pi \ge \dfrac{p(\omega + \sigma) - \dfrac{\delta}{p}(d-i)}{(1-r)} \\
\textbf{(iii)} & j^* = any\ j \in \{1,\ldots,N\}, & if & \pi \ge \dfrac{p\left(\dfrac{\omega}{1-p}\right) - \dfrac{\delta}{p}(d-i)}{(1-r)}
\end{array}
\tag{B.3}
$$

Otherwise, the patient should be booked into any open slot on day d.

PROOF. Assume that all patients booked on day *i* are from the same batch, and that day *d* is not fully booked. The two cases to consider are overbooking the additional patient on day *i* or booking the patient in an empty slot on day *d*. The proof proceeds by comparing the marginal benefit of the two cases, to determine on which day the patient should be overbooked. If the patient is overbooked on day *i*, we show which slot is optimal. If the patient is booked on day *d*, the patient should be booked in any available slot.

From Proposition 2, if the patient is overbooked on day *i* and shows for the appointment, the marginal change in day *i's* expected service reward is given by

$$
\pi p - \omega p^2\left(\frac{1-p^{N-j+1}}{1-p}\right) - \sigma p^{N-j+2} \tag{B.4}
$$

If the additional patient is overbooked on day *d*, the probability that the appointment will be completed is $p\alpha^{d-i}\beta^{d-i-1}\theta$. For succinctness, let $r=\alpha^{d-i}\beta^{d-i-1}\theta$. The waiting and overtime costs are not affected, because day *d* is not fully booked. With a service benefit of *π* and an indirect waiting penalty of *δ*(*d*-*i*), the marginal change in day *d's* expected service reward is given by

$$\pi r p-\delta\left(d-i\right) \tag{B.5}$$

To determine which case is optimal, we examine the slope of the difference between the two expected service rewards. If we subtract (B.4) from (B.5) and divide by *p*, the general expression for the difference between the expected service rewards is given by

$$\Delta R\left(\boldsymbol{S}\right)=\pi\left(r-1\right)-\frac{\delta}{p}\left(d-i\right)+p\left(\omega\left(\frac{1-p^{N-j+1}}{1-p}\right)+\sigma p^{N-j}\right) \tag{B.6}$$

When $\Delta R\left(\boldsymbol{S}\right)$ is positive, the patient should be overbooked on day *d*, when it is negative, the patient should be overbooked on day *i*, and when it equals zero, the patient can be booked on either day. The parameter *π* is the benefit received from seeing the additional patient. If we isolate *π* to compare the costs of overbooking to the benefit received, the general expression is given by

$$\Delta R\left(\boldsymbol{S}\right)=\pi>\frac{p\left(\omega\left(\frac{1-p^{N-j+1}}{1-p}\right)+\sigma p^{N-j}\right)-\frac{\delta}{p}\left(d-i\right)}{\left(1-r\right)} \tag{B.7}$$

The first term of the numerator is the cost of overbooking the additional patient on day *i*, and the second is the penalty for deferring the additional patient. The denominator is the probability that the service benefit will not be received if the patient is scheduled on day *d*. When this quantity is less than the service benefit, the patient is overbooked on day *i*.

Equation (B.7) can be reduced based upon the results from Proposition 2 to yield the formulas in Equation (B.3). For example, when $\sigma<\frac{\omega p}{(1-p)}$, the patient should be overbooked in slot *N* of day *i*. Substituting *j*=*N* into Equation (B.7) yields $\pi>\frac{p\left(\omega+\sigma\right)-\frac{\delta}{p}\left(d-i\right)}{\left(1-r\right)}$. For conciseness, we group the case where the patient can be scheduled on either day, π equal to the expected marginal change in cost, with when the patient should be scheduled on day i. This substitution yields expression (**i**) in (B.3). Expressions (**ii**) and (**iii**) in (B.3) can be obtained in a similar manner. Q.E.D.

PROPOSITION 5. *Let i denote a day with N booked appointments, for which an additional patient requests an appointment. Let d represent a day in the future of the scheduling horizon with appointment availability. As a function of θ, the patient should be booked into slot j*, according to the following rules:*

$$\begin{array}{llll}
(\mathbf{i})\ \textit{if} & j^{*}=1 & \textit{and} & \theta \le \dfrac{\pi - p\left(\omega\left(\dfrac{1-p^{N}}{1-p}\right)+\sigma p^{N-1}\right)+\dfrac{\delta(d-i)}{p}}{\pi\alpha^{d-i}\beta^{d-i-1}} \\
(\mathbf{ii})\ \textit{if} & j^{*}=N & \textit{and} & \theta \le \dfrac{\pi - p(\omega+\sigma)+\dfrac{\delta(d-i)}{p}}{\pi\alpha^{d-i}\beta^{d-i-1}} \\
(\mathbf{iii})\ \textit{if} & j^{*}=\textit{any } j\in\{1,\ldots,N\} & \textit{and} & \theta \le \dfrac{\pi - p\left(\dfrac{\omega}{1-p}\right)+\dfrac{\delta(d-i)}{p}}{\pi\alpha^{d-i}\beta^{d-i-1}}
\end{array} \tag{B.8}$$

Otherwise, the patient should be booked into any open slot on day d.

PROOF. This proof follows from the results in Proposition 4. The equations in (B.3) are increasing in θ, thus, if π is greater than the RHS for the largest expected value of θ, then it will be greater than the RHS for all θ, and the patient will always be overbooked on day i. Conversely, if π is less than the RHS for the smallest expected value of θ, then it will be less than the RHS for all θ, and the patient will always be overbooked on day d. When π equals the RHS, this is the point where the patient shifts from being booked on day d, to being overbooked on day i. The expression in (B.8) is derived from solving the equations in (B.3) for θ, and applying these rules. Q.E.D.

PROPOSITION 6. *Let i denote a day with N booked appointments, for which an additional patient requests an appointment. Let d represent a day in the future of the scheduling horizon with N booked appointments, and let $r=\alpha^{d-i}\beta^{d-i-1}\theta$. A clinic schedule achieves a maximal reward by overbooking the additional patient in slot j* on day i, according to the following rules:*

$$\begin{array}{lll} \textbf{(i)} \quad j^* = 1, & \textit{if} & \pi \geq p\left(\omega\left(\dfrac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right)-\dfrac{\delta(d-i)}{p(1-r)} \\ \textbf{(ii)} \quad j^* = N, & \textit{if} & \pi \geq p(\omega+\sigma)-\dfrac{\delta(d-i)}{p(1-r)} \\ \textbf{(iii)} \quad j^* = any\ j \in \{1,\ldots,N\}, & \textit{if} & \pi \geq p\left(\dfrac{\omega}{1-p}\right)-\dfrac{\delta(d-i)}{p(1-r)} \end{array} \tag{B.9}$$

Otherwise, the patient should be overbooked in slot j on day d.*

PROOF. The proof for Proposition 6 is similar to that of Proposition 4. Assume that all days in the scheduling horizon are booked, and there is an additional patient who needs to be overbooked within the scheduling horizon. From Proposition 3, the marginal expected change in day *i's* expected service reward is given by

$$\pi p - \omega p^2\left(\frac{1-p^{N-j+1}}{1-p}\right)-\sigma p^{N-j+2} \tag{B.10}$$

Given that day *d* is also fully booked, the marginal change in day *d's* expected service reward is now given by

$$\pi rp - \delta(d-i) - \omega rp^2\left(\frac{1-p^{N-j+1}}{1-p}\right)-\sigma rp^{N-j+2} \tag{B.11}$$

To determine which case is optimal, we examine the slope of the difference between the two expected service rewards. If we subtract (B.10) from (B.11) and divide by *p*, the general expression for the difference between the expected service rewards is given by

$$\Delta R(\boldsymbol{S}) = (1-r)\times\left[p\left(\omega\left(\frac{1-p^{N-j+1}}{1-p}\right)+\sigma p^{N-j}\right)-\pi\right]-\frac{\delta}{p}(d-i) \tag{B.12}$$

When $\Delta R(\boldsymbol{S})$ is positive, the patient should be overbooked on day *d*, when it is negative, the patient should be overbooked on day *i*, and when it equals zero, the patient can be booked on either day. If we rearrange the terms to isolate *π*, and account for slot placement as in Proposition 2, Expressions (**i**), (**ii**) and (**iii**) follow directly.

The RHS of the expressions in Proposition 6 are equal to the RHS of the expressions in Proposition 3, plus the penalty for making a patient incur indirect waiting. Thus, the RHS expressions of Proposition 6 are always less than the RHS expressions of Proposition 3 when a patient incurs indirect waiting. Given both expressions are compared to the service benefit, *π*, when it is optimal to overbook a patient on day *i*, it is only optimal on day *i*, and never optimal to make the patient incur indirect waiting. Q.E.D.

Proposition 7. *Let i denote a day with N+1 booked appointments, for which an additional patient requests an appointment. If the additional patient is to be overbooked on day i, then a clinic schedule achieves a maximal reward by overbooking the additional patient in slot j**, according to the following rules:*

$$\begin{aligned}
&(\mathbf{i}) \quad \text{if} \quad j^* = 1, && \text{then} \quad j^{**} = \left\lceil \frac{Ln\left[\left(-A+\sqrt{A^2+4A}\right)/2\right]}{Ln[p]} \right\rceil \quad \text{where} \quad A = p^N \left(\sigma/_{\omega}(1-p) - p \right) \\
&(\mathbf{iia})\ \text{if} \quad j^* = N \quad \text{and} \quad \frac{\sigma}{\omega} \geq \frac{(2p-1)}{(1-p)(2-p)}, && \text{then} \quad j^{**} = 1 \\
&(\mathbf{iib})\ \text{if} \quad j^* = N \quad \text{and} \quad \frac{\sigma}{\omega} < \frac{(2p-1)}{(1-p)(2-p)}, && \text{then} \quad j^{**} = N
\end{aligned} \tag{B.13}$$

Proof. The proof of Proposition 7 follows from the marginal expected increase in direct waiting time and overtime when two patients are overbooked on the same day. We assume that all patients, including the overbook requests, are "request day" patients, and thus all have probability *p* of showing. The marginal expected increase in direct waiting time (DWT) and overtime (OT) when one patient is overbooked is given by

$$\begin{aligned}
&DWT \quad \frac{\omega p^2}{1-p}\left(1-p^{N-j+1}\right) \\
&OT \qquad \sigma p^{N-j+2}
\end{aligned} \tag{B.14}$$

The DWT expression is the direct waiting time cost, *ω*, multiplied by a geometric series with *N-j+1* terms, where *N* is the number of slots in the day, and *j* is the slot placement of the overbooked patients. The OT expression is the final term of the DWT expression times the overtime cost, and represents the probability everyone from the overbooked slot to the end of the day, shows for her appointment. After combining and rearranging terms, the total marginal expected increase in service costs for adding an additional patient to a clinic day is $p\left(\omega\left(\frac{1-p^{N-j+1}}{1-p}\right)+\sigma p^{N-j}\right)$.

To calculate the marginal expected increase for adding two patients, we compose similar expressions based upon the slot placements of both overbooked patients. Let j_1 be the slot number of the first slot in the day that is overbooked, and j_2 the slot number of the second overbooked slot, $j_1 \le j_2$. The marginal expected increase in direct waiting time is a combination of geometric series, minus a term to account for the probability of patients not showing. First, if both patients in j_1 show, a unit of direct waiting is incurred by the additional patient. This unit of direct waiting cascades down the clinic schedule if all patients show. Thus, the first geometric series is given by

$$\frac{\omega p^2}{1-p}\left(1-p^{N-j_1+1}\right) \tag{B.15}$$

Second, if both patients in j_2 show, a unit of direct waiting is incurred by the additional patient, which cascades through the clinic day. Thus, the second geometric series is given by

$$\frac{\omega p^2}{1-p}\left(1-p^{N-j_2+1}\right) \tag{B.16}$$

Third, if both overbooked patients show, an additional unit of direct waiting is incurred for patients in j_2, and all subsequent patients. The number of patients who must show for the additional unit of direct waiting to be incurred is the number of patients between the two overbooked slots plus both patients in j_2 Thus, the third geometric series is given by

$$\frac{\omega p^{j_2-j_1+2}}{1-p}\left(1-p^{N-j_2+1}\right) \tag{B.17}$$

The third geometric series is contingent on both overbooked patients, and everyone after j_2 showing. The probability of these patients no-showing must also be considered. This is captured in a single term that represents the number of patients in the schedule who need to show for all units of waiting to be accrued, $N - j_1 + 3$, times the number of ways those patients can no-show and effect the waiting time, $N - j_2 + 1$. Additionally, there is one unit of waiting accrued in overtime if all patients show. Thus, the final term in the total marginal expected increase in direct waiting time when overbooking two patients in one day is given by

$$-\left(N-j_2\right)\omega p^{N-j_1+3} \tag{B.18}$$

The total marginal expected increase in overtime when overbooking two patients is equal to the total expected backlog at slot *N*. There will be a unit of overtime if both patients in j_2 and all subsequent patients show, and one unit of waiting if both patients in j_1 show, and a single person in every subsequent slot shows. The probability of patient no-show is captured similarly to the waiting time, thus, the total marginal overtime is given by

$$\sigma p^{N-j_2+2} + \left(N - j_2 + 2\right)\sigma p^{N-j_1+2} - \left(N - j_2 + 1\right)\sigma p^{N-j_1+3} \qquad \text{(B.19)}$$

Combining terms, the direct waiting time and overtime expressions are given by

$$\begin{array}{ll} DWT & \dfrac{\omega p^2}{1-p}\left(2 - 2p^{N-j_1+1} + p^{j_2-j_1} - p^{N-j_2+1}\right) - \left(N - j_2\right)\omega p^{N-j_1+3} \\ OT & \sigma p^{N-j_2+2} + \left(N - j_2 + 2\right)\sigma p^{N-j_1+2} - \left(N - j_2 + 1\right)\sigma p^{N-j_1+3} \end{array} \qquad \text{(B.20)}$$

Because we assume that the patients are being sequentially overbooked, we seek to find the marginal expected change in direct waiting and overtime when overbooking the second patient. This change is dependent on where the first patient is overbooked.

Case 1: Assume the first overbooked patient is overbooked in slot 1. Thus, $j_1 = 1$, and the second overbooked patient will be overbooked in $j_2 \geq 1$. Substituting $j = 1$ into Equation (B.14) and subtracting these terms from Equation (B.20) when $j_1 = 1$ yields

$$\begin{array}{ll} DWT & \dfrac{\omega p^2}{1-p}\left(1 - p^N + p^{j_2-1} - p^{N-j_2+1}\right) - \left(N - j_2\right)\omega p^{N+2} \\ OT & \sigma p^{N-j_2+2} + \left(N - j_2 + 1\right)\sigma p^{N+1} - \left(N - j_2 + 1\right)\sigma p^{N+2} \end{array} \qquad \text{(B.21)}$$

After rearranging the terms and labeling the slot placement of the second overbooked patient as *j***, the total expected marginal change in service costs when overbooking a second patient when the first patient is overbooked in slot 1 is given by

$$p^2\left(\omega\left(\frac{1-p^N}{1-p}\right) + \omega\left(\frac{p^{j^{**}-1} - p^{N-j^{**}+1}}{1-p}\right) + \sigma p^{N-1} - \sigma p^N\left(1 - p^{-j^{**}}\right) + p^{N-1}\left(N - j^{**}\right)\left(\sigma\left(1-p\right) - \omega p\right)\right)$$

(B.22)

The first and third terms are the total expected service costs when overbooking a patient in slot 1, as found in Proposition 3, and the subsequent terms represent the additional expected accrued waiting.

Case 2: Assume the first overbooked patient is overbooked in slot *N*. Thus $j_2 = N$, and the second overbooked patient will be overbooked in $j_1 \le N$. Substituting $j = N$ into Equation (B.14) and subtracting these terms from Equation (B.20) when $j_2 = N$ yields

$$\begin{array}{ll} DWT & \dfrac{\omega p^2}{1-p}\left(1-2p^{N-j_1+1}+p^{N-j_1}\right) \\ OT & \sigma p^{N-j_1+2}\left(2-p\right) \end{array} \tag{B.23}$$

After rearranging the terms and labeling the slot placement of the second overbooked patient as *j***, the total expected marginal change in service costs when overbooking a second patient when the first patient is overbooked in slot *N* is given by

$$p^2\left(\omega\left(\frac{1-p^{N-j^{**}}\left(2p-1\right)}{1-p}\right)+\sigma p^{N-j^{**}}\left(2-p\right)\right) \tag{B.24}$$

To determine the value of *j*** we evaluate the change in the day's expected service reward for each case when overbooking the second patient in slot *j*** versus *j**+1*.

Case 1: Assume the first overbooked patient is overbooked in slot 1. Then the marginal expected change in the objective function when overbooking a second patient is given by

$$\Delta R\left(\boldsymbol{S}[i, j^{**}]\right)_{j^*=1} = \pi p - p^2\left(\omega\left(\frac{1-p^N}{1-p}\right)+\omega\left(\frac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)+\sigma p^{N-1}-\sigma p^N\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma\left(1-p\right)-\omega p\right)\right) \tag{B.25}$$

Evaluating $\Delta R\left(\boldsymbol{S}[i, j^{**}+1]\right)_{j^*=1} - \Delta R\left(\boldsymbol{S}[i, j^{**}]\right)_{j^*=1} = 0$ leads to the optimal value of *j*** as shown in Equation (B.13**i**). Thus, optimal placement of the second overbooked patient is a function of the clinic parameters and the patient's probability of show. Given that *j*** can be calculated to be non-integer, we assign *j*** to the next greatest integer after the value is calculated.

Case 2: Assume the first overbooked patient is overbooked in slot *N*. Then the marginal expected change in the objective function when overbooking a second patient is given by

$$\Delta R\left(\boldsymbol{S}[i, j^{**}]\right)_{j^*=N} = \pi p - p^2\left(\omega\left(\frac{1-p^{N-j^{**}}\left(2p-1\right)}{1-p}\right) + \sigma p^{N-j^{**}}\left(2-p\right)\right) \qquad \text{(B.26)}$$

When $\Delta R\left(\boldsymbol{S}[i, j^{**}+1]\right)_{j^*=N} - \Delta R\left(\boldsymbol{S}[i, j^{**}]\right)_{j^*=N} \geq 0$, *j**=N*, and when the value is ≤ 0, *j**=1*. As in Proposition 2, the results of the calculation lead to *j*** equal to the end slots, based upon a comparison of the overtime cost, *σ*, with an expression involving the direct waiting cost, *ω*, and the probability of show, *p*. Q.E.D.

PROPOSITION 8. *Let i denote a day with N+1 booked appointments, for which an additional patient requests an appointment. Let d represent a day in the future of the scheduling horizon with N booked appointments, and let $r = \alpha^{d-i}\beta^{d-i-1}\theta$. A clinic schedule achieves a maximal reward by overbooking the additional patient in slot j** on day i, according to the following rules:*

$$\begin{array}{llllll}
\textbf{(i)} & j^{**} \text{ from Prop.7,} & \textit{if} & j^*=1 & \textit{and} & \pi \geq p\left(\omega\left(\dfrac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right)+\dfrac{p\left(\omega\left(\dfrac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)-\sigma p^N\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma\left(1-p\right)-\omega p\right)\right)-\dfrac{\delta}{p}\left(d-i\right)}{1-r} \\
\textbf{(ii)} & j^{**}=1, & \textit{if} & j^*=N & \textit{and} & \pi \geq p\left(\omega+\sigma\right)+\dfrac{p\left(\omega\left(\dfrac{1-p^{N-1}\left(2p-1\right)}{1-p}\right)+\sigma p^{N-1}\left(2-p\right)-\left(\omega+\sigma\right)\right)-\dfrac{\delta}{p}\left(d-i\right)}{1-r} \\
\textbf{(iii)} & j^{**}=N, & \textit{if} & j^*=N & \textit{and} & \pi \geq p\left(\omega+\sigma\right)+\dfrac{p\left(\omega+\sigma\left(1-p\right)\right)-\dfrac{\delta}{p}\left(d-i\right)}{1-r}
\end{array}$$

(B.27)

Otherwise, the patient should be overbooked in slot j on day d.*

PROOF. The proof for Proposition 8 is similar to that of Proposition 6. Assume that all days in the scheduling horizon are booked, a single patient is overbooked on day *i*, and there is an additional patient who needs to be overbooked within the scheduling horizon. From Proposition 7, the marginal expected change in day *i's* expected service reward is given by

$$\textbf{Case 1:}\quad \Delta R\left(\boldsymbol{S}[i]\right)=\pi p-p^2\left(\omega\left(\frac{1-p^N}{1-p}\right)+\omega\left(\frac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)+\sigma p^{N-1}-\sigma p^N\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma\left(1-p\right)-\omega p\right)\right)$$

$$\textbf{Case 2:}\quad \Delta R\left(\boldsymbol{S}[i]\right)=\pi p-p^2\left(\omega\left(\frac{1-p^{N-j^{**}}\left(2p-1\right)}{1-p}\right)+\sigma p^{N-j^{**}}\left(2-p\right)\right)$$

(B.28)

Given that Day *d* is also fully booked, the marginal change in day *d's* expected service reward is given by

$$\textbf{Case 1:}\quad \Delta R\left(\boldsymbol{S}[d]\right)=\pi rp-\delta\left(d-i\right)-rp^2\left(\omega\left(\frac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right) \qquad \text{(B.29)}$$

$$\textbf{Case 2:}\quad \Delta R\left(\boldsymbol{S}[d]\right)=\pi rp-\delta\left(d-i\right)-rp^2\left(\omega+\sigma\right)$$

To determine which case is optimal, we examine the slope of the difference between the two expected service rewards. If we subtract $\Delta R\left(\boldsymbol{S}[i]\right)$ from $\Delta R\left(\boldsymbol{S}[d]\right)$ and divide by *p*, the general expression for the difference between the expected service rewards is given by

$$\textbf{Case 1:}\quad \Delta R\left(\boldsymbol{S}\right)=\pi\left(r-1\right)+p\left(\omega\left(\frac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right)\left(1-r\right)+p\left(\omega\left(\frac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)-\sigma p^N\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma\left(1-p\right)-\omega p\right)\right)-\frac{\delta}{p}\left(d-i\right)$$

$$\textbf{Case 2:}\quad \Delta R\left(\boldsymbol{S}\right)=\pi\left(r-1\right)+p\left(\omega\left(\frac{1-p^{N-j^{**}}\left(2p-1\right)}{1-p}\right)+\sigma p^{N-j^{**}}\left(2-p\right)\right)-pr\left(\omega+\sigma\right)-\frac{\delta}{p}\left(d-i\right)$$

(B.30)

When $\Delta R\left(\boldsymbol{S}\right)$ is positive, the patient should be overbooked on day *d*, when it is negative, the patient should be overbooked on day *i*, and when it equals zero, the patient can be booked on either day. If we rearrange the terms to isolate *π*, and account for slot placement as in Proposition 7, Expressions (**i**), (**ii**) and (**iii**) follow directly. Q.E.D.

PROPOSITION 9. *Let i denote a day with N+1 booked appointments, for which an additional patient requests an appointment. Let d represent a day in the future of the scheduling horizon with N booked appointments. As a function of θ, the patient should be booked into slot j** on day i, according to the following rules:*

$$\begin{array}{lllll}
\textbf{(i)} & j^{**} \text{ from Prop.7}, & \textit{if} \quad j^{*}=1 & \textit{and} & \theta \leq \dfrac{\pi - p\left(\omega\left(\dfrac{1-p^{N}}{1-p}\right)+\omega\left(\dfrac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)+\sigma p^{N-1}-\sigma p^{N}\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma(1-p)-\omega p\right)\right)+\dfrac{\delta(d-i)}{p}}{\left(\pi - p\left(\omega\left(\dfrac{1-p^{N}}{1-p}\right)+\sigma p^{N-1}\right)\right)\alpha^{d-i}\beta^{d-i-1}} \\
\textbf{(ii)} & j^{**}=1, & \textit{if} \quad j^{*}=N & \textit{and} & \theta \leq \dfrac{\pi - p\left(\omega\left(\dfrac{1-p^{N-1}(2p-1)}{1-p}\right)+\sigma p^{N-1}(2-p)-(\omega+\sigma)\right)+\dfrac{\delta(d-i)}{p}}{\left(\pi - p(\omega+\sigma)\right)\alpha^{d-i}\beta^{d-i-1}} \\
\textbf{(iii)} & j^{**}=N, & \textit{if} \quad j^{*}=N & \textit{and} & \theta \leq \dfrac{\pi - p\left(2(\omega+\sigma)-\sigma p\right)+\dfrac{\delta(d-i)}{p}}{\left(\pi - p(\omega+\sigma)\right)\alpha^{d-i}\beta^{d-i-1}}
\end{array}$$

(B.31)

Otherwise, the patient should be booked into slot j on day d.*

PROOF. Proof for Proposition 9 follows from that of Proposition 5. For this Proposition we solve for the equations in (B.27) to obtain the bounds on when *θ* affects a schedule. Q.E.D.

PROPOSITION 10. *Let* i *denote a day with N+1 booked appointments, for which an additional patient requests an appointment. Let* d *represent a day in the future of the scheduling horizon with N booked appointments. It is optimal to overbook the additional patient, according to the following rules:*

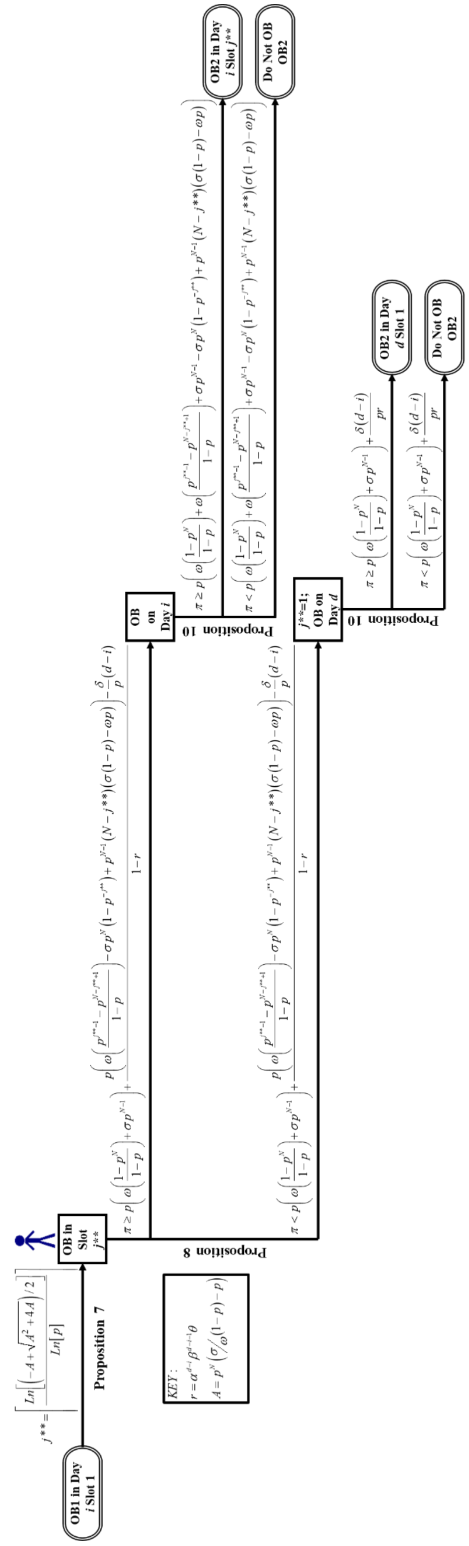
$$\begin{array}{llllll}
\textbf{(i)} & j^{**} \text{ from Prop.7}, & \textit{if} \quad j^{*}=1, & \text{Prop}\,7 \rightarrow \textit{Day } i, & \textit{and} & \pi \geq p\left(\omega\left(\dfrac{1-p^{N}}{1-p}\right)+\omega\left(\dfrac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)+\sigma p^{N-1}-\sigma p^{N}\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma(1-p)-\omega p\right)\right) \\
\textbf{(ii)} & j_{d}^{*}=1, & \textit{if} \quad j^{*}=1, & \text{Prop}\,7 \rightarrow \textit{Day } d, & \textit{and} & \pi \geq p\left(\omega\left(\dfrac{1-p^{N}}{1-p}\right)+\sigma p^{N-1}\right)+\dfrac{\delta(d-i)}{pr} \\
\textbf{(iii)} & j^{**}=1, & \textit{if} \quad j^{*}=N, & \text{Prop}\,7 \rightarrow \textit{Day } i, & \textit{and} & \pi \geq p\left(\omega\left(\dfrac{1-p^{N-1}(2p-1)}{1-p}\right)+\sigma p^{N-1}(2-p)\right) \\
\textbf{(iv)} & j_{d}^{*}=N, & \textit{if} \quad j^{*}=N, & \text{Prop}\,7 \rightarrow \textit{Day } d, & \textit{and} & \pi \geq p(\omega+\sigma)+\dfrac{\delta(d-i)}{pr} \\
\textbf{(v)} & j^{**}=N, & \textit{if} \quad j^{*}=N, & \text{Prop}\,7 \rightarrow \textit{Day } i, & \textit{and} & \pi \geq p\left(2(\omega+\sigma)-\sigma p\right) \\
\textbf{(vi)} & j_{d}^{*}=N, & \textit{if} \quad j^{*}=N, & \text{Prop}\,7 \rightarrow \textit{Day } d, & \textit{and} & \pi \geq p(\omega+\sigma)+\dfrac{\delta(d-i)}{pr}
\end{array}$$

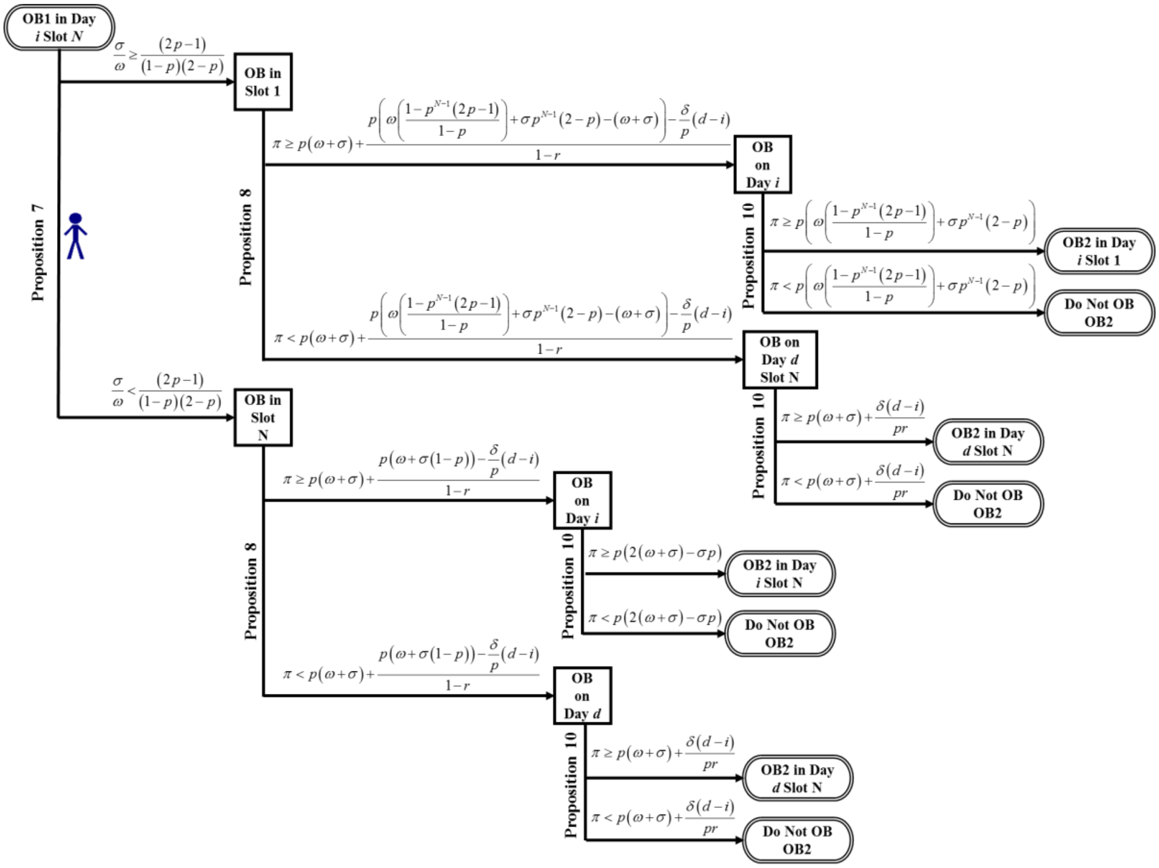
(B.32)

Where j_{d}^{*} *denotes the slot in which OB2 is overbooked on day d.*

PROOF. To determine if it is optimal to overbook a second patient on the preferred day, we evaluate if the marginal expected change in the service reward for that day, given the overbooking, is non-negative. The marginal expected change in the service reward when *j*=1* and *j*=N*, and the preferred day to overbook the second patient is day *i*, are given in Equation (B.25) and Equation (B.26), respectively. Substituting the optimal *j*** value and rearranging terms to compare the service costs to *π* yield the optimality rules in Proposition 8 when the optimal day is day *i*. When the optimal day to overbook is day *d*, the patient is the first overbook on that day, because the first overbooked patient is always overbooked on day *i*, and the function to evaluate is given in Equation (B.10). Q.E.D.

B.2 OVERBOOKING OB2 PROCESS FLOWS


OB1 in Day i Slot 1
$j^{**}=\left\lceil \frac{Ln\left[\left(-A+\sqrt{A^2+4A}\right)/2\right]}{Ln[p]} \right\rceil$
Proposition 7
OB in Slot j^{**}
Proposition 8
KEY:
$r=\alpha^{d-i}\beta^{d-i-1}\theta$
$A=p^N\left(\sigma/\omega(1-p)-p\right)$
$\pi \geq p\left(\omega\left(\frac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right)+\frac{p\left(\omega\left(\frac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)-\sigma p^N\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma(1-p)-\omega p\right)\right)-\frac{\delta}{p}(d-i)}{1-r}$
$\pi < p\left(\omega\left(\frac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right)+\frac{p\left(\omega\left(\frac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)-\sigma p^N\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma(1-p)-\omega p\right)\right)-\frac{\delta}{p}(d-i)}{1-r}$
OB on Day i
Proposition 10
$\pi \geq p\left(\omega\left(\frac{1-p^N}{1-p}\right)+\omega\left(\frac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)+\sigma p^{N-1}-\sigma p^N\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma(1-p)-\omega p\right)\right)$
$\pi < p\left(\omega\left(\frac{1-p^N}{1-p}\right)+\omega\left(\frac{p^{j^{**}-1}-p^{N-j^{**}+1}}{1-p}\right)+\sigma p^{N-1}-\sigma p^N\left(1-p^{-j^{**}}\right)+p^{N-1}\left(N-j^{**}\right)\left(\sigma(1-p)-\omega p\right)\right)$
OB2 in Day i Slot j^{**}
Do Not OB OB2
j^{**}=1; OB on Day d
Proposition 10
$\pi \geq p\left(\omega\left(\frac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right)+\frac{\delta(d-i)}{pr}$
$\pi < p\left(\omega\left(\frac{1-p^N}{1-p}\right)+\sigma p^{N-1}\right)+\frac{\delta(d-i)}{pr}$
OB2 in Day d Slot 1
Do Not OB OB2


OB1 in Day *i* Slot *N*
$\frac{\sigma}{\omega} \geq \frac{(2p-1)}{(1-p)(2-p)}$
OB in Slot 1
$\frac{\sigma}{\omega} < \frac{(2p-1)}{(1-p)(2-p)}$
OB in Slot N
Proposition 7
Proposition 8
$\pi \geq p(\omega+\sigma) + \frac{p\left(\omega\left(\frac{1-p^{N-1}(2p-1)}{1-p}\right) + \sigma p^{N-1}(2-p) - (\omega+\sigma)\right) - \frac{\delta}{p}(d-i)}{1-r}$
OB on Day *i*
$\pi < p(\omega+\sigma) + \frac{p\left(\omega\left(\frac{1-p^{N-1}(2p-1)}{1-p}\right) + \sigma p^{N-1}(2-p) - (\omega+\sigma)\right) - \frac{\delta}{p}(d-i)}{1-r}$
OB on Day *d* Slot N
Proposition 10
$\pi \geq p\left(\omega\left(\frac{1-p^{N-1}(2p-1)}{1-p}\right) + \sigma p^{N-1}(2-p)\right)$
OB2 in Day *i* Slot 1
$\pi < p\left(\omega\left(\frac{1-p^{N-1}(2p-1)}{1-p}\right) + \sigma p^{N-1}(2-p)\right)$
Do Not OB OB2
Proposition 10
$\pi \geq p(\omega+\sigma) + \frac{\delta(d-i)}{pr}$
OB2 in Day *d* Slot N
$\pi < p(\omega+\sigma) + \frac{\delta(d-i)}{pr}$
Do Not OB OB2
Proposition 8
$\pi \geq p(\omega+\sigma) + \frac{p(\omega+\sigma(1-p)) - \frac{\delta}{p}(d-i)}{1-r}$
OB on Day *i*
Proposition 10
$\pi \geq p(2(\omega+\sigma)-\sigma p)$
OB2 in Day *i* Slot N
$\pi < p(2(\omega+\sigma)-\sigma p)$
Do Not OB OB2
$\pi < p(\omega+\sigma) + \frac{p(\omega+\sigma(1-p)) - \frac{\delta}{p}(d-i)}{1-r}$
OB on Day *d*
Proposition 10
$\pi \geq p(\omega+\sigma) + \frac{\delta(d-i)}{pr}$
OB2 in Day *d* Slot N
$\pi < p(\omega+\sigma) + \frac{\delta(d-i)}{pr}$
Do Not OB OB2

BIBLIOGRAPHY

Alaeddini, A., Yang, K., Reeves, P. & Reddy, C. K. 2015. A hybrid prediction model for no-shows and cancellations of outpatient appointments. *IIE Transactions on Healthcare Systems Engineering,* 5**,** 14-32.

Bailey, N. T. 1952. A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. *Journal of the Royal Statistical Society. Series B (Methodological)***,** 185-199.

Bean, A. & Talaga, J. 1995. Predicting Appointment Breaking. *Journal of Health Care Marketing,* 15**,** 29.

Berchtold, A. 2005, October 18. Statistics: Research, Teaching, Consulting. Retrieved from http://andreberchtold.com/march.html on January 21, 2016

Berchtold, A. & Raftery, A. E. 2002. The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science,* 17**,** 328-356.

Berg, B. P., Denton, B. T., Erdogan, S. A., Rohleder, T. & Huschka, T. 2014. Optimal booking and scheduling in outpatient procedure centers. *Computers & Operations Research,* 50**,** 24-37.

Berger, J. O. 1985. *Statistical Decision Theory and Bayesian Analysis,* Springer Science & Business Media.

Beylkin, G. & Monzón, L. 2005. On approximation of functions by exponential sums. *Applied and Computational Harmonic Analysis,* 19**,** 17-48.

Beylkin, G. & Monzón, L. 2010. Approximation by exponential sums revisited. *Applied and Computational Harmonic Analysis,* 28**,** 131-149.

Box, G. E., Jenkins, G. M. & Reinsel, G. C. 2011. *Time series analysis: forecasting and control,* John Wiley & Sons.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review,* 78**,** 1-3.

Burnham, K. P. & Anderson, D. R. 2002. *Model selection and multimodel inference: a practical information-theoretic approach,* Springer.

Cayirli, T. & Veral, E. 2003. Outpatient scheduling in health care: a review of literature. *Production and Operations Management,* 12**,** 519-549.

Cayirli, T., Veral, E. & Rosen, H. 2006. Designing appointment scheduling systems for ambulatory care services. *Health care management science,* 9**,** 47-58.

Chariatte, V., Berchtold, A., Akré, C., Michaud, P.-A. & Suris, J.-C. 2008. Missed appointments in an outpatient clinic for adolescents, an approach to predict the risk of missing. *Journal of Adolescent Health,* 43**,** 38-45.

Chiu, T., Fang, D., Chen, J., Wang, Y. & Jeris, C. A robust and scalable clustering algorithm for mixed type attributes in large database environment. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2001. ACM, 263-268.

Cosgrove, M. 1990. Defaulters in general practice: reasons for default and patterns of attendance. *The British Journal of General Practice,* 40**,** 50.

Cox, D. R., Gudmundsson, G., Lindgren, G., Bondesson, L., Harsaae, E., Laake, P., Juselius, K. & Lauritzen, S. L. 1981. Statistical Analysis of Time Series: Some Recent Developments [with Discussion and Reply]. *Scandinavian Journal of Statistics,* 8**,** 93-115.

Daggy, J., Lawley, M., Willis, D., Thayer, D., Suelzer, C., Delaurentis, P.-C., Turkcan, A., Chakraborty, S. & Sands, L. 2010. Using no-show modeling to improve clinic performance. *Health Informatics Journal,* 16**,** 246-259.

Davis, J. & Goadrich, M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning, 2006. ACM, 233-240.

Delong, E. R., Delong, D. M. & Clarke-Pearson, D. L. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics,* 44**,** 837-845.

Fader, P. S., Hardie, B. G. S. & Shang, J. 2010. Customer-base analysis in a discrete-time noncontractual setting. *Marketing Science,* 29**,** 1086-1108.

Gallucci, G., Swartz, W. & Hackerman, F. 2005. Brief reports: Impact of the wait for an initial appointment on the rate of kept appointments at a mental health center. *Psychiatric Services*.

Garuda, S. R., Javalgi, R. G. & Talluri, V. S. 1998. Tackling no-show behavior: A market-driven approach. *Health Marketing Quarterly,* 15**,** 25-44.

Glowacka, K. J., Henry, R. M. & May, J. H. 2009. A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. *Journal of the Operational Research Society,* 60**,** 1056-1068.

Goffman, R., Harris, S. L., May, E. K., May, J. H., Tjader, Y. C., & Vargas, D. L., 2015. "Modeling Patient No-Show History and Predicting Future Outpatient Appointment Behavior." Working Paper, Joseph M. Katz Graduate School of Business, University of Pittsburgh.

Griva, I., Nash, S. G. & Sofer, A. 2009. *Linear and nonlinear optimization*, Siam.

Gupta, D. & Denton, B. 2008. Appointment scheduling in health care: Challenges and opportunities. *IIE transactions,* 40**,** 800-819.

Heckman, J. J. 1980. Sample selection bias as a specification error. *Female labor supply: Theory and estimation***,** 206-48.

Horn, R. A., and Johnson C. R. 2012. Matrix analysis. *Cambridge University Press*.

Huang, Y. & Hanauer, D. 2014. Patient No-Show Predictive Model Development using Multiple Data Sources for an Effective Overbooking Approach. *Appl Clin Inform,* 5**,** 836-860.

Huang, Y. & Zuniga, P. 2013. Effective cancellation policy to reduce the negative impact of patient no-show. *Journal of the Operational Research Society,* 65**,** 605-615.

IOM (Institute of Medicne). 2015. “Transforming health care scheduling and access: Getting to now.” Washington, DC: *The National Academies Press*.

Kenter, R., Warmerdam, L., Brouwer-Dudokdewit, C., Cuijpers, P. & Van Straten, A. 2013. Guided online treatment in routine mental health care: an observational study on uptake, drop-out and effects. *BMC psychiatry,* 13**,** 43.

Kim, S. & Giachetti, R. E. 2006. A stochastic mathematical appointment overbooking model for healthcare providers to improve profits. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on,* 36**,** 1211-1219.

Laganga, L. R. & Lawrence, S. R. 2007. Clinic overbooking to improve patient access and increase provider productivity. *Decision Sciences,* 38**,** 251-276.

Laganga, L. R. & Lawrence, S. R. 2012. Appointment overbooking in health care clinics to improve patient service and clinic performance. *Production and Operations Management,* 21**,** 874-888.

Lawrence, R. D., Hong, S. J. & Cherrier, J. Passenger-based predictive modeling of airline no-show rates. Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003. ACM, 397-406.

Li, W. 1994. Time series models based on generalized linear models: some further results. *Biometrics,* 50**,** 506-511.

Ling, C. X. & Li, C. Data Mining for Direct Marketing: Problems and Solutions. KDD, 1998. 73-79.

Liu, N., Ziya, S. & Kulkarni, V. G. 2010. Dynamic scheduling of outpatient appointments under patient no-shows and cancellations. *Manufacturing & Service Operations Management,* 12**,** 347-364.

Mehran, F. 1989. Longitudinal analysis of employment and unemployment based on matched rotation samples. *Labour,* 3**,** 3-20.

Morales, D. R. & Wang, J. 2010. Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research,* 202**,** 554-562.

Muthuraman, K. & Lawley, M. 2008. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions,* 40**,** 820-837.

Norris, J. B., Kumar, C., Chand, S., Moskowitz, H., Shade, S. A. & Willis, D. R. 2014. An empirical investigation into factors affecting patient cancellations and no-shows at outpatient clinics. *Decision Support Systems,* 57**,** 428-443.

Parizi, M. S. & Ghate, A. 2016. Multi-class, multi-resource advance scheduling with no-shows, cancellations and overbooking. *Computers & Operations Research,* 67**,** 90-101.

Patrick, J. 2012. A Markov decision model for determining optimal outpatient scheduling. *Health care management science,* 15**,** 91-102.

Pereyra, V. & Scherer, G. 2010. *Exponential Data Fitting and its Applications.* Bentham Science Publishers.

Prinzie, A. & Van Den Poel, D. 2006. Investigating purchasing-sequence patterns for financial services using Markov, MTD and MTDg models. *European Journal of Operational Research,* 170**,** 710-734.

Quinlan, R. 2004. Data mining tools See5 and C5. 0.

Raftery, A. E. 1985. A model for high-order Markov chains. *Journal of the Royal Statistical Society. Series B (Methodological),* 47**,** 528-539.

Reid, M., Cohen, S., Wang, H., Kaung, A., Patel, A., Tashjian, V., Williams Jr, D., Martinez, B. & Spiegel, B. 2015. Preventing patient absenteeism: validation of a predictive overbooking model. *The American journal of managed care,* 21**,** 902-910.

Rust, C. T., Gallups, N. H., Clark, W. S., Jones, D. S. & Wilcox, W. D. 1995. Patient appointment failures in pediatric resident continuity clinics. *Archives of Pediatrics & Adolescent Medicine,* 149**,** 693.

Samorani, M. & Laganga, L. R. 2015. Outpatient appointment scheduling given individual day-dependent no-show predictions. *European Journal of Operational Research,* 240**,** 245-257.

Startz, R. 2008. Binomial autoregressive moving average models with an application to US recessions. *Journal of Business and Economic Statistics,* 26**,** 1-8.

Vasilakis, C. & Marshall, A. H. 2005. Modelling nationwide hospital length of stay: opening the black box. *Journal of the Operational Research Society,* 56**,** 862-869.

Xie, H., Chaussalet, T. J. & Millard, P. H. 2005. A continuous time Markov model for the length of stay of elderly people in institutional long-term care. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 168**,** 51-61.

Zacharias, C. & Pinedo, M. 2014. Appointment Scheduling with No-Shows and Overbooking. *Production and Operations Management,* 23**,** 788-801.

Zadrozny, B. & Elkan, C. Learning and making decisions when costs and probabilities are both unknown. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, 2001. ACM, 204-213.

Zeger, S. L. & Qaqish, B. 1988. Markov regression models for time series: a quasi-likelihood approach. *Biometrics,* 44**,** 1019-1031.

Zeng, B., Turkcan, A., Lin, J. & Lawley, M. 2010. Clinic scheduling models with overbooking for patients with heterogeneous no-show probabilities. *Annals of Operations Research,* 178**,** 121-144.