

PROBABILISTIC COST MODEL TO ESTIMATE THE COST EXPECTED VALUE OF NO-SHOW IN OUTPATIENT CLINICS

by

Samuel A. Bonet Olivencia

A thesis submitted in partial fulfillment of the requirements for the degree of

**MASTERS OF SCIENCE
in
INDUSTRIAL ENGINEERING**

**UNIVERSITY OF PUERTO RICO
MAYAGUEZ CAMPUS
2016**

Approved by:

Mayra Méndez-Piñero, PhD
President, Graduate Committee

Date

Sonia Bartolomei-Suárez, PhD
Member, Graduate Committee

Date

Betzabé Rodríguez-Álamo, PhD
Member, Graduate Committee

Date

Abigail Matos Pagán, PhD
Representative of Graduate Studies

Date

Viviana Cesaní-Vázquez, PhD
Chairperson of the Department

Date

ABSTRACT

Efforts have been made to reduce the costs associated to healthcare systems. No-shows to medical appointments impact healthcare systems finances and efficiency. Outpatient clinics have used diverse methods to decrease no-show rates; however, it has not been eradicated. This work presents a framework for the development of a cost model that integrates patients' probabilities of no-show and cost information to determine the cost expected value of an appointment slot as a representation of the cost of a no-show to the system. The cost model has led to the development of a procedure for the evaluation of scheduling scenarios with the purpose of identifying the ones that perform better. A prototype of an interactive platform in Excel has been created to demonstrate the application of the developed methodology, and also to provide a tool that may help an outpatient facility in the scheduling process.

RESUMEN

Múltiples esfuerzos se han realizado con el propósito de reducir los costos asociados a los sistemas de salud. Las ausencias a las citas médicas, específicamente los pacientes que se ausentan sin aviso previo (“no-shows”), impactan los sistemas de salud en términos financieros y de eficiencia. Clínicas ambulatorias han utilizado varios métodos para reducir la incidencia de ausencias, sin embargo, las ausencias a citas médicas no han sido erradicadas. Este trabajo de investigación presenta un esquema general para el desarrollo de un modelo de costos que integra probabilidades de “no-show” con un análisis de costos, con el propósito de determinar el valor esperado del costo de una cita médica, que a la misma vez sirve de estimado del costo que un “no show” representa a una clínica ambulatoria. El modelo estocástico de costos se ha utilizado como base para el desarrollo de un procedimiento para la evaluación de diferentes esquemas de itinerarios, para así identificar posibles escenarios que presenten mejor desempeño en términos del costo total al sistema. Adicional, se introduce un prototipo de una herramienta para generar itinerarios de citas médicas, el cual se presenta en forma de una plataforma interactiva desarrollada utilizando Excel. La plataforma interactiva demuestra la aplicación de la metodología desarrollada en este trabajo, además provee una herramienta que puede servir de ayuda a las clínicas ambulatorias durante el proceso de realizar el itinerario de citas médicas.

DEDICATION

To my parents Julio C. Bonet and Josefa Olivencia, because your love and unconditional support have been essential during this process. Thank you for letting me dream big and for showing me the value of education. I appreciate your wise advice and your words of encouragement in the difficult times. You are my inspiration and my role model.

ACKNOWLEDGMENTS

First of all, I want to thank God for giving me the strength to complete this journey. Also, I want to thank my family for their love and support. Special thanks to my parents, Julio C. Bonet and Josefa Olivencia, for believing in me and supporting me in every step I take. Thanks to my sisters Tamara and Xiomara, and my brother Sammy for their good energy and support.

The development of this thesis was possible thanks to the support and academic advice of my Masters Degree advisor, Dr. Mayra Méndez Piñero, PhD. Thanks for believing in me, for the words of encouragement and the guidance in every step of the way. Thanks for being a friend and sharing with me the ups and downs of this journey. I will always be grateful for your personal and professional advice throughout my academic career.

Special thanks to Daniel Cortés and Gabriela Salvat, for their research assistance and collaboration during this process. Additionally I want to thanks my graduate committee, Dr. Betzabé Rodríguez and Dr. Sonia Bartolomei, for their input and guidance. Very special thanks to Dr. Saylisse Dávila for giving me academic advice, important for the completion of this work. I want to thanks the Industrial Engineering Department for providing me the financial support during my Masters Degree. Thanks to Dr. Sonia Bartolomei and Dr. Omell Pagán for their coaching and trust during my time as their teaching assistant.

Finally I want to thanks to my friends Jorlys Alvarado, Yaileen Méndez, Esmeralda Niño, Nitza García, Katia Camacho, Yaritza Santiago, Juan Rosas and Cesar Salazar for their support, motivation and unconditional friendship. They are an outstanding group of people that have made this journey more fun and pleasant.

TABLE OF CONTENTS

1	INTRODUCTION	1
1.1	Significance	3
1.2	Justification and Contribution.....	7
1.3	Thesis Organization	8
2	LITERATURE REVIEW	10
2.1	Predictors and Modeling.....	10
2.1.1	Factors Selection	10
2.1.2	Use of Data Base Information and Statistical Modeling.....	14
2.2	Scheduling Applications	16
2.3	Financial Considerations.....	19
3	METHODOLOGY	24
3.1	Variable Selection.....	24
3.2	Data Simulation	27
3.3	Classification and Regression Tree (CART)	29
3.4	Economic Analysis	34
3.5	Cost Model for the Expected Cost of the Appointment Slot	48
3.6	Test Bed	51
3.7	Interactive Platform in Excel	52
4	RESULTS	54

4.1	Cost Model Results	54
4.2.1	Overflow Probabilities	54
4.2.2	Value Lost Due to the Non-Utilization Cost Expected Value ($E(V)$)	67
4.2.3	Waiting Cost Expected Value ($E(W)$)	69
4.2.4	Personnel Overtime Cost Expected Value ($E(OP)$)	69
4.2.5	Stochastic Cost Model.....	70
4.3	Test Bed-Scenarios Simulation.....	71
4.4	Interactive Platform-Appointment Scheduler	79
5	CONCLUSION AND FUTURE WORK	82
	REFERENCES	85
	APPENDIX A.....	89
A.1	Classification and Regression Trees Steps	89
A.1.1	Tree Growing	89
A.1.2	Tree Pruning.....	90
A.1.3	Tree Performance Validation	91
A.2	Gradient Boosted Trees	94
A.3	R packages for CART and Gradient Boosted Trees	96
A.3.1	R-part.....	96
A.3.2	Caret and gbm	97

APPENDIX B	99
B.1 CART Results-Representative Example.....	99
B.2 Gradient Boosted Trees Results-Representative Example.....	107
APPENDIX C	110

TABLE LIST

Table 1. Summary of overall relevant factors found in seven publications.....	25
Table 2. Attributes categorized according to the “determinants” of broken appointments.....	26
Table 3. Final fourteen significant attributes identified.....	27
Table 4. Illustrative example of the generated data	29
Table 5. Cost to be considered in the economic analysis.....	35
Table 6. Financial Costs Possible Cost Drivers.....	39
Table 7. Appointment Schedule for Illustrative Example-2 Patients Overbooking.....	55
Table 8. Results of the Scheduling Policies Simulation.....	75
Table 9. Error Rates for Fully Grown Trees-Unknown Pattern Data Example.....	102
Table 10. Error Rates for Prune Trees-Unknown Pattern Data Example.....	103
Table 11. Error Rates for Fully Grown Trees-Patterned Data Example.....	105
Table 12. Error Rates for Prune Trees-Patterned Data Example.....	106
Table 13. Comparison Error Rates - CART Prune Trees vs. GBM Trees.....	108
Table 14. Replication Runs for Policy 1.....	110
Table 15. Statistics Results for the Replication Runs of Policy 1.....	110
Table 16. Replication Runs for Policy 2.....	111
Table 17. Statistics Results for the Replication Runs of Policy 2.....	111
Table 18. Replication Runs for Policy 3.....	112
Table 19. Statistics Results for the Replication Runs of Policy 3.....	112
Table 20. Replication Runs for Policy 4.....	113
Table 21. Statistics Results for the Replication Runs of Policy 4.....	113

FIGURE LIST

Figure 1. No-Show Rate in Studies Held in Clinics.....	5
Figure 2. Research Methodology.....	24
Figure 3. Representation of a Classification Tree.....	31
Figure 4. One Patient per Slot Representation.....	50
Figure 5. Overbooking Representation.....	51
Figure 6. Input Representation-Interactive Platform.....	52
Figure 7. Output Representation-Interactive Platform.....	53
Figure 8. Initial View-Interactive Platform.....	80
Figure 9. Steps for Scheduling an Appointment-Interactive Platform.....	90
Figure 10. Patient Assignment to Generate Appointment-Interactive Platform.....	81
Figure 11. Illustrative Example of a Confusion Matrix.....	92
Figure 12. Representation of the Folding Procedure.....	99
Figure 13. Example-Confusion Matrix and Performance Measures using Training Data for the Unknown Pattern Data Example.....	101
Figure 14. Example-Confusion Matrix and Performance Measures using Testing Data for the Unknown Pattern Data Example.....	101
Figure 15. Example-Cross-Validated Error vs. Tree Size Plot.....	101

1 CHAPTER – INTRODUCTION

The rise of healthcare costs has been a topic that has taken notoriety in the recent years. As Muthuraman and Lawley established [2], the environment of raising costs, limited capacity and an increasing demand for services has caused many clinics to shift from inpatient to outpatient facilities. Access to outpatient facilities is controlled through appointment scheduling, and tend to confront problems of patients non-attendance. **Non-attendance** to a clinical appointment can be described as the action of a patient failing to appear for a scheduled appointment [3]. Failed appointments can be divided in two groups: cancellations and no-shows. **Cancellations** are appointments that have been cancelled prior to the due time. Since these are known previous to the appointment date, the clinic can devise a strategy to use the available capacity. **No-show** is the term used to describe patients who fail to appear for the scheduled appointment without previous cancellation. Of the two groups, no-shows tend to have the greatest impact in healthcare operation systems because its occurrence is not known in advance, which results in an immediate loss of capacity.

No-shows can result in challenges in determining appropriate staffing levels, affecting productivity and efficiency due to the under-utilized clinical capacity, among other consequences on operational aspects of a clinic. Also, it is argued to be a problem that has consequences for the cost of healthcare due to their effect in social and financial costs [3]. The study of patient's attendance can help in mitigating healthcare costs by reducing inefficiencies [4]. Throughout the years, studies and investigations have been held to find methods to reduce failed appointments, such as reminder systems like phone calls, emails, among others [9]. However, non-attendance has not been eradicated. Researchers have demonstrated the use of patients' demographic attributes and characteristics of the

appointment system as a way to characterize patients' susceptibility to fail a clinical appointment [4, 5, 7, 8, 16, 17, 19-24]. Then, that information has served as an input for the construction of the appointment schedules, providing satisfactory results [5, 19, 25, 26]. Recently, one of the questions that have been raised regarding this issue is related to how to estimate, as accurate as possible, the real cost of a no-show to the system [3]. Clinical appointment schedules are structured by slots of fixed or variable time length. Its construction can be as simple as the assignment of one patient per slot or as complex as assigning multiple patients per slot, which is known as **overbooking**. In the simplest form, the cost of the physical resources and personnel that are used during the time interval of duration can be used to estimate the cost of the slot. However, when overbooking occurs new aspects emerge, as waiting time and personnel overtime, for example. When multiple patients arrive for the same appointment slot, appointments begin to fall behind; a chain effect occurs across all the slots and new direct and indirect costs arise as a consequence. This is the reason why it is not desired to schedule in the same slot patients with a high chance of attending to the appointment.

This document presents the development of a cost model that integrates patients' probabilities of no-show and cost information to determine the cost expected value of an appointment slot as a representation of the estimation of the cost of a no-show to the system. A methodology has been designed in order to achieve this objective. It includes using a Classification Trees approach to construct a classification model, with the purpose of predicting the patients' no-show probability. Also, it involves performing a cost analysis including financial costs that can be directly estimated and social costs which add complexity to the analysis because they have to be indirectly estimated and allocated. The predictions

(probabilities) from the classification model and the outcomes of the cost analysis are integrated in a cost model that estimates the expected cost of an appointment slot for a scheduling scheme constructed using overbooking. Then, this stochastic cost model is used to assist in the process of evaluation of different scheduling schemes, with the purpose of identifying possible scenarios that perform better in terms of the total no-show cost to the system. Its application in real life would benefit outpatient practices during the scheduling process; this is demonstrated through the creation of a prototype for an interactive appointment scheduling platform in Excel.

1.1 Significance

Outpatient clinics offer services of medical procedures or tests that can be done in a medical center without an overnight stay. Usually they cost less because of the shorter length of stay, which implies the elimination or reduction of several procedures and costs associated when a patient is hospitalized. In general, outpatient clinics can be divided in four service categories: Wellness and Prevention, Diagnosis, Treatment, and Rehabilitation. Wellness and Prevention centers focus in the orientation and application of preventive medicine. Diagnosis centers include radiology services and laboratory tests, among other services. The Treatment centers operate minor surgical interventions and provide treatments for specific medical conditions such as cancer. Rehabilitation centers provide services to treat patients with psychological conditions, patients who require physical therapy, among other services. It is important to clarify that an outpatient clinic can provide in their facilities more than one of the services described. Since outpatient clinics work at an appointment basis, they tend to be impacted by the absences of patients.

Non-attendance to clinical appointments implies a negative impact in the healthcare systems because it has consequences in costs, productivity, resources utilization and patients' flow, among others [5-8]. Besides the economic impact of non-attendance, other aspects of the operation of the system are affected by failed appointments [5]. Failed appointments interrupt the flow of patient care and the clinic productivity declines [6]. Also, no-shows cause relative longer waiting periods for appointments [7]. Scheduling conflicts and interrupting continuity of care are other effects of failed appointments [8]. Loss of available clinic capacity and higher times of the patients waiting on the clinic are examples of social consequences due to non-attendance, which can also be expressed in monetary terms as social costs [3]. As it can be seen, failed appointments may have an impact in financial, social and medical aspects.

By 1979, studies about hospital clinics with low socioeconomic populations presented appointments fail rates between nineteen and twenty eight percent and studies about family practice centers reported appointments fail rates which vary from five to eleven percent [10]. In 1999, Hixon et al. [11] performed a survey study on family practice residency programs. In the results, the authors show that more than one third of the clinics that answered the survey reported a no-show rate of more than twenty one percent. According to a study held by Moore et al. in 2001, no-shows and cancellations represented 31.1% of scheduled appointments and 32.2% of scheduled time [6]. This problem extends to today since, in a study published in 2014 by Lofti and Torres, held on a physical therapy clinic, clinical records showed that the typical monthly no-show rate was approximately 16% for all patients and 21% for new patients while the cancellation rate was approximately 22% for all patients and 27% for new patients [5].

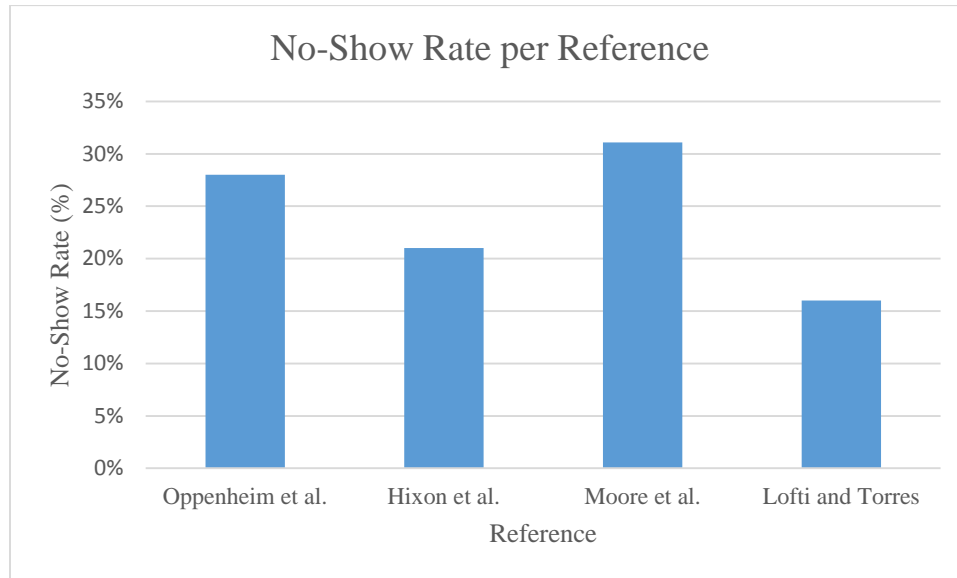


Figure 1. No-Show Rate in Studies Held in Clinics

In 2007, the healthcare industry expenditures represented approximately 15% of the gross domestic product (GPD) of the United States, a figure that increased to 17.4% in the year 2013 and is projected to increase 1.9% by 2023 [1]. According to the Centers for Medicare and Medicaid Services, the National Healthcare Expenditures (NHE) grew 3.6% from 2012 to 2013, accounting to 2.9 trillion of dollars, and is projected to grow at an average rate of 5.7% from 2013 to 2023 [1]. The increase in expenditures can be ascribed to the influence of several factors such as the introduction of new expensive technologies to the system, unfavorable trends in population demographics and legal expenses, among others [2]. In economic terms, non-attendance has impacted public expenditures around the world. The United Kingdom National Health Service (NHS), in 1984, reported a cost of up to 266 million pounds caused by broken appointments, and the Department of Health reported that as much as 360 million of pounds would be wasted each year [12]. From 1996 to 1997 the UK accounted for an average of 366 million of pounds due to missed appointments [12]. In the United States, information can be found about lost due to non-attendance in healthcare

clinics. In 2001, Moore et al. showed that total revenue shortfalls could range from 3 percent to 14 percent of total clinic revenue [6]. David Keefer, a health system specialist at the Lyster Army Health Clinic, expressed that “no-shows” cost the clinic \$450,000 to \$900,000 per year in lost revenue [13].

Outpatient clinics have used several methods to decrease non-attendance rates, for example, mailed and telephone reminders [9]. Also, they have implemented other techniques to decrease the impact of non-attendance. The use of no-show probability, extracted from data bases, as an input for scheduling appointments, has been researched and implemented through case studies. One of the disadvantages of this method is the fact that most of the approaches include overbooking, which involves scheduling an additional fixed number of patients each day based on the clinic no-show rate. It is known that overbooking can cause increases patient wait times, which impact patients experience in the system, and can cause provider overtime, which also impacts the clinic profit [14]. Since the magnitude of the number of additional patients to be scheduled depends on the clinic no-show rate, it is important to have a robust methodology and framework to assess the no-show probability. Lofti and Torres express in their article:

“Hence, a reliable procedure is needed to estimate the probability of show for patients who are scheduled too close to one another which is a form of overbooking. This will allow the scheduler to avoid overbooking patients with a relatively high probability of show. It will also allow the scheduler to implement a targeted overbooking of patients with high probability of no-show, minimizing the expected number of unused timeslots due to no-shows, thereby increasing the overall utilization of the Clinic.” [5]

The fact that non-attendance is a current problem in outpatient clinics emphasizes the need of research in this area in order to use the information available to assess the status of the system and identify areas of opportunities for improvement. In future research, that information could be used to strategize and develop better practices to counteract their impact in the system.

1.2 Justification and Contribution

Several efforts have been made within the research community by using historical information in the databases of healthcare outpatient centers or semi-structured interviews, with the purpose of identifying the probabilities of no-show of patients. Their main objective is using that information as a source of knowledge to use for many applications, such as using it as an input to determine an efficient medical appointment schedule. For this reason, many researchers have investigated possible factors that influence whether a person is sensitive to fail an appointment with or without prior notice. Besides the efforts, however, in the relevant literature regarding this topic one could identify the following areas of opportunity:

1. It is not evident that previous publications perform a deep investigation about the relationship (internal dependency) among factors and have not presented a general framework that can be applied into different scenarios.
2. Few studies considered or focused on exploring the costs of non-attendance; the ones that have been published are not as deep and extensive as needed.
3. Few studies considered the assessment of non-attendance, using no-show probabilities and cost information as a way to describe the expected cost of a failed appointment slot, and use that information as input in the scheduling process.

This research focuses in developing a general methodology that will overcome the areas of opportunity mentioned before. Specifically this work:

1. Develops a cost model that integrates patients' probabilities of no-show and cost information to determine the expected value of an appointment slot as a representation of the estimation of the real cost of a no-show to the system.
 - a. The cost model will integrate cost information in two major cost categories, the financial costs and the social costs. Few publications present research estimating the social cost of a patient absence. The ones that include this cost category estimate it from the patient point of view. On the contrary, in this work it has been identified allocation basis to estimate the social costs from the clinic point of view to determine the indirect economic impact in the clinic of a patient absence to a medical appointment.
2. Evaluates different scheduling scenarios in terms of costs by using stochastic cost model developed.
3. Demonstrates the cost model usefulness through the construction of an interactive platform that allows the scheduler to choose the attributes of the patients to calculate their probabilities of no-show and the estimated cost of the appointment slots when patients are assigned.

1.3 Thesis Organization

This thesis is organized in chapters. The second chapter consists of a literature review regarding the most relevant recent published articles related to the topic under research. A description of the methodology followed is given in the third chapter. The fourth chapter

presents a detailed explanation of the results. General conclusions, additional comments, and recommendations for future research in this area of interest are discussed in the fifth chapter.

2 CHAPTER – LITERATURE REVIEW

Non-attendance to clinical appointments has been a subject investigated for many years by the academic research community interested in healthcare operations. The work done about this topic can be divided in three main categories. The first category can be titled as Predictors & Modeling. Other research articles can be categorized as Scheduling Applications. The last category of research in this topic can be titled as Financial Considerations. A summary of the relevant literature related to these categories is presented next.

2.1 Predictors and Modeling

Researchers under this category have concentrated their efforts in the investigation of the factors that significantly influence patients to fail an appointment, they use this information as predictors to determine the probability of no-show of a patient [4, 7, 8, 15-18]. Their methodologies vary from structured interviews to patients, to retrospective studies using information from the data bases of outpatient clinics. Using statistical tools and techniques, researchers have constructed predictive models to determine the probability of no-show of a patient.

2.1.1 Factors Selection

Many reasons or factors that contribute to missed or broken appointments have been pointed out by researchers in this field of investigation. One of the most complete literature reviews in this topic was held by Deyo and Inui [18] where they present a table of “determinants” of broken appointments. The authors classified the “determinants” in eight categories: features of the patient, features of the medical

provider, features of the disease or reason for appointment, features of the patient-provider interaction, features of the therapeutic regimen, features of the medical facility and administrative process, features of access to the facility and environmental factors. They identified nearly eighty-six factors or reasons that can contribute to dropout and broken appointments. From the factors identified by the Deyo and Inui, only those related to demographic features of the patients and features of the appointment scheduling process are being considered in this work.

The majority of the studies related to this topic have been performed through data analysis from adult medical clinics, psychiatric and pediatric populations, where most of the publications until 1980 focused their work [18]. Interviews have also been used as a method to gather information. As Deyo and Inui [18] present, many investigators have presented results of studies held by interviewing patients by telephone, by mailed questionnaires, or in person. Patient forgot, did not know about the appointment, or misunderstood are the most frequent reasons given by the patients. Other reasons related to transportation, economic situation and time or work conflicts, also resulted significant. In [15], Lacy et al. held a semi-structured interview study on a university-affiliated family practice clinic. The purpose of the interview was to identify the patient perception about this topic. Results revealed three major issues related to missing appointments without previous notification. The first one is emotional barriers; according to the patients, the negative emotions about going to see the doctor were greater than the perceived benefit of it. The second reason is perceived disrespect from the healthcare staff. The last issue is related to the patient lack of understanding of the scheduling system; the patients seemed

unaware of the impact of missed appointments. As Deyo and Inui established [18], it is important to consider that the validity of data and results from interviews is open to questions because it depends on the respondent's attitude towards the interviewer. Also, the sample interviewed may or may not represent the population perception about the topic. The positive aspect about interviews is that people's opinion can be known, however it can be incorrect to use results from the interview as definitive or the norm.

Not necessarily the same factors are considered in all publications presented, also not necessarily all of them present the same results; it can vary among the scenarios. However, in general, several factors have resulted to be consistently significant in most of the articles reviewed. Age, socioeconomic status and history of non-attendance have been the factors that consistently resulted significant in the majority of the articles published [4, 5, 7, 8, 16, 17, 19-24]. In general, patients who are sensitive to fail an appointment tend to be young, of lower socioeconomic status or unemployed and with a previous history of non-attendance. In addition, age has also been relevant in relation to patients' calling to make an appointment cancellation [7]. Deyo and Inui [18] confirmed in their review that, in general, the majority of the studies present that age, education and socioeconomic status are probably the most consistent and important demographic reasons on appointment-keeping behavior. In the case of patient's mode of payment, it has been found important in some studies but not in others, the significance of this factor depends on the setting and the scenario under analysis [4, 17, 22, 23]. Factors like gender [5, 8, 22] and time between scheduling and the actual appointment date (lead time) [4, 5, 8, 15-17, 19, 22-25] have

resulted as relevant in several publications. Other factors such as transportation, waiting time, misunderstanding in scheduling, patients' lack of knowledge about the process, among others, have been mentioned in literature but with less relevance. It can be seen that factors related to demographic features of the patients and features of the appointment scheduling process tends to be more relevant in relation to patients non-attendance. This is the reason why this work focuses on them. The consideration of factors and their significance is dependent on the scenario (the particular characteristics of the system being analyzed) and the availability of information.

Dependence between factors is important to be considered; for example, the patient's payment method can be strongly related or influenced by several demographic factors like age and socioeconomic status. In summary, Deyo and Inui literature review reveals two important characteristics that need to be considered in the analysis: interdependency between factors and that factors significance can vary among scenarios. In this work, in order to deal with the possible interdependency between factors, a classification technique called Classification and Regression Trees is used because provides prediction models that are not affected in the presence of multi-collinearity. A literature review approach has been considered to identify those factors or attributes that consistently resulted as significance in previous studies held on healthcare clinics.

As mentioned before, non-attendance can be divided in two groups: cancellations and no-shows. Of the research literature, in most of the cases, cancellations and no-shows are grouped together and studied under one category usually called non-attendants, or only no-shows are considered in the analysis. Cancellations are

important for service operations because a clinic could devise a strategy to recover the capacity, something that cannot happen with no-shows [4]. This is the reason why Norris et al. [4] performed an empirical investigation of the factors affecting patient cancellations and no-shows at outpatient clinics. They examined patient attendance using three discrete outcomes: no-shows, cancellations and patients who arrive. In [7], Shaparin et al. also considered cancellations in their analysis. From their results it can be highlighted the fact that the same factors can be studied for no-show and cancellations, however not necessarily the same factors will result significant in both categories. This reinforces the necessity of considering and analyzing cancellations apart from no-shows.

2.1.2 Use of Data Base Information and Statistical Modeling

Once the possible factors that influence whether a person is sensitive to fail an appointment without prior notice have been investigated, statistical screening has been essential in order to consider the variables or factors that show statistical importance with respect to the response; that in this case is known as patient sensitivity to fail a medical appointment. In general, the majority of the studies published use statistical techniques to determine the importance of the factors and for the construction of predictive models in order to calculate the probability of no-show. Lacy et al. [15] is an exception because the researchers used semi-structured interviews to patients in order to obtain the data and used an immersion-crystallization organizing style to analyze the data. Dove & Schneider used an interactive computer language for statistical analysis on data analysis and variables screening [16]. Predictive modeling with multinomial logistic regression and decision trees has been

techniques used in several studies [4-19]. For a study held on an academic pain clinic in Newark, New Jersey, researchers used 2-sample t-test for continuous variables comparisons and Pearsons chi-square test or Fisher exact test for categorical variables. They modeled the data using a marginal logistic regression [7]. Lee et al. [8] used univariate analysis to determine variables significance and multivariate analysis with a multiple logistic regression to construct a prediction model. Also, they used a receiver-operating characteristic (ROC) curve to assess the model's discrimination ability. Artificial Neural Networks have also been used to construct models and predict the risk of no-show for a patient appointment; Dravenstott et al. [17] accompanied the technique with a sensitivity analysis with the purpose of eliminating the variables that were not sensitive. Lofti and Torres used Classification and Regression Tree (CART) for data analysis, as we do, they compared the results with the ones obtained of a Bayesian Network and Neural Network models [5].

The literature reviewed reveals that Dove and Schneider [16], in 1981, were ones of the first researchers to use a tree model approach to develop a predictive model of no-shows for outpatient clinics. They used the patients' individual characteristics as variables to predict the number of patients who could be expected to keep their appointments. The statistical analysis revealed that only four variables (patient's age, travel distance, appointment interval, and previous no-show record) were significant in relation to the binary dependent variable patient's last appointment. Those four variables were considered under a decision tree structure to develop a predictive model to estimate the expected number of patients who will show out of all patients. The purpose of their model was not to make a prediction for each individual patient,

as is intended in this research. Their decision tree is not complex and the fact that only four variables are considered makes the problem simpler. Lofti and Torres [5] expand Dove and Schneider work with a more complex approach using Classification and Regression Tree (CART), assessing the relative predictive powers of four different decision tree techniques; which at the same time are compared with the predictive power of models obtained from Bayesian Networks (BN) and Artificial Neural Network (ANN) models. The authors compared the four decision tree techniques using measures as sensitivity (portion of patients correctly classified as show), specificity (portion of correctly classified as no-show patients), and risk estimate (overall portion of the cases that are misclassified). Decision tree analysis exhibited superior performance when compared with those of BN and ANN.

2.2 Scheduling Applications

Publications under this category have gone more far in their investigation, studying the use of no-show modeling as a tool to improve clinic performance [5, 19, 25, 26]. These researches utilize the no-show probabilities obtained from the models as an input of advanced scheduling methods.

Considering no-show probabilities in patients' appointments scheduling has been an approach used in order to reduce the impact of no-shows in clinic efficiency. The scheduling approach includes two methods: overbooking (the most used) and short lead time scheduling [19]. Overbooking, which consists in booking multiple patients in the same appointment slot, can result in a negative impact because it can cause clinic personnel overtime and higher patients' waiting time, among other consequences. The short lead time scheduling minimizes the time between the appointment making and the appointment date. This method has

worked for some scenarios, but not for others. There is a current need for research in this area, starting with the assessment of the input data needed and finishing with the development of new scheduling methods. Most articles about this topic, before doing the schedule, show the use of factor screening and no-show model construction because they provide part of the input needed to develop the scheduling methodologies or algorithms. Several research efforts related to this topic can be found in the academic literature. Most of them are from 2009 to the present. It is a relevant issue to what is happening in the healthcare systems.

Daggy et al. in [19] demonstrated the utility of using no-show probabilities as an input for appointments scheduling by comparing a regular schedule with a schedule developed using the Mu-Law method. The Mu-Law method use no-show probabilities, service time, slot length information in conjunction with cost and revenue information to assign appointment slots to patients. The objective of the method is to optimally balance patient waiting times, clinic overtime and revenue. It is a dynamic method that can overbook or leave slots unassigned because it is dependent of the sequence of appointment calls. The algorithm stops when the increment in marginal costs by the addition of one patient is higher than the marginal revenue [19]. To feed the algorithm, in this article, the authors performed a logistical regression to determine the no-show probabilities. Results from the study revealed that, in terms of physician utilization and overtime, the Mu-Law based schedule performed better. However, in terms of patients' waiting time, the regular schedule performed better. It was expected because the Mu-Law method tends to overbook. Also, considering the dynamic call in sequence of patients and their probability of now-show, Tsai et al. [26] presented a stochastic appointment scheduling system with multiple resources. In their algorithm, the authors considered a fixed number of slots of equal length, the probability

of no-show and the fact that in the visit the patient can require more than one resource. The no-show rate was used to classify the patients in different classes. The algorithm is designed with the purpose of maximizing total profit, considering patient's waiting cost and physicians' overtime costs. The stochastic scheduling system developed by them was then compared with traditional scheduling systems and with systems that consider no-show rate in a homogenous way; it performed better in profitability terms.

Lofti and Torres, in [5], used the attendance conditional probabilities obtained from the CART approach they applied to assist with scheduling. They developed an algorithm considering time slots, attendance probabilities and the expected number of patients schedule in timeslots. The authors analyzed five scenarios, from a regular one-slot per patient schedule to several patients per slot schedule according to their attendance probability (which allows overbooking). As a result, the scenarios that allow overbooking performed better in terms of number of timeslots with no-shows and in terms of clinic's utilization. However, the authors do not consider or discuss the effects in patients waiting time caused by overbooking. In [27], Tang et al. present two approaches to develop an appointment schedule considering no-show probability. They propose an exact deterministic service time method to find an optimal schedule. They also propose a heuristic algorithm, considering exponential distributed service time, which provides a local optimal solution to develop a schedule. They consider two types of patients: routine patients and urgent patients. The algorithms were designed to minimize the average patient waiting time, the physician idle time and overtime.

The methodology developed in this work consider no-show modeling, as the publications mentioned previously do, but also include financial information, all integrated in a probabilistic cost model. This cost model contribute in assessing the cost of a no-show

to an outpatient clinic, which allows compare different scheduling policies that may or may not include overbooking in order to identify a better procedure to schedule patients appointments. The scheduling policies are compared in terms of clinic utilization, level of overbooking and cost, among other performance measures.

2.3 Financial Considerations

Researchers under this category have focused their studies in the economic impact of non-attendance [3, 6, 12, 28]. The vast majority of research in appointments non-attendance considers cost impact in their analysis. However, very few perform a deep economic analysis considering all the cost components involved, instead a basic analysis is presented. As Bech established, few studies have explored this topic category and many of the studies done are rather unsophisticated or even misleading [3]. Areas of opportunity for research in this topic are considered.

Few studies consider or focus on exploring the costs of non-attendance; the ones that have been published are not as extensive as required. For the year 2005, Mickael Bech, in "The economics of non-attendance and the expected effect of charging to fine on non-attendees" [3], establishes that very few studies have explored the costs of "non-attendance". In addition, he indicates that the majority of the studies made in previous years were unsophisticated or confusing. Even the majority of studies and estimates were not up to par with the standards of an economic analysis, the majority over-estimated in its techniques. However, these studies provide an idea of the magnitude of the economic impact that the absences to medical appointments cause in healthcare centers. An extensive and complete cost analysis is required in order to fully assess the cost/benefit balance.

After 2005, very few studies on the economic effects of the absences to medical appointments can be found. Within the most recent is "Impact of missed appointments for outpatient back on cost, efficiency and patients' recovery" [28]. In this study a basic cost analysis considering gains and losses is performed, and even within its own set of limitations it clarifies that the estimated value of the cost per treatment used does not represent the actual cost.

As established by Michael Bech in [3], the costs of non-attendance to medical appointments can be divided into two types of costs, which in part overlap. These are the social costs and the financial costs to their healthcare providers. Within the social costs are the value lost by the non-utilization of resources, resulting in low productivity and loss of benefits. It also includes the staff time not used, the equipment not used and the utilities this entails, and the "good-will lost" because patients can wait more because there is a tendency to "over-scheduling" for counteracting the effects of the no-show. The failure to redistribute resources vacancies by the absence of the patients is also a social cost. Opportunity cost is a significant social cost; it can be seen as the benefits that could have been generated if the time and the resources were used in other activities. In addition to the already mentioned, the delay in waiting times can cause serious clinical results because it is known that the effectiveness of treatments depends on the time frame between the diagnosis and treatment. As Deyo & Inui established, few studies have attempted to show increased hospitalization rates for those who break appointments [18]. There are social costs associated with this aspect, which should be investigated. The social costs are often ignored by healthcare providers because they do not necessarily affects them directly, however, they can have an indirect impact. This work incorporated both cost categories in the development of the

probabilistic cost model. The financial costs are strictly considered as the operational costs directly related to the service provided to the patient. In relation to the social costs, this work studies the “loss of good-will” cost, referred in this document as the waiting cost, and also consider the cost of the value lost due to the non-utilization of the resources.

The financial costs to the healthcare providers are related to the loss of income due to the absence of the patient, given that when warning ahead from the scheduled patient, the providers could not fill the space with another patient. As is well known, the medical providers (doctors) give a service, which is then reimbursed economically by the well-known "third party payers" or health insurance companies. In addition, the medical insurance may not cover the entire operation to the patient and therefore the patient has to pay a deductible for the service. The providers that are paid in charge for their services or by schema of cases are the ones most affected. It is important to consider that in many occasions a no-show does not mean a financial loss due to the "walk-in" patients that can fill that space. "Walk-in" is patients seeking appointment the same day, and as the space is empty by the absence of the no-show, the slot can be filled immediately. Researchers who published the article "Time and Money: Effects of No-Shows at a Family Practice Residency Clinic" [6] explored the balance between the no-show and the “walk-in” patients. Within their findings is that the "walk-in" patients generate less revenue than the one that could have been generated with the no-show patients that were in the itinerary. Additional to that, there may be loss of productivity and resources, among others, because, if for example, the patient in the itinerary that failed the appointment was cited for a procedure of thirty minutes, and the "walk-in" patient requires a procedure of fifteen minutes; there are still fifteen idle minutes that impacts in a negative way.

In general, we can say that the majority of investigations related to the economic impact of failed appointments focused on the financial costs, for the purpose of presenting the economic impact on the income by the absence without prior notice of the patient. Berg et al. consider social costs in their discrete event simulation study conducted for an outpatient endoscopy clinic [29]. They evaluated different scheduling policies and different overbooking levels, in order to determine the schedule that maximize the net gain of the clinic. In the study they included the traditional operational costs, estimated as a percent of the clinic reimbursement, and also included the cost of no-show, the overtime cost and the patients' waiting cost. The overtime cost was estimated based on expert opinion. The cost of no show was quantified based on the difference between the revenue obtain from the procedure and the operational costs of managing the procedure, which they call net gain. Finally, they quantified the patient waiting cost in terms of the average wage the patient loose by waiting on the clinic instead of being productive at work. It can be said that the authors are accounting this social cost from the patients' point of view, and not from the clinic point of view as will be seen in this research work. The authors simulated the scenarios using general statistics about the patients' attendance in a length of time. Their methodology do not focus in the assignment of each individual patient in the appointment schedule. They do not consider each patient susceptibility (probability) to fail an appointment in order to use that information for the assignment, as will be presented in this research work. To the best of our knowledge, there has not been more research or publications that conducted a more in-depth study (considering social costs, financial and opportunity costs) on the economic impact of patients' non-attendance.

Few methods or heuristics consider the economic factor as a determinant aspect at the moment of making the schedule of appointments, once you have the probability of no-show. The Mu-Law algorithm, a method constructed, developed and published by Muthuraman and Lawley, considers the probability of no-show of patients but it also calculates the expected income for each block of time and accommodates the patient in the block of time that maximizes profit [2]. They constructed an optimization model for the call-in scheduling problem with a profit maximization objective function. The objective function presented by the authors is unimodal, meaning that is non-decreasing until a peak is reached and then is monotone decreasing afterwards. This implies that the optimization model assign until the cost of assigning one patient outweigh the revenue generated, which can be seen as a stopping criterion. It is a probabilistic optimization model because it considers patients' probability of no-show and the probability that the assignment of a patient to a slot may increase the overflow from one slot to another. Sands et. al [19] used the Mu-Law algorithm in their investigation and by the results it is presented that the optimization model leave some slots unassigned, something that can be seen as a disadvantage. Even when the Mu-Law methodology performed better than the one-patient per slot scenario, it is important to highlight the fact that it is designed for scheduling construction and not for the evaluation of different possible scenarios. Although it has an economic based objective function, the authors do not present a methodology to identify, define, assign and allocate the direct and indirect financial and social costs involved. Also, it does not provide a cost model that could be applied to evaluate different scheduling scenarios, which is the purpose and contribution of the methodology that will be presented in the following section.

3 CHAPTER – METHODOLOGY

A methodology has been developed with the purpose of structuring a plan to accomplish the objectives of this research. It consists of a series of steps that will be explained in detail in the following sub-sections. Figure 2 presents an illustration of the research methodology.

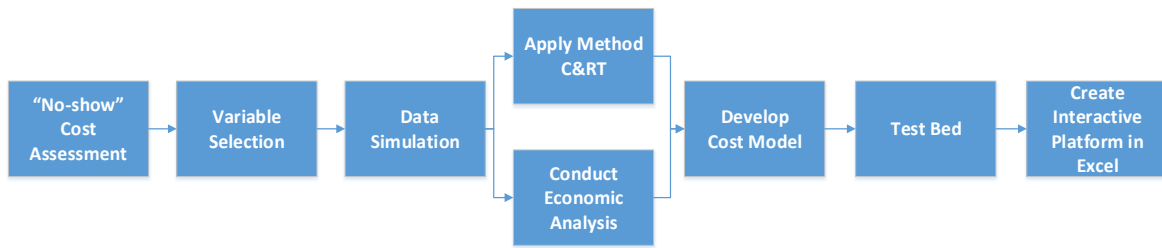


Figure 2. Research Methodology

3.1 Variable Selection

As mentioned in Sub-section 1.1.1, several research publications have presented studies related to the attributes associated or significant in relation to patients' non-attendance to medical appointments. A literature review of the recent publications related to this topic has been conducted with the purpose of identifying the attributes that has resulted constantly significant in relation to predict a patient susceptibility to fail an appointment. Twelve published articles, in total, under the category of "Predictors and Modelling" were reviewed. Table 1 presents a sample of seven of the twelve published articles reviewed, showing the overall factors considered in those publications, highlighting the factors that resulted significant from the statistical screening. The fact that an attribute is considered in a study does not imply that it has to result statistical significant, that is the reason why the data is analyzed using a screening technique. As it can be seen, since the publications present some

variables or factors in common, it can be said that the significance of factors may vary among scenarios.

Table 1. Summary of overall relevant factors found in seven publications

Factor	Article						
	Norris et al.	Lofti et al.	Shaparin et al.	Lee et al.	Dove et al.	Sands et al.	Dravenstott et al.
Age	X	X	X	X	X	X	X
Appointment Length							X
Continuity of care(office visit number)			X				
Date of Last Visit		X					
Distance		X	X	X	X	X	X
Education		X					
Gender		X	X	X			X
Having an interventional procedure scheduled and performed in connection to the appointment			X				
Hospital admission during the appointment or between scheduling and appointment				X			
Hospital Department				X			X
Lead Time	X	X		X	X	X	X
Marital Status						X	X
Number of days since patients' last completed appointment						X	X
Particular complaint			X			X	
Patients' Employment Status		X	X				X
Patients' length of time being seen in the health system							X
Patients' Primary Language			X				
Prior attendance history	X	X		X	X	X	X
Provision of cellphone number				X		X	
Race		X	X	X	X		
Referring Physician			X				
Schedule (Day of week, hour, time of the day)	X					X	X
Season						X	X
Service connected disability: priority given to patients with disease or disability that occurred during active duty					X		
Type of Payment (Insurance)	X		X			X	X
Type of Visit					X		X
Weather	X						

Factors marked with an X were considered in their respective analysis; the ones highlighted resulted significant after the statistical screening.

Since a considerable number of attributes were identified in the literature, it was convenient to organize them in categories. To accomplish that, the twenty-one attributes were categorized in six of the eight “determinants” of broken appointments as established by Deyo and Iniu [18]. Table 2 present the attributes categorized and show (highlighted in gray) in which of the twelve published articles those attributes resulted statistical significant.

Table 2. Attributes categorized according to the “determinants” of broken appointments

Factors	Research Articles											
<i>Demographic Attributes</i>	[4]	[5]	[7]	[8]	[16]	[17]	[19]	[20]	[21]	[22]	[23]	[24]
Age												
Education												
Gender												
Marital Status												
Patients' Employment Status												
Patients' Primary Language												
Race												
Socioeconomic Status												
Type of Payment (Insurance)												
<i>Access Attributes</i>												
Cellphone number												
Distance												
<i>Facility and Administrative Process Attributes</i>												
Appointment Length												
Lead Time												
Hospital Department												
Number of days since patients' last completed appointment												
Continuity of care(office visit number)												
<i>Environment Attributes</i>												
Schedule (Day of week, hour, time of the day)												
Season												
<i>Sociobehavioral Attributes</i>												
Prior attendance history												
<i>Other Attributes</i>												
Having an interventional procedure scheduled and performed in connection to the appointment												
Particular complaint												

The number of times a factor is mentioned in the literature as significant has been used as the criteria of identification because each author has used different techniques for statistical screening and prediction model construction. Not all the publications provides an attribute ranking that inform about the importance of an attribute in comparison with the others. Analyzing Table 2, it has been identified that thirteen attributes were consistently mentioned as significant across most of the articles. These are related to demographic factors, scheduling factors, and prior attendance history. Demographic attributes refer to particular characteristics of the patient, such as gender and race. The scheduling attribute is related to

the lead time; the time from making the appointment to the appointment date. Prior attendance history depends on the patients' previous no-show record. Table 3 show the final fourteen attributes identified by the literature review.

Table 3. Final fourteen significant attributes identified

No-Show Attributes
Age
Education
Gender
Marital Status
Patients' Employment Status
Patients' Primary Language
Race
Socioeconomic Status
Type of Payment (Insurance)
Provision of Cellphone Number
Distance
Appointment Length
Lead Time
Prior Attendance History

Since the purpose of this research is to provide an output that could be used in outpatient clinics, it is important to consider variables for which the clinics have information or can be accessed in their databases. The literature review has served as a method to determine the variables to be considered in the study.

3.2. Data Simulation

The reviewed literature related to this topic focus their research scope to particular settings or scenarios. Using data base information from healthcare systems, researchers have applied statistical tools to screen the variables with the greatest impact and for model construction. Norris et al. [4], Lofti and Torres [5] and Shaparin et al. [7] used data from outpatient facilities from Medical Schools departments in their respective analysis. Dravenstott et al. [17] used Primary Care and Endocrinology data set from a healthcare

system. Dove and Scheneider work [16], focused on a sample data collected from a Medical Center in the United States. Veterans' outpatient clinics from the United States have also served as scenarios under study [19]. A study with data collected from an outpatient clinic in Singapore, have also been published [8].

In the case of this research study, simulated data is being used to apply the methodology developed. The purpose is to show how the procedures resulting from this research could be replicated in a real scenario of an outpatient clinic. The attributes and variables for which the data has been generated were identified and selected based on literature review, also taking in consideration the information that could be accessed or collected in real life. As would be seen in the next chapter, several cases will be analyzed in order to take in consideration the possible scenarios that could be encountered in an outpatient clinic environment. Details about how the data was generated is presented next.

A total of thirty sets of data were generated in Excel for each example. Each data set contains five hundred samples for each of the attributes under consideration. Those five hundred samples were then divided in five folds because in the analysis a cross-validation approach is being used in order to qualify the prediction model and to assert the level of precision we can obtain from it. From the thirteen possible significant no-show attributes identified, which are mentioned in Sub-section 3.1, only eight of them were selected as the predictive attributes for the examples: age, distance, race, lead time, type of patient, marital status, primary language and gender. It was identified from a real outpatient clinic that information about these eight variables are usually available in the data bases of the system, and the idea of this research is to present a methodology that could be used with the information that is already available in healthcare facilities. The response variable considered

is binary, taking a value of {0} for Show and {1} for No-Show. The five hundred samples of each data set were generated creating half of them with a Show response and the other half with a No-Show response. Table 4 presents an illustrative example of the data sets generated in Excel, showing the first 30 samples. Each row of the data sets represents a patient medical appointment. The first column is the patient ID. Columns two through ten contain the attributes information for each particular patient. The last columns contain the binary response, indicating a {0} if the patient showed to that particular appointment, and {1} if not.

Table 4. Illustrative example of the generated data

ID	l time	a length	age	gender	race	m status	p language	distance	t patient	attendance
1001	9	25	36	A	A	A	B	14	B	1
1002	2	11	48	A	A	A	A	12	A	1
1003	12	12	47	A	C	A	A	6	A	1
1004	27	26	33	A	B	A	A	0	A	0
1005	18	16	63	B	A	B	A	12	A	1
1006	4	21	59	B	C	B	A	12	A	0
1007	11	28	44	B	A	A	A	6	B	1
1008	27	10	60	A	A	B	A	14	A	1
1009	20	15	62	A	B	A	B	6	B	1
1010	30	10	65	B	A	A	B	19	A	0
1011	1	28	41	A	C	A	C	8	A	0
1012	20	24	26	A	A	B	B	19	A	1
1013	17	22	23	B	B	B	B	0	B	1
1014	14	23	22	B	A	A	A	0	A	1
1015	26	28	46	B	C	B	C	14	A	0
1016	1	24	35	B	B	A	B	6	A	0
1017	23	18	69	B	A	A	B	14	A	1
1018	30	14	22	A	C	A	B	8	B	0
1019	26	14	35	A	C	B	B	19	A	1
1020	14	26	40	B	B	A	A	14	B	0
1021	6	23	57	B	B	A	B	17	B	1
1022	1	27	62	A	C	A	B	0	B	0
1023	25	21	68	B	C	A	B	14	A	0
1024	18	27	69	A	C	B	B	6	B	0
1025	21	29	53	B	A	A	A	6	B	1
1026	24	30	53	A	C	B	A	19	A	1
1027	1	20	26	B	C	B	B	8	A	0
1028	27	11	70	A	C	A	A	14	A	1
1029	2	12	57	A	C	A	A	12	A	1
1030	5	22	23	B	B	B	A	12	A	0

3.3 Classification and Regression Tree (CART)

Classification tree method is the selected technique to assess the no-show to clinical appointments through the calculation of no-show probabilities. Considering a combination of continuous and categorical variables, it is pretended to construct a model to predict a binary

dependent variable, which is the classification of a patient as a show or a no-show, and to obtain a conditional probability.

Classification is the task of using a set of attributes (x 's) to assign objects to one of several predefined classes (y 's) [30, 34]. It is a tool that helps to distinguish between objects of different classes. As in the regression setting, the classification setting counts with a set of training observations (records) that can be used to construct a classification model [31, 35]. The main difference between both settings lies in the fact that regression methods assumes that the response variable (Y) is quantitative, while classification methods assess the situation where a response variable is instead qualitative.

Classification Tree is a Decision Tree-Based Classification Method which consists in the segmentation of the data, better known as predictor space, into regions in order to make a prediction or classify a given record (observation) into a class (response) [35]. As a classification method, it is used to predict a qualitative response. Its methodology consists of a series of splits of the data in order to stratify the predictor space. The splitting process can be visually represented as a tree (Figure 3). Basically, this technique classify each record by predicting that it belongs to the most commonly occurring class of training observations in the region to which it belongs [35].

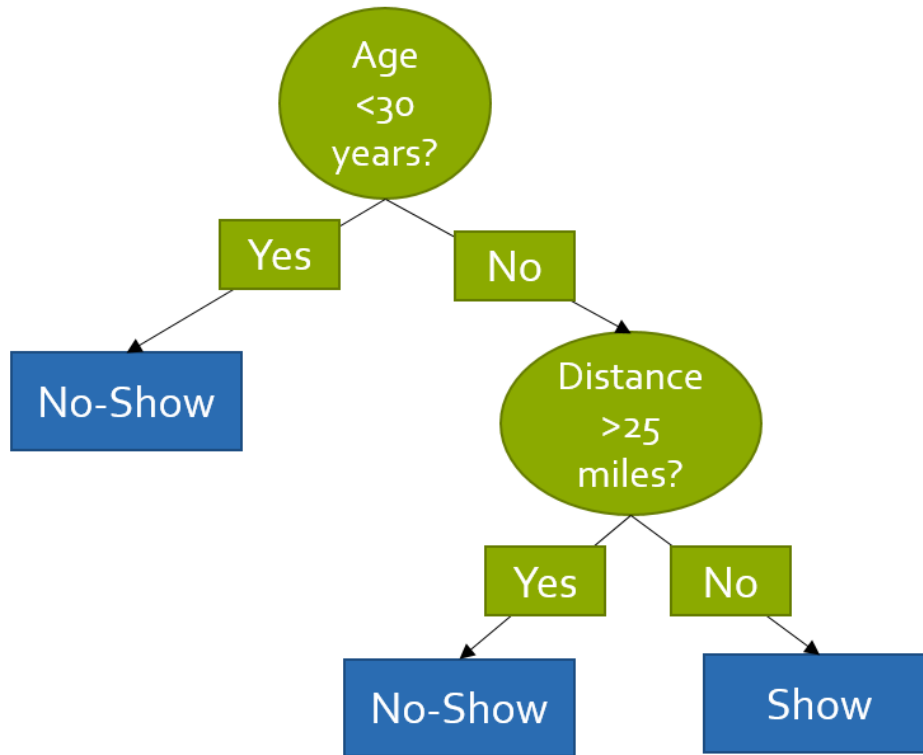


Figure 3. Representation of a Classification Tree

This method has several advantages over other classification techniques. First of all, CART is a non-parametric method, implying that no assumptions regarding to the probability distribution that the data follows has to be known a priori [32]. The second advantage that can be mentioned is the simplicity of results, which facilitate its interpretation and make it easier to present to management or personnel not familiarized with data mining techniques. It is important to mention the ability of the technique to manage missing values [32]. One of the most relevant advantages in relation to this research is the fact that the presence of multicollinearity does not affect the performance of the method, as in other similar techniques such as logistic regression. The majority of the relevant attributes associated with no-shows are demographic factors, which tend to present high correlation among them. Selecting a

prediction technique that can deal with the internal dependencies among the variables is essential, and CART accomplishes that.

Trees are composed of nodes, categorized in three types. The first type of node is known as **root node**, which is the initial node from which the splitting process begins. It has no incoming edges, but can have multiple outgoing edges [34]. **Internal nodes** are the second type of nodes. These are non-terminal nodes that continue the splitting process, since they contains attribute test conditions that allow the segmentation of records that have different characteristics. Each internal node consists of exactly one incoming edge and multiple outgoing edges [34]. When a node cannot be split anymore, because the number of observations in the node does not allows to differentiate the records in order to classify them or because a stopping criteria is reached, it becomes a **leaf node** which is the last type of nodes. Leaf nodes have exactly one incoming edge, no outgoing edges and a class label is assigned to each one [34]. Once the tree is constructed, this method provide a straightforward and simpler method of classifying records. It is a matter of following the node path that matches the records' attributes until a leaf node is reached and the record can be classified with the class label assigned to that node.

A classification tree is structured based on a set of attributes of interest and a response variable that serve as the class label to classify the records. The algorithms that has been developed for the construction of the classification trees provides a suboptimal solution by usually employing a greedy strategy. By this term, it is intended to describe the process of growing the decision tree by making optimum decisions about the attribute that will be used to partition the data in purer subsets [34]. Several algorithms have been developed, the majority of them have been generated using as basis the Hunt's Algorithm. This algorithm

simply considers the training records that belong to more than one class, and use an attribute test condition in order to stratify the records into smaller subsets by creating child nodes. The algorithm is recursively applied to each child node until no more partition of the data is possible.

The process of constructing a classification tree can be summarized in three basic steps: tree growing, tree pruning and tree performance validation. The tree growing step consists in using the recursive splitting of the training data until a stopping criteria is reached. Once the tree is fully grown, the tree pruning step allows to prune it back in order to obtain a smaller tree (subtree) that presents better performance, since larger trees tends to be complex and are likely to over-fit the data. The last step in the process consists in validating the model obtained from the previous steps, by using a new set of records (testing data) and assessing how the model performs predicting the class label of a set of observation never seen before. Each of the steps just mentioned are explained in more detail in Appendix A (see section A.1). Also, Appendix I-section A.2 contains the detailed explanation about the Gradient Boosted Trees technique which is a slow learning approach that improves the prediction power of a resulting decision tree [38]. Finally, a brief explanation of the packages R-part, Caret and gbm, from the computer language R, is presented in the section A.3 of the Appendix A [36]. These packages facilitate the task of applying CART and Gradient Boosted Trees to large sets of data [37, 39].

Two different representative examples were developed to demonstrate the procedure of using classification trees to predict the probability of no-show of a patient and also to test the functionality of the technique. The first example consists of an analysis using generated data with an unknown pattern, with the purpose of assessing how well the technique could adjust

a model in the absence of a known pattern in the data. For the second example, the data was managed creating a pattern of no-show for patients below thirty years old who live further than fourteen miles from the clinic. It has been identified in previous publications that younger patients and patients that live further from the clinic present a higher risk of no-showing to the medical appointments. The purpose of this example is to assess the predictive power of the model constructed, identifying how well the technique capture the pattern in the data, which was introduced on purpose. First of all, both examples were analyzed using the CART methodology. The patterns that the outpatient clinic real data will present is not known, therefore it may be true that applying the CART methodology is sufficient, resulting in lower error rates values. However, the case where applying the CART methodology is not enough can be obtained and it may be necessary to enhance the power of the analysis by implementing a decision tree technique that uses trees as building blocks to construct more powerful prediction models, such as Gradient Boosted Trees. Appendix B contains the results from the application of both techniques in the data of the two representative examples.

3.4 Economic Analysis

The economic analysis consists of determining the different costs and benefits associated with the two categories that are going to be investigated: social and financial. Subsection 1.3 provides a detailed description of the two economic categories. The financial costs considered are strictly related to operational expenditures. Since the research is being held by using generated data and simulating possible scenarios, the emphasis of the economic analysis section is suggesting possible ways that outpatient clinics can use to select cost drivers to assign direct costs and allocation basis to allocate indirect costs. With that information, an economic analysis has to be done in order to determine the cost of a failed

appointment to the clinic. Table 5 enlist the costs, from both economic categories, that will be considered as part of the economic analysis.

Table 5. Cost to be considered in the economic analysis

Financial Costs	Social Costs
Rent	Value lost by non-utilization of resources
Utilities	
Supplies	
Direct Personnel Services	"Loss of Good-will" or Waiting Time Cost
Indirect Personnel Services	
Personnel Overtime	

The economic analysis has been held using the framework presented by Hepard, Hodgkin and Anthony, which is based on the procedures of the UNICEF manual for analysis of district health service costs and financing [33]. This framework consists of seven steps, summarized in six steps to be used in this work. This serves as a guide to identify, define, assign, allocate, and compute unit costs for centers or departments in healthcare systems. The six steps are:

1. Define the final product.
2. Define cost centers.
3. Identify the full cost for each input.
4. Assign inputs to cost centers and allocate all costs to final costs centers.
5. Compute total and unit cost for each final cost center.
6. Report results.

The steps will be described next. Also, how each step will be adapted and aligned to the research methodology of this work will be explained.

1. *Define the final product*

It is essential to define the services or departments of interest. The department of interest can be defined as a cost center, or can be composed of many cost centers. Cost centers are the centers of activity in the healthcare center to which direct and/or indirect costs will be assigned [32]. At the same time, the cost centers are defined by unit of output. The unit of output can be defined as the activity for which the costs will be allocated. Defining the final product help to determine if the unit cost should be calculated for each department separately or if a single unit cost for the entire health center is sufficient. Also, it helps to verify the data availability in each of the areas of interest, since the unit cost calculation is dependent on the aggregation or disaggregation of the required data [32]. Finally, the data availability will influence the selection of the time period for which the data will be collected in order to calculate the unit cost.

The scope of the analysis that is being held in this research work is directed towards analyzing a single area or service. If an outpatient clinic provides more than one service and each one is divided by areas, it is compounded of several cost centers and the analysis can be replicated for each one of them. In the analysis it is expected to estimate the cost of an appointment slot to the outpatient clinic. Each appointment slot is defined by the inputs that are used in order to provide services to a patient in that interval of time. The unit of output will be defined as “time slot”, since the cost of the service provided to a patient will be estimated in that time interval.

2. *Define cost centers*

Identifying cost centers allows to trace the route of the costs through the entire process. A healthcare cost center can be classified in three categories: patient care,

intermediate clinical care, and overhead centers [32]. Patient care cost centers are the areas where the patient receives direct care. Intermediate clinical care cost centers provide support to patient care units, but are organized as separate departments, for example, the radiology department of a hospital. Overhead centers support the previously mentioned centers, for example, the finance department. Patient care centers and intermediate clinical care centers differentiate from the overhead centers in the fact that they generate revenue. Usually the direct and indirect expenses from the overhead centers and other general centers are allocated to those centers that produce revenue.

The methodology established in this research work is focused on patient care cost centers, or better known as **direct patient care departments**, because these are the cost centers that provide direct care to the patients. The links that the direct patient care cost center have with other secondary service departments are not being taken into consideration. The purpose is to identify the immediate costs incurred in providing services to the patient in the space that occupies the cost center. For that reason using a Direct Allocation Method is reasonable since it would not be necessary to consider the costs that arise from other secondary service departments that are indirectly linked to the services provided in the primary direct cost center. The word “Direct” in the term Direct Allocation Method does not refer to the way of allocating the costs (directly or indirectly), instead it makes reference to the fact that the costs allocated are directly related to the service provided in the cost center under analysis, secondary costs from external cost centers are ignored [40].

3. Identify the full cost for each input

The expenditures that will be counted as cost should be determined taking in consideration the resources involved in the service provided in the cost center of interest. In general, the major direct cost categories or inputs are related to salaries and supplies. Indirect cost inputs may include depreciation, utilities, rent and allocated costs from other non-revenue departments.

4. Assign inputs to cost centers and allocate all costs to final cost centers

Each input should be assigned to cost centers. This process can be simple for some inputs, since several of them can be completely assigned to a cost center because the expenditures belongs or comes from activities perform on a particular cost center. However, other inputs assignment is more difficult, since alternative methods have to be applied in order to estimate the cost. One example is the cost of staffing. When the staff is shared by several departments, it is essential to determine what proportion of the cost is assignable to each cost center. Usually, this is assessed using time as the cost driver and determining through administrative data or direct measuring the proportion of the time that the staff work on each center.

Assigning costs consists of linking costs with their respective cost objective, which can be a product or a department [41]. This can be done by using a cost driver as the cost allocating base. The accuracy of the cost measurement depends on the data availability, when the data is incomplete or unreliable, estimates are made. The availability is related to the fact that some resources can be provided internally by the healthcare center, but others can be externally provided by a third-way party.

As mentioned in Section 3.4, two cost categories will be evaluated in this work: financial and social. The financial costs are strictly related to the operational expenditures of the cost center. They are classified according to how accurately they can be attributed to the cost object. If they can be traced with accuracy and little effort, then it can be said that it is a direct cost. In contrast, if the cost is associated with multiple cost objects and cannot be individually traced with accuracy to a cost object, then it is classified as an indirect cost. Table 6 presents the financial costs enlisted and classified as direct or indirect. Also, possible cost drivers that could be used for an effective allocation of the costs have been identified.

Table 6. Financial Costs Possible Cost Drivers

Financial Cost	Direct (D) Indirect (I)	Possible Cost Driver
Rent	I	Per Sq. Feet
Utilities	I	Per Sq. Feet Per Consumption
Supplies	I	No. Utilized Per Patient Per Slot
Direct Personnel Services*	D	Labor Hours
Indirect Personnel Services	I	Labor Hours
Personnel overtime	D or I	Labor Hours

*It may be necessary to estimate the proportion of the total time spent providing services to the patient.

Rent, utilities and supplies are indirect costs because they can be used by more than one cost center of an outpatient clinic, so they have to be indirectly traced to the cost center of interest. Laborers wages or personnel services have been divided in two groups. The first one is Direct Personnel Services including personnel that are directly involved in providing services to the patients, such as the nurses and doctors.

The second group includes the administrative personnel that work in the cost center but is not directly in contact with the patients, such as the receptionists and secretaries. Since direct personnel have several responsibilities and perform different activities, it would be necessary to estimate the proportion of the total labor time the personnel spend providing services to the patient. In terms of cost drivers, rent and utilities can be allocated to the cost objective by using the same cost drivers, per area, as it is suggested in books related to the field of study [41]. When this happens it can be said that they are a cost pool. The same happens with the equipment maintenance and the personnel services. The personnel overtime cost could be direct or indirect because it may cause an effect in both types of personnel.

In the case of the social costs, they need to be allocated indirectly since they are not associated to a tangible aspect of the service provided to the patient. They can be defined as a penalization to the outpatient clinic if there is a total no-show in the slot, causing an immediate loss of capacity, or a penalization due to overbooking the slot and causing higher waiting times and personnel overtime. The first social cost that is being considered in the analysis is the value lost due to the non-utilization of the resources, and it results if there is a total-no show in the slot causing no activity during that interval of time. This cost can be seen and estimated as an opportunity cost. An opportunity cost can be defined as “the potential benefit that is given up as you seek an alternative course of action” [42]. In this case the potential benefit given up is going to be considered in terms of the average profit lost due to not overbooking the slot or due to not assigning another patient to the slot. The value lost due to the non-utilization of the resources can be calculated as the difference between the

average revenue (RV) per patient and the average operational costs (OC) per patient, which is the average profit (P).

$$\textbf{Value lost due to the non-utilization} = P = RV - OC \quad [\text{Equation 1}]$$

The waiting cost is the second social cost being taken in consideration in this work. In this case it is being considered that higher waiting times in an outpatient clinic represent a risk to the system in two aspects: the patient can start to no-show to future appointments because of the unpleasant experience or the patient decides to switch of the healthcare provider. Both scenarios represent a loss of benefits for the system, which leads to visualize the waiting cost as a loss-of-profit if those two scenarios occur as a consequence of overbooking the appointment slots. Since it is a social cost and cannot be directly assigned, it can be allocated in terms of the profit loss multiplied by the average number of times a person may visit a healthcare provider. According to statistics by the CDC National Center for Health Statistics, people in the U.S. visit hospital outpatients and emergency departments at a rate of nearly four visits per person annually [43]. The cost per patient can be estimated in terms of the average profit (P) and the number of times a patient might visit in a year.

$$\textbf{Waiting Cost} = P * \text{Average \# of Times a Patient Visits} \quad [\text{Equation 2}]$$

This estimate of the waiting cost may seem to lack precision because it assumes that after one occurrence of delay in the appointment the patient would consider to not show-up or change of healthcare provider. This may not be true at all, because the patient may be willing to wait any required time due to the healthcare provider good public image about his/her work. For future research, this cost estimate

could be refined by investigating, through polls or patients interviews, how many delays a patient is willing to accept before deciding switching of healthcare provider. The total waiting cost for an appointment slot is subjected to the probabilities of overflow from one slot to another. It will be explained with details in sub-sections 4.2.1 and 4.2.3.

5. *Compute total and unit cost for each final cost center*

In this step, the data about utilization is integrated into the cost analysis, with the purpose of determining the cost per unit. The utilization is defined in terms of the unit of service that is intended to represent. The total costs of each cost center is divided by the unit of service. For example, it may be desired to know the cost per patient serviced on X department.

Since it is desired to estimate the cost of an appointment slot which have a duration of certain time length, as established in Step 1, the unit of output is “time”. According to what has been explained in Step 4, some costs are expressed in terms of area, others in terms of labor hours and others are expressed “per patient”. It is necessary to express all the costs in terms of the unit of output. Next, a detail explanation for each cost is presented.

a. Rent Cost per Time Slot (R)

Usually rent is an expense paid at a monthly rate. The total monthly rate can be divided by the total area that occupies the outpatient clinic in order to obtain the cost by square feet. Since the analysis corresponds to evaluating only one cost center, the area of the cost center is required. The rent cost per square feet is multiplied by the cost center area, and the rent

cost is allocated to the cost center. As established before, it is being assumed that a working day is divided in appointment slots of equal time length. To obtain the cost expressed in terms of time, the cost center rent cost is divided by the working time of the cost center. Then that cost per time unit is multiplied by the time length of the appointment slot, in order to express it in terms of time slot length.

$$R = \frac{\text{Monthly Rent (\$)}}{\text{Total Sq. Feet}} * \left(\frac{\text{Cost Center Sq. Feet}}{\text{Cost Center Working Time}} \right) * \left(\frac{\text{Time Length of the Appointment Slot}}{\text{the Appointment Slot}} \right) \quad [\text{Equation 3}]$$

b. Utilities Cost per Time Slot (U)

If the cost driver being used is square feet, then the utilities cost per time slot is calculated with a procedure similar to the presented for the rent cost (Equation 13). Using “consumption” as the driver the total monthly utilities cost can be divided by the clinic total consumption (kWh) and then multiplied by the cost center consumption (kWh). To obtain the cost expressed in terms of time slot, the cost center utilities cost is divided by the working time of the cost center and then multiplied by the time length of the appointment slot.

$$U = \frac{\text{Monthly Utilities (\$)}}{\text{Total consumption (kwh)}} * \left(\frac{\text{Cost Center Consumption (kwh)}}{\text{Cost Center Working Time}} \right) * \left(\frac{\text{Time Length of the Appointment Slot}}{\text{the Appointment Slot}} \right) \quad [\text{Equation 4}]$$

c. Supplies Cost per Time Slot (S)

The variable x_i represents the number of supplies of the category i used on a regular basis per patient per appointment, where i goes from 1 to n

categories. If the unitary cost c_i is known for each category i , then the supplies cost per patient per appointment slot can be calculated as

$$S = \sum_{i=1}^{i=n} c_i * x_i \quad [\text{Equation 5}]$$

d. Direct Personnel Cost per Time Slot (DP)

Direct Personnel Services includes personnel, such as the nurses and doctors, that are directed involved in providing services to the patients. Commonly, doctors or physicians are paid by annual salaries. In general, it can be estimated that a physician work two thousand hours per year. This is based on a basic workload of fifty weeks per year, forty hours per week. From the total time length of the appointment slot, the physician spends only a fraction of time f with the patient. That fraction of time may vary per patient, but an estimated can be calculated by performing a time study. The physician labor cost (PLC) per patient per time slot can be expressed as:

$$\text{PLC} = \frac{\text{Physician Salary } (\$)}{2,000 \frac{\text{work-hours}}{\text{year}}} * \left(\frac{\text{Time Length of}}{\text{the Appointment Slot}} \right) * f \quad [\text{Equation 6}]$$

The nurses are commonly paid a salary at an hourly rate. The nurses provide services to the patient since the patient check in and also provide support and assistance to the physician during the examination. The nurse labor cost (NLC) can be expressed as:

$$\text{NLC} = \frac{\text{Nurse salary } (\$)}{\text{hour}} * \left(\frac{\text{Time Length of}}{\text{the Appointment Slot}} \right) \quad [\text{Equation 7}]$$

The total Direct Personnel Cost can be estimated by adding the physicians and the nurses labor cost:

$$DP = PLC + NLC \quad \text{[Equation 8]}$$

e. Indirect Personnel Cost per Time Slot (IP)

Clerical personnel do not provide direct healthcare services to the patients but are involved in administrative activities that are indirectly related to the services provided to the patients, for example, managing the appointment schedule, registering the patients and filing the patients' medical records, among others activities. Administrative personnel are paid at an hourly basis. The cost for Indirect Personnel Services can be expressed as:

$$IP = \frac{\text{Administrative Personnel salary (\$)}}{\text{hour}} * \left(\frac{\text{Time Length of the Appointment Slot}}{\text{hour}} \right) \quad \text{[Equation 9]}$$

f. Personnel Overtime Cost per Time Slot (OP)

Overtime work can result as a consequence of overbooking the appointment slot and since the regular work time is not sufficient to provide services to all the patients booked, extra time is required. Compensation for overtime worked hours does not apply to doctors, but it does for nurses and administrative personnel. Paid for overtime hours is regulated by law. For example, in Puerto Rico a laborer is paid regular hourly salary for an eight hours daily working load. The laborer only can

work in excess of that daily limit, if paid at least one and a half times the regular salary. If the overtime hours worked exceeds forty hours a week, the worker should be paid double the regular salary [44]. Taking this into consideration, the overtime personnel cost can be calculated by multiplying the overtime salary by the number of extra hours worked:

$$OP = \frac{\text{Overtime Salary (\$)}}{\text{hours}} * \left(\frac{\text{Number of extra}}{\text{the Appointment Slot}} \right) \quad [\text{Equation 10}]$$

Since the attendance of the patients to the appointments is not known with certainty, the number of possible overtime hours required cannot be calculated. However, it can be estimated by calculating an expected value of the cost by using the probabilities of overflow in the last appointment slot of the day. Details about this procedure will be explained in Sub-section 4.2.4.

g. Value lost due to the non-utilization Cost per Time Slot (V)

If calculated as established in Equation 1, a cost per patient per appointment time slot is obtained. No further computations are required, since it is already expressed in terms of the unit of output. Since the parameters R , U , S , DP and IP (from Equation 3 to 9) represent the operational costs of the cost center, and assuming that the parameter RV is the average revenue the clinic generates from providing services to a patient, Equation 1 can be expressed as:

$$V = RV - (R + U + S + DP + IP) \quad [\text{Equation 11}]$$

This cost occurrence is subjected to the fact that the slot is completely empty because patients no-show to the appointment. An expected value of this cost can be estimated by using the probability of total no-show in a slot, which can be obtained from the overflow calculation analysis that will be presented in Sub-section 4.2.1. For that reason, further explanation of the computation of the expected value of this cost will be explained with details in Sub-section 4.2.2.

h. Waiting Cost (W)

Since the cost of the value lost due to the non-utilization of the resources was expressed in terms of the revenue and the operational expenses (Equation 11). The waiting cost (Equation 2) can also be expressed in the same terms as:

$$W = V * \text{Average \# of Times a Patient Visits} \quad [\text{Equation 12}]$$

The waiting cost depends on the occurrence of overflow from one slot to another. The cost per patient per appointment time slot, as presented in Equation 12, will be elaborated with more details in Sub-section 4.2.3 by adding the overflow probabilities (Sub-section 4.2.1) in order to obtain the expected value of the cost.

Once all the individual costs are expressed in terms of the unit of output, they can be integrated as a cost model. This will be presented in details in Sub-section 4.2.5.

6. Report results

The report should be redacted in a way that it is clear to the reader how the costs were categorized, assigned and allocated. Also, it should present which costs were included in the unit cost calculation, and which were not.

3.5 Cost Model for the Expected Cost of the Appointment Slot

Access to outpatient facilities is controlled through scheduled appointments. Usually, a day is divided in n number of slots. Those slots are intervals of time that may or may not be of equal length. The assignment of patients to slots can go from the simplest form of assigning one patient per slot, to overbooking, where more than one patient is assigned per slot. Figure 4 and Figure 5 present an illustrative representation of both scheduling scenarios. The assumption that we have n number of slots and each slot have a time length s will be used. For the case in which only one patient is assigned to each slot (Figure 4), a patient $Patient(i,l)$ is assigned to each slot. The parameter $Patient(i,l)$ represents the patient booked i ($i=1 \dots m$) assigned to the slot l ($l=1 \dots n$), where m is the total number of patients booked in slot l . In the case of only one patient per slot, the index i is not relevant since only one patient per slot is booked, but it is relevant for the overbooking case. Each patient has been assigned a probability of no-show $P(i,l)$, which is obtained from the classification model constructed using Classification Trees methodology. Physical resources and personnel services related to the service provided in that length of time can be assigned and their cost can be estimated. If the patient shows to the appointment, revenue that overcome those costs is generated. If the patient fails to show (no-show) an immediate loss of capacity occurs, and related financial costs can be assigned and social costs can be allocated.

Complexity is encountered when the overbooking scenario (Figure 5) is analyzed. It is similar to the one person per slot case, with the difference that multiple patients can be assigned to the same slot. Even when it is desired to overbook avoiding the assignment of patients with high probability of show in the same slot, more than one patient can appear at the same time. If more than one patient shows to the same slot, appointments begin to fall behind, a chain effect occurs across all the slots and new direct and indirect costs arise as a consequence. Cost of waiting and cost of overtime should be considered, since once the patients arrive to the appointment, all of them should be serviced. Waiting is considered in this study as a social indirect cost since it cannot be directly estimated, while the overtime cost is considered as a financial indirect cost. The waiting cost can be visualized as a penalty cost incurred by the system due to overbooking that slot. In order to estimate the expected waiting time per slot it would be necessary to calculate the probability of overflow from slot l to slot $l+1$. To calculate the overtime cost, it would be necessary to estimate the additional time that could be required to provide the service to all the patients.

Since information about the patients' probability of no-show and information about the economic impact of failed appointments is generated from this research methodology, for a given schedule configuration an expected cost of an appointment slot can be estimated. For that reason, a cost model considering the financial and social cost categories has been developed. The cost that will be estimated is referred as an expected cost because the model integrates the patients' probabilities of no-show and the possible probability of overflow per slot. This work contributes to the academic literature and the creation of knowledge because it responds to the necessity for further research in this

area. As Lofti and Torres identified [5] as an area for future research, the potential impact of incurring overtime cost and the cost of waiting caused by overbooking can be compared versus the loss of potential income due to no-shows.

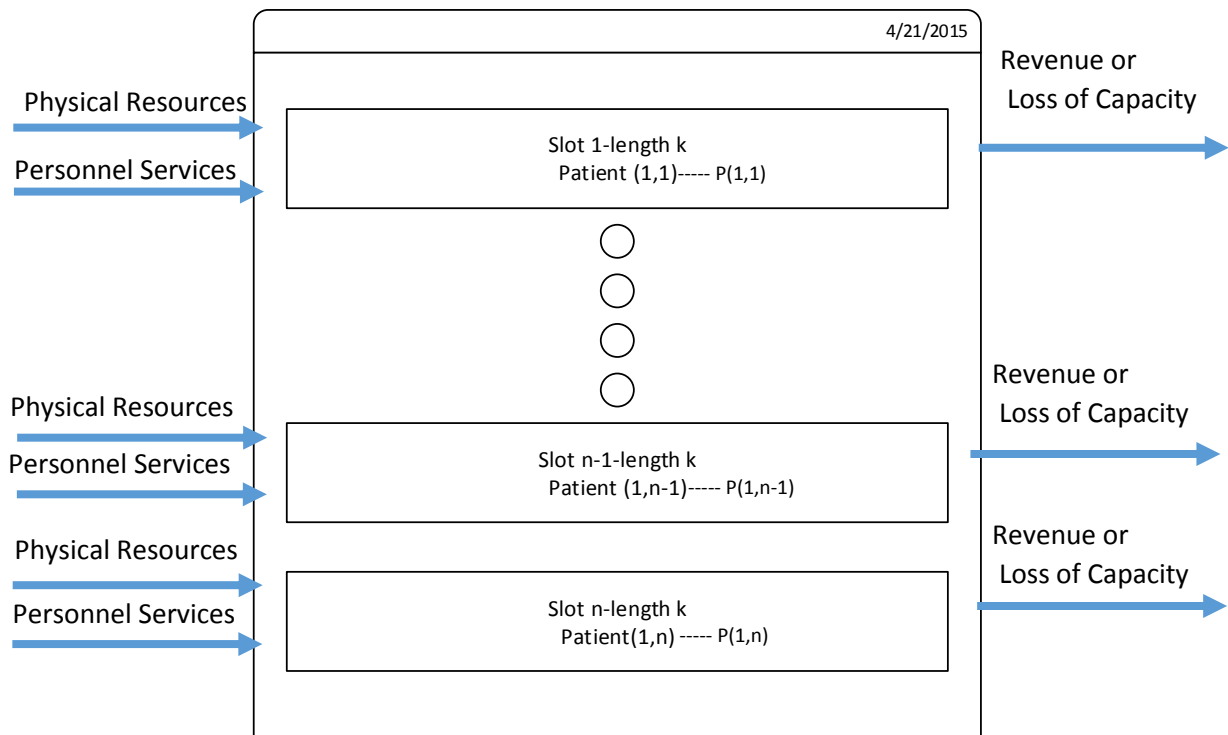


Figure 4. One Patient per Slot Representation

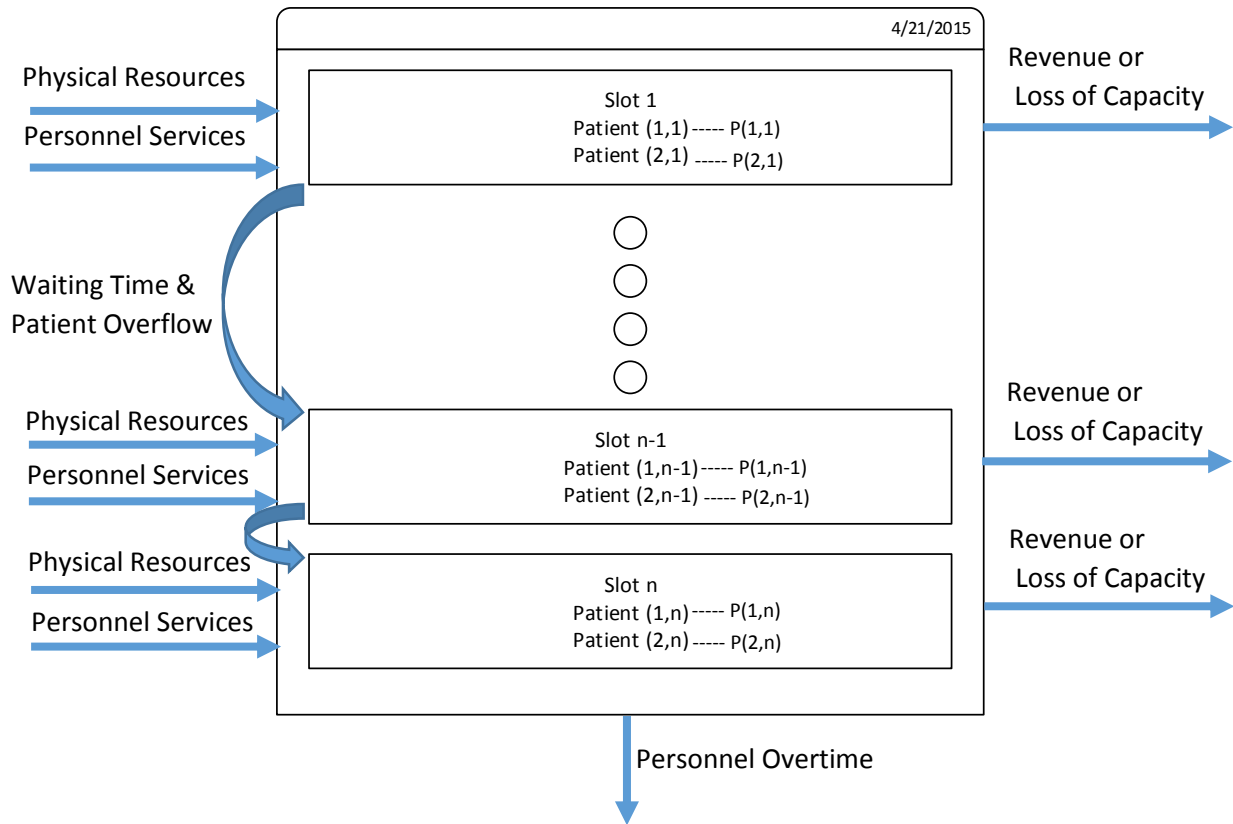


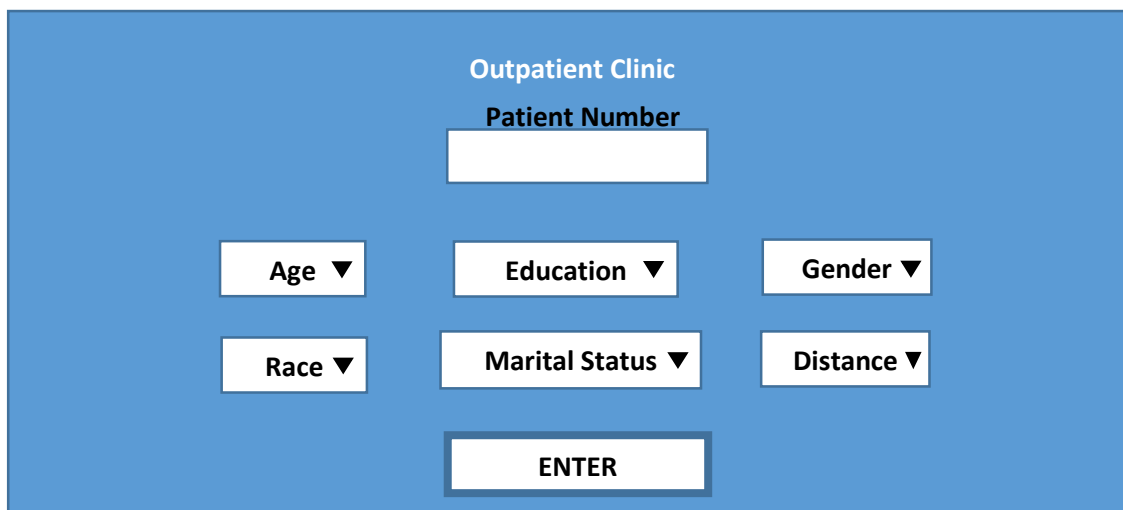
Figure 5. Overbooking Representation

3.6 Test Bed

Information related to the type of data that a clinic have about a patient and information related to the current procedure used for appointment scheduling has been collected through interviews with authorized personnel of an outpatient clinic. Real data about the patients attributes could not be provided for this work due to private policies of the clinic. The methodology has been verified, demonstrated and applied through a test bed. It is denominated a test bed because even though the data being used is generated, the analysis is performed using real parameters from an outpatient clinic. The objective is to simulate different scheduling scenarios with the purpose of assessing the no-show cost for each one of them. This will allow to perform a comparison between scenarios and will contribute to the verification and validation of the methodology developed.

3.7 Interactive Platform in Excel

The applicability of the cost model is presented in the form of an interactive platform in Excel, with the purpose of demonstrating the usefulness of the developed methodology. The objective is to provide a tool that may help the outpatient facility to estimate the cost of no-show and assist during the appointment scheduling process. In the platform, the user can indicate the attributes of the patient (Figure 6) and automatically the associated probability of no-show will be calculated (Figure 7). Once the patients are assigned to the appointment slots, the tool generate the estimated expected value of the cost (Figure 7). Figure 6 and 7 are an illustrative representation of the initial idea developed for the interactive platform.



Outpatient Clinic

Patient Number

Age ▼

Education ▼

Gender ▼

Race ▼

Marital Status ▼

Distance ▼

ENTER

Figure 6. Input Representation-Interactive Platform

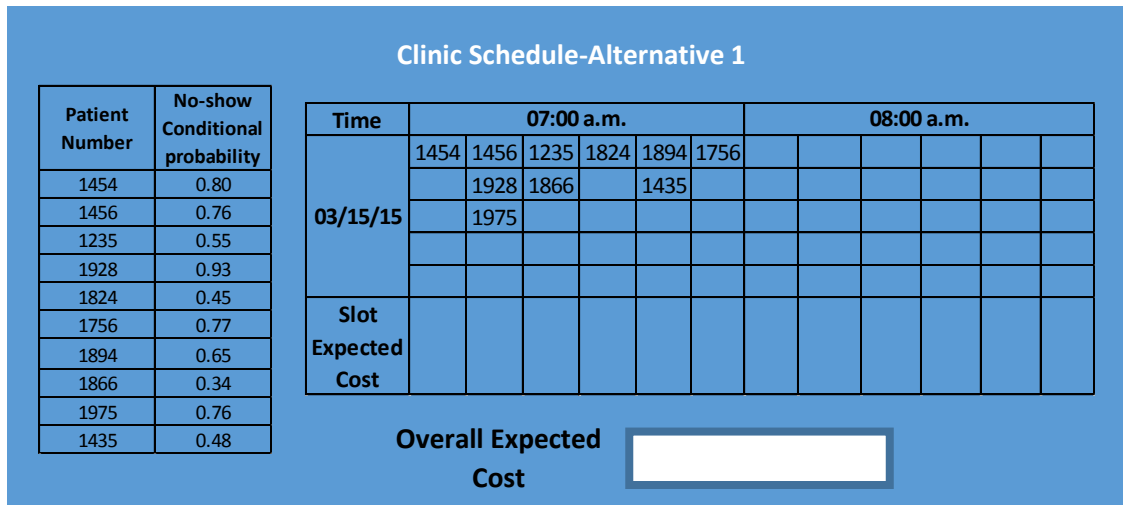


Figure 7. Output Representation-Interactive Platform

4 CHAPTER – RESULTS

4.1 Cost Model Results

As mentioned in Sub-section 3.5, estimating the expected waiting cost per slot for an overbooking scheduling scenario requires the computation of the probability of overflow from one slot to another. This is correct under the assumption that if more than one of the patients booked for a certain slot shows to the appointment, only one of them will be serviced during that period of time and the others have to wait to be serviced in the next available time slot. The problem is that the next available time slot have other patients booked, which can or cannot show to the appointment. A chain effect of patients overflow occurs, causing higher waiting times and personnel overtime. Calculating the overflow probabilities can help in evaluating patients scheduling scenarios, in order to identify those that have a lower probability of overflow from one slot to another. Then those overflow probabilities can be translated in terms of costs, to be used as a performance measure.

4.2.1 Overflow Probabilities

An example evaluating an overbooking of up to two patients per appointment slot will be considered to illustrate the process of calculating the overflow probabilities. Two important assumptions are considered in this analysis. First of all, the appointment slots are of equal time length. Second, exactly m patients are going to be assigned on each time slot; in the case of this illustrative example m is equal to two patients. Table 7 presents a hypothetical schedule to be used for the example. It consists of three time slots with exactly two different patients booked on each one. A total of six different patients are booked in the three slots, two patients on each one. A patient i is booked at a slot n , and at the moment of the appointment the patient can

be in any of two conditions j : Show ($j=1$) or No-Show ($j=2$). Each patient has a probability for each condition j , according to the data mining analysis using the CART methodology (explained in Section 3.3).

Table 7. Appointment Schedule for Illustrative Example-2 Patients Overbooking

Slot 1			Slot 2			Slot 3		
Patient	P(show)	P(no-show)	Patient	P(show)	P(no-show)	Patient	P(show)	P(no-show)
1	0.8	0.2	1	0.6	0.4	1	0.6	0.4
2	0.7	0.3	2	0.9	0.1	2	0.3	0.7

Basically, estimating the overflow probabilities requires to consider all the possible scenarios that can occur. These scenarios are mutually exclusive, since only one of them can actually occur at each time slot. The initial number of scenarios t that can occur depends on the number of patients m assigned to the slot n , and can be calculated according to the following combinations equation [45]:

$$t = \sum_{r=0}^m \frac{m!}{(m-r)!(r!)} \quad [\text{Equation 13}]$$

A total of four initial mutually exclusive scenarios ($t=4$) are possible for the case of two patients overbooking:

Scenario #1: Patient 1 shows to the appointment and Patient 2 does not show.

Scenario #2: Patient 1 does not show to the appointment and Patient 2 shows.

Scenario #3: Both patients show the appointment.

Scenario #4: Both patients do not show to the appointment.

Since each patient can be visualized as an independent event, the probability of occurrence of each scenario can be calculated by multiplying the probabilities of each event [45]. Let's assume that P_{ijl} represents the probability of a patient i being in response j on slot l , where $i=1,\dots,m$, $j=1$ (Show) or 2 (No-Show) and $l=1,\dots,n$. The probability of occurrence of the four scenarios presented before can be represented as:

$$\mathbf{P(Scenario \#1)}_l = P_{11l} * P_{22l}$$

$$\mathbf{P(Scenario \#2)}_l = P_{12l} * P_{21l}$$


$$\mathbf{P(Scenario \#3)}_l = P_{11l} * P_{21l}$$

$$\mathbf{P(Scenario \#4)}_l = P_{12l} * P_{22l}$$

To have a better understanding of the process, from now on, the procedure is going to be explained for each time slot individually.

Slot #1 ($n=1$)

Since this is the first slot, no overflow from previous slots needs to be considered in the analysis. The probabilities of the four scenarios, according to the values in Table 7 are:

$\mathbf{P(Scenario \#1)}_l = P_{111} * P_{221} = 0.80 * 0.30 = 0.24 \rightarrow \text{Overflow of 0}$	 <p>Scenarios when the previous outflow is 0 patients ($k=0$), because there is no previous slot.</p>
$\mathbf{P(Scenario \#2)}_l = P_{121} * P_{211} = 0.20 * 0.70 = 0.14 \rightarrow \text{Overflow of 0}$	
$\mathbf{P(Scenario \#3)}_l = P_{111} * P_{211} = 0.80 * 0.70 = 0.56 \rightarrow \text{Overflow of 1}$	
$\mathbf{P(Scenario \#4)}_l = P_{121} * P_{221} = 0.20 * 0.30 = 0.06 \rightarrow \text{Overflow of 0}$	

The occurrence of each one of these scenarios could have an effect in relation to the overflow from this slot to the next one. For example, in the scenarios #1 and #2 only one of the two patients booked show to the appointment, leading to the fact that if one of these two scenarios occurs, no overflow results. The same happens if Scenario #4 occurs because none of the two booked patients' shows to the appointment, the difference is that in this scenario an immediate loss of capacity results because no patient is serviced. In contrast, Scenario #3 results in an overflow of one patient because both patients show to the appointment and only one can receive services in that interval of time. In summary, from the first slot can result either no overflow or and overflow of 1 patient. Since the four scenarios are mutually exclusive, the probability of no overflow or an overflow equal to 0 is simply the sum of the three scenarios that could result in that output.

Assuming that POF_{kl} represents the probability of overflow of k patients in slot l , where for each slot l the index $k=0,\dots,l$; in this first slot POF_{0I} and POF_{1I} can be calculated.

$$POF_{0I} = P(\text{Scenario \#1})_I + P(\text{Scenario \#2})_I + P(\text{Scenario \#4})_I = 0.44$$

$$POF_{1I} = P(\text{Scenario \#3})_I = 0.56$$

These two probabilities of overflow will be considered when analyzing the next slot, Slot #2.

Slot #2 ($n=2$)

Two different patients are booked in this slot. The same four scenarios discussed above applies to this slot; but with a slightly difference. Now, those four scenarios

have to be analyzed taking in consideration that an overflow from the previous slot can occur. In this case the overflow from Slot #1 to Slot #2 can either be 0 patients or 1 patient. Two times the number of initial scenarios has to be evaluated for this slot, in this case a total of eight scenarios.

The computations of the probabilities of the four initial scenarios for Slot #2, according to the values in Table 7, are:

$$\mathbf{P(Scenario \#1)_2: P_{112} * P_{222} = 0.60 * 0.10 = 0.06}$$

$$\mathbf{P(Scenario \#2)_2: P_{122} * P_{212} = 0.40 * 0.90 = 0.36}$$

$$\mathbf{P(Scenario \#3)_2: P_{112} * P_{212} = 0.60 * 0.90 = 0.54}$$

$$\mathbf{P(Scenario \#4)_2: P_{122} * P_{222} = 0.40 * 0.10 = 0.04}$$

Assuming that S_{plk} represents the probability of the scenario p in the time slot l with a previous overflow of k patients, where $p = 1, \dots, t$, $l = 1, \dots, n$ and $k = 0, \dots, l-1$. Also, assuming that the order of the scenarios will not be altered, then the probability of the scenarios will be $S_{plk} = P(\text{Scenario } p)_l * \text{POF}_{k,l-1} \forall k, p$ where $k=0, \dots, l-1$ and $p=1, \dots, t$.

$$\mathbf{S_{120} = P(\text{Scenario \#1})_2 * \text{POF}_{01} = 0.06 * 0.44 = 0.0264 \rightarrow \text{Overflow of 0}}$$

$$\mathbf{S_{220} = P(\text{Scenario \#2})_2 * \text{POF}_{01} = 0.36 * 0.44 = 0.1584 \rightarrow \text{Overflow of 0}}$$

$$\mathbf{S_{320} = P(\text{Scenario \#3})_2 * \text{POF}_{01} = 0.54 * 0.44 = 0.2376 \rightarrow \text{Overflow of 1}}$$

$$\mathbf{S_{420} = P(\text{Scenario \#4})_2 * \text{POF}_{01} = 0.04 * 0.44 = 0.0176 \rightarrow \text{Overflow of 0}}$$

Scenarios
when the
previous
outflow is
0 patients
($k=0$).

$$\begin{array}{l}
S_{121} = P(\text{Scenario \#1})_2 * POF_{11} = 0.06 * 0.56 = 0.0336 \rightarrow \text{Overflow of 1} \\
S_{221} = P(\text{Scenario \#2})_2 * POF_{11} = 0.36 * 0.56 = 0.2016 \rightarrow \text{Overflow of 1} \\
S_{321} = P(\text{Scenario \#3})_2 * POF_{11} = 0.54 * 0.56 = 0.3024 \rightarrow \text{Overflow of 2} \\
S_{421} = P(\text{Scenario \#4})_2 * POF_{11} = 0.04 * 0.56 = 0.0224 \rightarrow \text{Overflow of 0}
\end{array}
\left. \vphantom{\begin{array}{l} S_{121} \\ S_{221} \\ S_{321} \\ S_{421} \end{array}} \right\} \begin{array}{l} \text{Scenarios} \\ \text{when the} \\ \text{previous} \\ \text{outflow is} \\ \text{1 patient} \\ \text{(k=1).} \end{array}$$

In this slot there are eight scenarios, that depending of which one occur, it can cause an overflow that can range from 0 to 2 patients. For example, in the scenario S_{321} both booked patients show to the appointment, and since an overflow of one patient is received from Slot #1, only one patient can receive service and the other two will overflow the next time slot. The probabilities of overflow POF_{kl} for the second slot are:

$$POF_{02} = S_{120} + S_{220} + S_{420} + S_{421} = 0.2248$$

$$POF_{12} = S_{320} + S_{121} + S_{221} = 0.4728$$

$$POF_{22} = S_{321} = 0.3024$$

These three probabilities of overflow will be considered when analyzing the next slot, Slot #3.

Slot #3 (n=3)

Since three possible overflows can result from the previous slot, the number of scenarios to analyze in this slot is three times the number of the initial scenarios. In this case, twelve scenarios will be evaluated. The computations of the probabilities of the four initial scenarios for Slot #3, according to the values in Table 7, are:

$$\mathbf{P(Scenario \#1)}_3: P_{113} * P_{223} = 0.60 * 0.70 = 0.42$$

$$\mathbf{P(Scenario \#2)}_3: P_{123} * P_{213} = 0.40 * 0.30 = 0.12$$

$$\mathbf{P(Scenario \#3)}_3: P_{113} * P_{213} = 0.60 * 0.30 = 0.18$$

$$\mathbf{P(Scenario \#4)}_3: P_{123} * P_{223} = 0.40 * 0.70 = 0.28$$

The twelve scenarios, taking in consideration the possible overflow values from the previous slot are:

$$\begin{aligned} \mathbf{S_{130}} &= \mathbf{P(Scenario \#1)}_3 * \mathbf{POF_{02}} = 0.42 * 0.2248 = 0.0944 \rightarrow \text{Overflow of 0} \\ \mathbf{S_{230}} &= \mathbf{P(Scenario \#2)}_3 * \mathbf{POF_{02}} = 0.12 * 0.2248 = 0.0269 \rightarrow \text{Overflow of 0} \\ \mathbf{S_{330}} &= \mathbf{P(Scenario \#3)}_3 * \mathbf{POF_{02}} = 0.18 * 0.2248 = 0.0404 \rightarrow \text{Overflow of 1} \\ \mathbf{S_{430}} &= \mathbf{P(Scenario \#4)}_3 * \mathbf{POF_{02}} = 0.28 * 0.2248 = 0.0629 \rightarrow \text{Overflow of 0} \end{aligned}$$

Scenarios when the previous outflow is 0 patients ($k=0$).

$$\begin{aligned} \mathbf{S_{131}} &= \mathbf{P(Scenario \#1)}_3 * \mathbf{POF_{12}} = 0.42 * 0.4728 = 0.1985 \rightarrow \text{Overflow of 1} \\ \mathbf{S_{231}} &= \mathbf{P(Scenario \#2)}_3 * \mathbf{POF_{12}} = 0.12 * 0.4728 = 0.0567 \rightarrow \text{Overflow of 1} \\ \mathbf{S_{331}} &= \mathbf{P(Scenario \#3)}_3 * \mathbf{POF_{12}} = 0.18 * 0.4728 = 0.0851 \rightarrow \text{Overflow of 2} \\ \mathbf{S_{431}} &= \mathbf{P(Scenario \#4)}_3 * \mathbf{POF_{12}} = 0.28 * 0.4728 = 0.1323 \rightarrow \text{Overflow of 0} \end{aligned}$$

Scenarios when the previous outflow is 1 patient ($k=1$).

$S_{132} = P(\text{Scenario \#1})_3 * POF_{22} = 0.42 * 0.3024 = 0.1270 \rightarrow \text{Overflow of 2}$	$\left. \begin{array}{l} \\ \\ \\ \end{array} \right\} \text{Scenarios when the previous outflow is 2 patients (k=2).}$
$S_{232} = P(\text{Scenario \#2})_3 * POF_{22} = 0.12 * 0.3024 = 0.0362 \rightarrow \text{Overflow of 2}$	
$S_{332} = P(\text{Scenario \#3})_3 * POF_{22} = 0.18 * 0.3024 = 0.0544 \rightarrow \text{Overflow of 3}$	
$S_{432} = P(\text{Scenario \#4})_3 * POF_{22} = 0.28 * 0.3024 = 0.0846 \rightarrow \text{Overflow of 1}$	

The probabilities of overflow POF_{kl} for the third slot are:

$$POF_{03} = S_{130} + S_{230} + S_{430} + S_{431} = 0.3165$$

$$POF_{13} = S_{330} + S_{131} + S_{231} + S_{432} = 0.3802$$

$$POF_{23} = S_{331} + S_{132} + S_{232} = 0.2483$$

$$POF_{33} = S_{332} = 0.0544$$

These four probabilities of overflow will be considered when analyzing the next slot, Slot #4.

In general (for two patients overbooking (m=2))

If the order of the initial scenarios is maintained as presented, several observations can be pointed out in relation to patterns in the calculation of the overflow probabilities slot per slot for the case of two patients overbooking schedule.

1. For each slot l calculate the probabilities of the four initial scenarios

$$P(\text{Scenario \#1})_l = P_{11l} * P_{22l}$$

$$P(\text{Scenario \#2})_l = P_{12l} * P_{21l}$$

$$P(\text{Scenario \#3})_l = P_{11l} * P_{21l}$$

$$P(\text{Scenario \#4})_l = P_{12l} * P_{22l}$$

2. For each slot l calculate the probabilities of the scenarios S_{plk} by taking in consideration the values of overflow (k) that can result from the previous time slot:

$$S_{plk} = P(\text{Scenario } p)_l * \text{POF}_{k,l-1} \forall k, p \text{ where } k=0, \dots, l-1 \text{ and } p=1, \dots, t \text{ and } k \geq 0$$

3. The possible resulting overflow of each scenario S_{plk} can be determined in the following way:

For $k=0$:

- $S_{1/0} \rightarrow$ Overflow of k
- $S_{2/0} \rightarrow$ Overflow of k
- $S_{3/0} \rightarrow$ Overflow of $k+1$
- $S_{4/0} \rightarrow$ Overflow of k

For $k>0$:

- $S_{1/k} \rightarrow$ Overflow of k
- $S_{2/k} \rightarrow$ Overflow of k
- $S_{3/k} \rightarrow$ Overflow of $k+1$
- $S_{4/k} \rightarrow$ Overflow of $k-1$

4. For each slot, the possible overflow that can result and affect to the next time slot goes from 0 to l .

$$\text{POF}_{k,l} \text{ where } k=0, \dots, l$$

5. The probability of overflow POF_{kl} is the sum of the probabilities of the scenarios S_{plk} which if occurring could lead to an overflow of k patients from the slot l to the slot $l+1$.

In general

The same procedure explained before for the case of overbooking two patients per slot, can be reproduced for any number of patients m . The following basic steps can

be followed. The procedure will be exemplified by using the case of overbooking three patients ($m=3$).

1. Calculate the number of initial scenarios t by using Equation 13.

$$\text{Example: } t = \sum_{r=0}^m \frac{m!}{(m-r)!(r!)} = \sum_{r=0}^3 \frac{3!}{(3-r)!(r!)} = 8 \text{ initial scenarios}$$

2. Establish the t initial scenarios in the order that will be analyzed across the slots evaluation.

Example:

Scenario #1: Patient #1 shows, Patient #2 and Patient #3 do not show.

Scenario #2: Patient #2 shows, Patient #1 and Patient #3 do not show.

Scenario #3: Patient #3 shows, Patient #1 and Patient #2 do not show.

Scenario #4: Patient #1 does not show, Patient #2 and Patient #3 shows.

Scenario #5: Patient #2 does not show, Patient #1 and Patient #2 shows.

Scenario #6: Patient #3 does not show, Patient #1 and Patient #2 shows.

Scenario #7: All patients show to the appointment.

Scenario #8: All patients do not show to the appointment.

3. Beginning with the first slot, calculate the probabilities ($\mathbf{P}(\text{Scenario } p)_n$) of each of the t initial scenarios by multiplying the probabilities of the independent events, defined by the probabilities of show and no-show of the m patients booked in the slot l , P_{ijl} where $i=1, \dots, m$, $j=1$ (Show) or 2 (No-Show) and $l=1, \dots, n$.

Example:

$$\mathbf{P(Scenario \#1)}_n = P_{11n} * P_{22n} * P_{32n}$$

$$\mathbf{P(Scenario \#2)}_n = P_{12n} * P_{21n} * P_{32n}$$

$$\mathbf{P(Scenario \#3)}_n = P_{12n} * P_{22n} * P_{31n}$$

$$\mathbf{P(Scenario \#4)}_n = P_{12n} * P_{21n} * P_{31n}$$

$$\mathbf{P(Scenario \#5)}_n = P_{11n} * P_{22n} * P_{31n}$$

$$\mathbf{P(Scenario \#6)}_n = P_{11n} * P_{21n} * P_{32n}$$

$$\mathbf{P(Scenario \#7)}_n = P_{11n} * P_{21n} * P_{31n}$$

$$\mathbf{P(Scenario \#8)}_n = P_{12n} * P_{22n} * P_{32n}$$

4. Compute the scenarios probabilities for the first slot and determine the possible overflow that can result if each one occurs.

Example: First Slot

$$\mathbf{P(Scenario \#1)}_I = P_{111} * P_{221} * P_{321} \rightarrow \text{Overflow of 0}$$

$$\mathbf{P(Scenario \#2)}_I = P_{121} * P_{211} * P_{321} \rightarrow \text{Overflow of 0}$$

$$\mathbf{P(Scenario \#3)}_I = P_{121} * P_{221} * P_{311} \rightarrow \text{Overflow of 0}$$

$$\mathbf{P(Scenario \#4)}_I = P_{121} * P_{211} * P_{311} \rightarrow \text{Overflow of 1}$$

$$\mathbf{P(Scenario \#5)}_I = P_{111} * P_{221} * P_{311} \rightarrow \text{Overflow of 1}$$

$$\mathbf{P(Scenario \#6)}_I = P_{111} * P_{211} * P_{321} \rightarrow \text{Overflow of 1}$$

$$\mathbf{P(Scenario \#7)}_I = P_{111} * P_{211} * P_{311} \rightarrow \text{Overflow of 2}$$

$$\mathbf{P(Scenario \#8)}_I = P_{121} * P_{221} * P_{321} \rightarrow \text{Overflow of 0}$$

5. For the first time slot, compute POF_{kl} , the probabilities of overflow of k patients in slot l , by adding the probabilities of those scenarios that could result in the same output of an overflow of k patients. Those probabilities will be used in the evaluation of the next time slot.

Example: First Slot

$$POF_{01} = \mathbf{P(Scenario \#1)}_1 + \mathbf{P(Scenario \#2)}_1 + \mathbf{P(Scenario \#3)}_1 + \mathbf{P(Scenario \#8)}_1$$

$$POF_{11} = \mathbf{P(Scenario \#4)}_1 + \mathbf{P(Scenario \#5)}_1 + \mathbf{P(Scenario \#6)}_1$$

$$POF_{21} = \mathbf{P(Scenario \#7)}_1$$

6. Since there is a probability of receiving overflow from the previous slot, from the second slot onwards, (t *number of possible POF 's from the previous slot) number of scenarios S_{plk} are going to be evaluated. The initial scenarios are evaluated in the slot when the previous overflow is $k=0$ and so on until assessing all the possible k 's that can be evaluated from the previous slot.

Example: Second Slot

Since the overflow probabilities from the first slot could be either zero, one or two patients, a total of $8*3=24$ scenarios are evaluated.

$$\begin{aligned} S_{12k} &= P(\text{Scenario \#1})_2 * POF_{k,1-1} \\ S_{22k} &= P(\text{Scenario \#2})_2 * POF_{k,1-1} \\ S_{32k} &= P(\text{Scenario \#3})_2 * POF_{k,1-1} \\ S_{42k} &= P(\text{Scenario \#4})_2 * POF_{k,1-1} \\ S_{52k} &= P(\text{Scenario \#5})_2 * POF_{k,1-1} \\ S_{62k} &= P(\text{Scenario \#6})_2 * POF_{k,1-1} \\ S_{72k} &= P(\text{Scenario \#7})_2 * POF_{k,1-1} \\ S_{82k} &= P(\text{Scenario \#8})_2 * POF_{k,1-1} \end{aligned} \quad \left. \vphantom{\begin{aligned} S_{12k} \\ S_{22k} \\ S_{32k} \\ S_{42k} \\ S_{52k} \\ S_{62k} \\ S_{72k} \\ S_{82k} \end{aligned}} \right\} \begin{array}{l} \text{Scenarios should} \\ \text{be evaluated for} \\ k=0, k=1 \text{ and } k=2. \end{array}$$

7. Compute POF_{kl} , the probabilities of overflow of k patients in slot l , by adding the probabilities of those scenarios that could result in the same output of an overflow of k patients. Those probabilities will be used in the evaluation of the next time slot. Maintaining the order of the scenarios as established in Step 2, discover the pattern in the POF's in order to facilitate the analysis.

Example:

For any value of k

- $S_{1/0} \rightarrow \text{Overflow of } k$
- $S_{2/0} \rightarrow \text{Overflow of } k$

- $S_{3/0} \rightarrow$ Overflow of k
- $S_{4/0} \rightarrow$ Overflow of $k+1$
- $S_{5/0} \rightarrow$ Overflow of $k+1$
- $S_{6/0} \rightarrow$ Overflow of $k+1$
- $S_{7/0} \rightarrow$ Overflow of $k+2$
- $S_{8/0} \rightarrow$ Overflow of k

8. Use those probabilities POF_{kl} to evaluate the next slot.

4.2.2 Value Lost Due to the Non-Utilization Cost Expected Value (E(V))

For this research work, value is lost due to the non-utilization of the resources when there is a complete no-show in the appointment slot, in other words, no patient received service during that time interval. As established in Section 3.4, the cost related is a social cost that can be defined as an opportunity cost defined by the average profit lost due to not overbooking the slot or due to not assigning another patient to the time slot (Equation 11). Equation 11 presents a way to estimate the cost per patient per slot. However, this cost is dependent of the occurrence of a total no-show in the slot. This cannot be assessed with certainty, since it is unknown if a patient will show to an appointment, but probabilities of show and no-show can be calculated as explained in sections 3.3 and 4.1. In order to estimate the expected value of the cost of the value lost due to the non-utilization of the resources, it is necessary to consider the probability of total no-show in each slot. That probability is estimated for the case of two patients overbooking, as explained in Sub-section 4.2.1, when two patients are booked in an appointment slot, four initial mutually exclusive scenarios are possible:

Scenario #1: Patient 1 shows to the appointment and Patient 2 does not show.

Scenario #2: Patient 1 does not show to the appointment and Patient 2 shows.

Scenario #3: Both patients show the appointment.

Scenario #4: Both patients do not show to the appointment.

In the first appointment slot (Slot #1) there is no overflow from previous slots, then a complete no-show happens if Scenario #4 occurs, assuming that the order of the scenarios is not altered. The probability of a total no-show in Slot #1 is equal to the probability of occurrence of the Scenario #4 in that slot, which is $P(\text{Scenario \#4})_1 = P_{121} * P_{221}$. Refer to Sub-section 4.2.1 for more information. For each following appointment slot, the probability of total no-show is influenced by the probability of overflow from previous slots. A total no-show will occur if Scenario #4 happens and the overflow from the previous slot is zero patients. Using the notation presented in Sub-section 4.2.1, in general for each slot, the probability of total no-show for slots other than Slot#1 can be defined as $S_{4|0} = P(\text{Scenario\#4})_l * \text{POF}_{0,l-1}$. Refer to Sub-section 4.2.1 for more details about the parameters and the notation. Now, the expected value of the cost of value lost due to the non-utilization of the resources, for each slot, can be estimated as:

$$\mathbf{E}(V)_l = V * S_{4|0} \quad \text{[Equation 14]}$$

Where V is the cost per patient per time slot of the value lost due to the non-utilization of resources (Equation 11) and S_{4l0} is the probability of the scenario # 4 in the slot l when the previous overflow is zero patients.

4.2.3 Waiting Cost Expected Value ($E(W)$)

In Section 3.4 a way to allocate the waiting cost per patient was presented. Since the objective is to estimate the cost per appointment slot, it is necessary to express the cost in terms of that unit of output. It is not known with certainty the number of patients that will overflow from one slot to another, however, in Subsection 4.2.1 a procedure to estimate the overflow probabilities from one slot to a consecutive slot has been developed by using the probabilities of show and no-show of the patients booked. The procedure has been established for the case of overbooking two patients per slot, and it has been identified that on each slot l we could have a probability of overflow POF_{kl} where k has values that goes from 0 to l . Taking this into consideration, for the two patients overbooking scenario, the expected value of the waiting cost ($E(W)_l$) can be estimated as:

$$E(W)_l = \sum_{k=0}^l W * k * POF_{kl} \quad [\text{Equation 15}]$$

where W is the waiting cost per patient (Equation 12), k is the number of patients that could overflow from that slot and POF_{kl} is the probability of overflow of the k patients from slot l to slot $l+1$.

4.2.4 Personnel Overtime Cost Expected Value ($E(OP)$)

The overtime can be defined as the additional time required to provide services to patients that could not be attended during the regular work time. Since a day is

divided in n number of time slots, overtime will be required if more than one patient in the last appointment slot (n) are left unattended yet. That value cannot not be estimated with certainty in order to calculate the cost, but the probabilities of overflow in the slot n (Refer to sub-section 4.2.1) can be used to estimate the expected value of the personnel overtime cost. Taking this into consideration, the personnel overtime cost expected value in the last appointment slot $E(OP)_n$ can be expressed as:

$$E(OP)_n = \sum_{k=0}^n OP * k * POF_{kn} \quad [\text{Equation 16}]$$

Where OP is the personnel overtime cost per time slot (Equation 10) and POF_{kn} is the probability of having k patients left unattended at the end of the last time slot n .

4.2.5 Stochastic Cost Model

All the costs has been expressed in terms of the unit output and the cost expected value of each social cost has been calculated. When evaluating an existing appointment schedule, the total cost of an appointment slot C_l for all the slots in the schedule can be expressed as:

$$C_l = R + U + S + DP + IP + V * r + W * k \quad [\text{Equation 17}]$$

Where r is a binary variable that takes a value of 1 if the appointment slot was a complete no-show, and a value of 0 otherwise. The parameter k is the number of patients that will overflow from slot l to slot $l+1$. In the case of estimating the no-show cost when generating an appointment schedule a priori, which add an stochastic element to the cost model, the total expected value of the cost of an appointment slot ($E(C_l)$), for slot 1 to slot $n-1$, can be expressed as:

$$E(C_l) = R + U + S + DP + IP + E(V)_l + E(W)_l \quad [\text{Equation 18}]$$

For the last time slot (n) of the appointment schedule, the personnel overtime cost is added to the equation.

$$E(C_n) = R+U+S+DP+IP+ E(V)_n + E(W)_n + E(OP)_n \quad [\text{Equation 19}]$$

The details about each parameter in the stochastic cost model can be found in Section 3.4 and in Section 4.2, Sub-sections 4.2.1 through 4.2.4. In the next section, this stochastic cost model will be used to evaluate several appointment scheduling scenarios in terms of the total no-show cost to the system.

4.3 Test Bed-Scenarios Simulation

As mentioned before, outpatient clinics construct their appointment schedules by using a single patient per slot approach or by overbooking the appointment slots. When overbooking, the vast majority of the clinics do not have established a specific procedure and instead they assign the patients randomly. The purpose of developing a stochastic cost model in this research work is to create a methodology to assess the economic impact that patients' absences have in healthcare systems, but also to use it as a tool for the evaluation of different scheduling policies, in terms of the total cost to the system.

The work of Lofti and Torres [5] has been used as reference for the policies simulations. They performed a simulation of five scheduling policies. In the policies that included overbooking, their methodology assigned patients in a slot until the expected number of patients reached a value of one. In other words, each slot was overbooked until the sum of the probabilities of show of the patients assigned to the slot reached a value of one. That approach works by overbooking patients with a low probability of showing (high probability of no-show) in the same slot. The authors recognize that this approach could be

harmful to the system, in terms of the waiting time of the patients and the personnel overtime, because patients with a low probability of showing can actually show up to the appointments. However, in contrast to what is being done in this research work, Lofti and Torres are not considering these two aspects in their evaluation; instead they are only comparing the policies in terms of the clinic utilization. In this research work, the policies are being compared in terms of the clinic utilization but also in terms of the total cost to the system, which includes the waiting cost, the personnel overtime cost, and the cost of the resources non-utilization, among other costs. Instead of overbooking until the expected number of patients in the slot reach one, a threshold value of 2 patients overbook per appointment slot will be used.

The following four scheduling policies have been simulated in Visual Basic:

- Policy 1: Assign one patient per appointment slot.
- Policy 2: Avoid overbooking consecutive appointment slots. Overbook one appointment slot and not the next one.
- Policy 3: Overbook all the appointment slots without taking in consideration the patients probabilities of attendance.
- Policy 4: Overbook all the appointment slots by assigning a patient with a high probability of showing ($P(\text{show}) \geq 60$) and a patient with a low probability of showing ($P(\text{show}) \leq 40$) in the same appointment slot.

The four policies has been simulated using data from a data set from the illustrative example that provided the lowest generalized cross-validated error when the model was constructed using the Generalized Boosted Trees technique. The model provides a seventy percent of accuracy in the prediction. The probabilities of show and no-show were obtained for each sample in the data set based on the prediction provided by the classification model.

It is known if the patient actually arrived to the appointment, for each sample. Basically, five hundred medical appointments (samples) have been simulated. The simulation algorithm selects patients from the full data set at a random order and assign them according to the specifications of the scheduling policy being evaluated. It is being considered that one clinic day consist of ten appointment slots. Once all the time slots of a day are full with assigned patients, the algorithm begins a new day; so on until all the patients are assigned to an appointment slot. For the first policy (Policy 1) a total of fifty clinic days has been simulated. Thirty-four days has been simulated for the second policy (Policy 2). Finally, twenty-five days has been simulated for the third (Policy 3) and fourth (Policy 4) policies.

Six performance measures has been calculated based on the output results collected for each policy simulated. The first performance measure is the *clinic utilization*. Lofti and Torres [5] define the clinic utilization based on the proportion of time slots that had at least one patient assigned that showed to the appointment (Equation 20). This measure do not consider patients overflow from previous slots, is only based on patients assignments. Since an empty slot can be filled out due to an overflow of a patient from a previous slot, it has been decided to calculate the *overall clinic utilization* (Equation 21), which is the second performance measure. The third performance measure to consider is the *fraction of slots with overflow* from one slot to a consecutive slot (Equation 22). The fourth performance measure is the *fraction of show overbooks*, which collect the proportion of slots that were overbooked with assigned patients that both showed to the appointment (Equation 23).

$$\text{clinic utilization (\%)} = \left(1 - \frac{\text{Assigned Empty Slots per day}}{\text{Total number of slots per day}}\right) * 100 \quad [\text{Equation 20}]$$

$$\text{overall clinic utilization} = \left(1 - \frac{\text{Real Empty Slots per day}}{\text{Total number of slots per day}}\right) * 100 \quad [\text{Equation 21}]$$

$$\text{fraction of slots with overflow} = \frac{\text{Overflow slots per day}}{\text{Total number of slots per day}} \quad [\text{Equation 22}]$$

$$\text{fraction of show overbooks} = \frac{\text{Assigned show overbooked slots per day}}{\text{Total number of slots per day}} \quad [\text{Equation 23}]$$

The fifth performance measure is the **expected cost to the system**, calculated by evaluating each schedule with the stochastic cost model (Equation 18 and Equation 19), which are based on the probabilities of attendance of the patients. By using the developed stochastic cost model, each policy is also being evaluated in terms of the expected waiting of the patients, the expected non-utilization of the resources and the expected personnel overtime. Finally, the sixth performance measure is the **actual cost to the system**, which is calculated based on the real output of the schedules. The objective is to compare the expected total cost, which is based on probabilities, with the actual total cost, which is based on what would have really happened if it was the real schedule. This will allow to assess how well the stochastic cost model estimated the cost of the schedule under each policy. A set of hypothetical values were used to assess the different cost values considered in the cost model.

A stochastic element is present in the simulation due to the randomization of the patients at the beginning, which cause variability in the results of the different performance measures of interest on each simulation run. It is incorrect to make inferences about the values of the performance measures with only one simulation run, this is the reason why several replication runs are used instead. The required number of replication runs (n) depends on the statistical relative error (e_r) willing to be accepted for the confidence level desired. The replication runs allows to estimate the average values of the performance measures (\bar{x}) by calculating the average value among the replications. The average value do not provide information about the precision of the estimate. To address this issue, each

average value is accompanied with a half-width (h) that provides information about the error in the sampling. In this analysis, an initial set of fifty replication runs has been generated for each policy. This initial sample of replications has been used to estimate if more replication runs (n) are required for a confidence level of 95%, with a relative error (e_r) of 0.05, according to Equation 24 where s is the standard deviation.

$$n = \left(\frac{t_{1-\alpha/2, n-1} * s}{e_r \bar{x}} \right)^2 \quad [\text{Equation 24}]$$

The parameter n has been estimated for all the performance measures, selecting the higher value as the final number of replications to be run. When no further replications are needed, the average value of each performance measure has been estimated and presented accompanied with their respective half-width (h), calculated based on Equation 25.

$$h = \frac{t_{1-\alpha/2, n-1} * s}{\sqrt{n}} \quad [\text{Equation 25}]$$

Individual details about the simulations results for each policy are presented in Appendix A.

Table 8 presents a summary of the results for the scheduling simulations.

Table 8. Results of the Scheduling Policies Simulation

Policy	Clinic Utilization (%)		Overall Clinic Utilization (%)		Fraction of slots with overflow		Fraction of Show Overbooks		Expected Cost to the System (\$)		Actual Cost to the System (\$)	
	Average	h	Average	h	Average	h	Average	h	Average	h	Average	h
1	51	0	51	0	0	0	0	0	26297.5	0	26054.02	0
2	64.51	0.35	74.13	0.24	0.25	0.01	0.13	0.004	23342.09	1076.83	20813.41	177.10
3	75.91	0.39	87.46	0.38	0.55	0.01	0.26	0.004	21517.85	24.31	22741.54	340.38
4	76.82	0.51	87.53	0.47	0.53	0.01	0.25	0.004	21441.85	16.49	22318.44	243.70

Policy 1 requires the higher number of days and slots, fifty days and five hundred slots, to assign all the patients. Two hundred and forty-five (245) empty slot results from patients that did not show to the appointment. The number of real empty slots is equal to the number of assigned empty slots because one patient is assigned per slot and there is no

overflow from one appointment slot to a consecutive appointment slot. As a consequence the clinic utilization is equal to the overall clinic utilization, with a value of fifty-one percent (51%), which implies a non-utilization of almost half of the clinic capacity. No overbooked and overflow slots results from this policy. In terms of the cost to the system, it presents the lower average cost per day in comparison with the other policies, but since it requires a higher number of days and slots to assign all the patients, it results as the policy with the higher total cost to the system. There is a 0.92% of difference between the estimated total expected cost and the total real cost of the schedule. The stochastic cost model provides a suitable estimate of the actual cost of the scheduling policy.

Policy 2 is a scheduling policy with a conservative overbooking scheme, trying to counteract the possible effect of overbooking one appointment slot by not overbooking the consecutive appointment slot. A total of thirty-four days and three hundred forty slots were required to assign all the patients. It resulted in an approximate of one hundred and twenty (120.68 ± 1.20) assigned empty slots, on average. The clinic utilization is approximately sixty-four percent ($64.51\% \pm 0.35\%$) of the clinic capacity. The overbooking of some slots caused overflow on a total of eighty-six (86.04 ± 2.21) appointment slots on average, twenty-five percent ($25\% \pm 0.01\%$) of the total number of slots. This leads to the result that on average only eighty-seven (87.96 ± 0.83) of the one hundred twenty (120.68 ± 1.20) are real empty slots, which lead to a higher overall clinic utilization of approximate seventy-four percent ($74.13\% \pm 0.24\%$). Two patients that showed to the appointment were assigned in the same appointment slot in thirteen percent ($13\% \pm 0.004\%$) of the assigned overbooked appointment slots, on average. The total cost to the system of this policy is lower than the cost of the Policy 1 due to the fact that fewer empty slots results due to the overbooking of

certain slots, causing a higher utilization of the clinic capacity. The reduction in the cost of the value lost due to the non-utilization of the resources counteracted the effect of the cost of waiting and overtime due to the overflow caused by overbooking, also it is due to the balance caused by the pattern created by overbooking one slot and not the consecutive one. A 10.83 percent of difference lies between the estimated total expected cost and the actual cost to the system. This policy resulted in the higher variability in estimating the expected cost among the replication runs. Not overbooking certain slots can results in a buffer if two show patients were assigned in a previous slot. However, it can result in a higher number of empty slots, than when overbooking all slots, even more when patients are assigned randomly. This cause higher variability in the estimate of the cost. Also, there is higher probability of empty slots in this scenario, than in the policies 3 and 4, reason why the expected cost estimated by the stochastic cost model is higher for this policy.

The third policy, Policy 3, consists of overbooking all the appointment slots by assigning the patients randomly. The probabilities of attendance of each patient are not taken under consideration at the moment of the assignment. Twenty five days and, two hundred and fifty slots have been simulated for this policy. Approximately sixty (60.22 ± 0.97) appointment slots resulted as assigned empty slots, however, the real number of empty slots is thirty-one (31.36 ± 0.96), due to one hundred and thirty-seven (137.24 ± 2.50) slots that had overflow caused by overbooking. This leads to an overall clinic utilization of approximately eighty-seven percent ($87.46\% \pm 0.38\%$). In comparison with the previous policies, Policy 3 provides a higher clinic utilization. However, in terms of the fraction of slots with overflow ($55\% \pm 0.01\%$) and the fraction of assigned show overbooked slots ($26\% \pm 0.004\%$), it resulted in a poorer performance. It was expected due to the fact that it is a

general consequence of overbooking all the appointment slots. This is reflected in the total cost to the system. The percentage of difference between the expected total cost and the actual total cost is 5.69 percent, which implies that the stochastic cost model is capturing well the cost behavior of the policy. The total actual cost for this policy is higher than the cost for the second policy (Policy 2) because the increase in the number of overflow slots cause an increase in the waiting cost and the overtime cost. However, the percentage of difference of the total cost between both policies range between a seven percent (7%) and a nine percent (9%), which is not alarming taking in consideration the crucial improvement in the clinic utilization. This result leads to realize that the stochastic cost model is designed to penalize more for the patients waiting time and personnel overtime, than penalizing for the loss of capacity due to a completely empty slot.

The last policy simulated, Policy 4, overbook the appointment slots assigning patients by taking in consideration their probabilities of attendance. The objective is to assign a patient with a high probability of showing with a patient with a low probability of showing. By the results, a slightly improvement in the clinic utilization ($76.82\% \pm 0.51\%$) and the overall clinic utilization ($87.53\% \pm 0.47\%$) is obtained in comparison with Policy 3. Also, on average, fewer overflowed slots (133.02 ± 2.53) and assigned show overbook slots (62.96 ± 0.99) results from this policy. In terms of the total cost to the system, Policy 4 is the policy that resulted with the lower average total expected cost because it balance the patients assignment better, reducing the overflows which cause an increase in the waiting cost and the overtime cost. The actual cost to the system is lower than the resulting cost of the Policy 3 because it balance the patients assignment better, reducing the overflows which cause a decrease in the waiting cost and the overtime cost. The percentage of difference between the

expected total cost and the actual total cost is 4.08 percent, which is lower is comparison with Policy 2 and Policy 3. Better results would have been obtained if the data fitted a classification model with a higher prediction accuracy. The model used have a seventy percent accuracy, which is the best that could be obtained, but a model with a higher accuracy would have allowed better assignment and therefore better results.

In overall it is important to highlight that in terms of the performance measure of relevance in this work, which is the total cost to the system, the results show that Policy #4 is the less costly when estimated with anticipation (Expected Cost to the System). In terms of the Actual Cost to the System, Policy #2 presents the lowest cost. However, even when Policy #4 is slightly more costly, it allows to book the same amount of patients in less days (25 days) providing a higher overall utilization. When comparing the expected cost values with the actual cost values, the higher percent of difference is 10.83% for the Policy #2. Taking this in consideration it can be established that the stochastic cost model developed in this work provides a suitable estimate of the actual cost of the scheduling policies.

4.4 Interactive Platform-Appointment Scheduler

A prototype of an appointment scheduler has been developed in Excel using VBA and applying the methodology developed in this work. Figure 8 present the initial view. It contains the data base with the information of the patients, necessary to predict the probability of attendance of the patients based on the prediction model that should have been constructed with anticipation. If the user press the “Start” button, then it can be started the process of adding a new patient to the data base or the process of creating an appointment for an existent patient.

ID	Social Secun	age	gender	race	m_stat	p_langua	distanc	t_patient	attendance	t_tim	County	Distance (mile)
1001	2322	22	Male	White	Single	Spanish	Mayaguez	Dependent	0	6	Quebradillas	19
1002	1006	64	Male	Other	Married	Spanish	San Germa	Veteran	1	13	San German	8
1003	1023	42	Female	Other	Married	Spanish	Hormiguer	Veteran	1	18	Mayaguez	0
1004	1024	44	Female	White	Married	Other	Aguada	Veteran	1	8	Anasco	6
1005	1026	66	Female	African Ar	Married	Other	San Germa	Veteran	1	21	Isabela	17
1006	1039	33	Male	White	Single	English	Quebradill	Dependent	0	8	Rincon	12
1007	1058	19	Male	Other	Married	Spanish	San Germa	Veteran	1	25	Sabana Grande	14
1008	1063	25	Male	African Ar	Married	Spanish	Mayaguez	Veteran	1	24	Hormigueros	6
1009	1069	38	Male	White	Married	Spanish	Isabela	Veteran	0	21	Aguada	14
1010	1073	36	Female	White	Single	Spanish	San Germa	Dependent	0	9		
1011	1076	45	Male	Other	Married	English	Anasco	Veteran	0	4		
1012	1085	40	Male	African Ar	Married	Spanish	Rincon	Dependent	0	14		
1013	1087	40	Female	White	Married	Spanish	Isabela	Veteran	1	4		
1014	1088	20	Male	African Ar	Married	English	Isabela	Veteran	1	2		
1015	1112	39	Female	Other	Single	Other	San Germa	Veteran	1	17		
1016	1196	28	Male	African Ar	Single	Other	Sabana Gr	Veteran	0	22		
1017	1222	55	Female	Other	Single	Spanish	Hormiguer	Dependent	0	28		
1018	1225	62	Male	African Ar	Single	Spanish	San Germa	Veteran	0	18		
1019	1242	70	Male	African Ar	Married	Spanish	Anasco	Veteran	1	14		
1020	1251	54	Female	Other	Married	Spanish	San Germa	Veteran	0	10		
1021	1266	27	Female	Other		Spanish	Isabela	Dependent	0	8		
1022	1293	51	Male	African Ar	Married	English	San Germa	Veteran	0	12		

Figure 8. Initial View-Interactive Platform

Figure 9 shows the steps that the user follows to create the appointment. After pressing the “Book Patient Appointment” button, the user should indicate the existent patient ID. Using the information (attributes) of the indicated patient, the algorithm run the classification tree prediction model adjusted for that group of patients, and an output box appears with the information of the patient and the respective probability of attendance. Then, the user should agree to make an appointment for the patient.

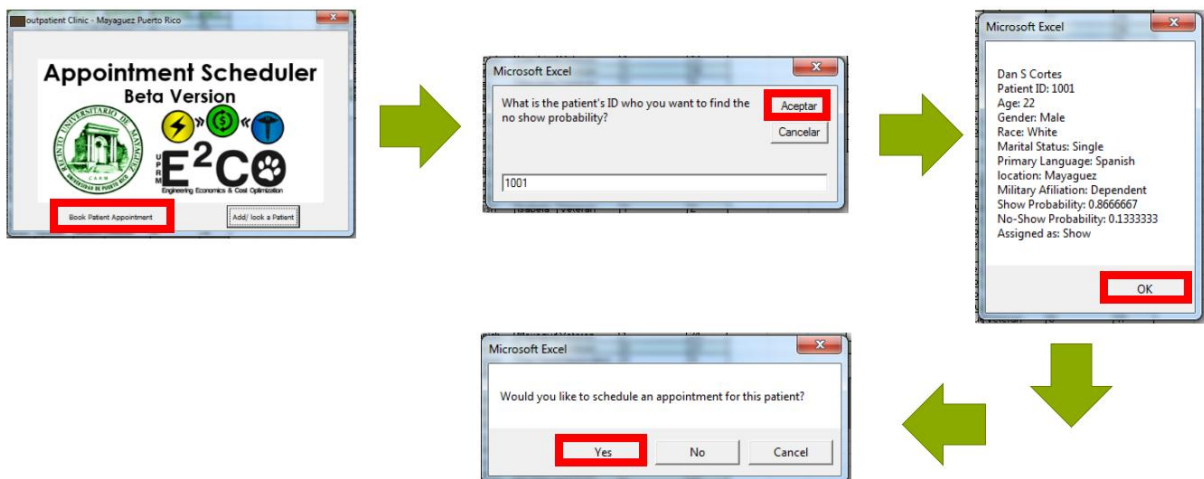


Figure 9. Steps for Scheduling an Appointment-Interactive Platform

The patient assignment process works by booking two patients per appointment slot by following the procedure of Policy #4 in the Test Bed (Section 4.3) where a patient with a high probability of showing is matched with a patient with a low probability of showing. In Figure 10 is presented an illustrative example of the assignment of three patients, whose names are fictitious. The first patient is Dan S Cortes, who is assigned in the first available empty appointment slot. The second patient assigned is Ezequiel Delbert and since is not a probability match with Dan S Cortes it cannot be assigned in the same slot. Therefore, the patient is assigned in the next available slot. Finally, the third patient (Rhett Lynwood) is a match with Dan S Cortes and can be assigned in the same appointment slot. So on, the algorithm continues this logic in the assignment of the patients to appointment slots. This a Beta version of the interactive platform, and it can be subjected to improvements.

Patient Visit Schedule					
Slot #	Time	Monday	Tuesday	Wednesday	Thursday
1	8:00 AM	Dan S Cortes			
2	9:00 AM				
3	10:00 AM				
4	11:00 AM				
5	12:00 AM				
6	1:00 PM				
7	2:00 PM				
8	3:00 PM				
9	4:00 PM				
10	5:00 PM				
Total Cost		\$ 400.59			

Patient Visit Schedule					
Slot #	Time	Monday	Tuesday	Wednesday	Thursday
1	8:00 AM	Dan S Cortes			
2	9:00 AM	Ezequiel Delbert Dicaprio			
3	10:00 AM				
4	11:00 AM				
5	12:00 AM				
6	1:00 PM				
7	2:00 PM				
8	3:00 PM				
9	4:00 PM				
10	5:00 PM				
Total Cost		\$ 416.07			

Patient Visit Schedule					
Slot #	Time	Monday	Tuesday	Wednesday	Thursday
1	8:00 AM	Dan S Cortes Rhett Lynwood Danieli			
2	9:00 AM	Ezequiel Delbert Dicaprio			
3	10:00 AM				
4	11:00 AM				
5	12:00 AM				
6	1:00 PM				
7	2:00 PM				
8	3:00 PM				
9	4:00 PM				
10	5:00 PM				
Total Cost		\$ 424.07			

Figure 10. Patient Assignment to Generate Appointment-Interactive Platform

5 CHAPTER – CONCLUSION AND FUTURE WORK

In this thesis the performance of four appointment scheduling scenarios for outpatient clinics has been evaluated in terms of the total no-show cost to the system. An emphasis was made on evaluating scheduling scenarios constructed by overbooking the appointment slots, which have consequences on the patients' waiting time and personnel overtime. The no-show cost has been assessed by the development of a stochastic cost model to determine the expected value of an appointment slot as a representation of the estimation of the cost of a no-show to the system. The stochastic cost model integrates the patients' probabilities of attendance with an economic analysis. A Classification and Regression Tree approach has been used as a procedure for the prediction of the probabilities of attendance. In the economic analysis, possible drivers for allocation and estimation of costs have been identified for the cost parameters in two cost categories: financial and social costs.

Through an illustrative example it has been demonstrated that, as Lofti and Torres presented in [5], the Classification and Regression Tree (CART) approach is a beneficial technique for the classification and prediction of patients' attendance. To enhance this statement, a Gradient Boosted Tree approach was applied to the same data of the illustrative example, resulting in an improvement in the prediction power of the model of up to twenty percent in terms of the generalized prediction error. The only disadvantage that presents the latter is that the technique provides a black-box model as an output, something that affects the required reproducibility of the model for the application in the prototype of an interactive platform that has been developed for the assistance in the scheduling process of an outpatient clinic.

One of the most relevant contributions of this work is in the consideration of social costs in the economic analysis, and their integration in the cost model. To the best of our knowledge, this has not been addressed in any other research publication. It has been identified that the social costs can be estimated, from the outpatient clinic point of view, in terms of the profit lost by the system due to the higher waiting as a consequence of overbooking, or due to the value lost by the non-utilization of the resources when there is a complete no-show in an appointment slot. Since both aspects are dependent on the attendance of the patient to the appointment, which is uncertain; a deterministic estimate of those costs cannot be done but the probabilities of attendance obtained by the classification techniques CART and Gradient Boosted Trees has been used to estimate the probabilities of patients' overflow from one appointment slot to a successive appointment slot and to estimate the probability of a complete no-show in an appointment slot. As a result, the social costs have been estimated in terms of their expected value. The deterministic financial operational costs have been integrated with the stochastic social costs in a cost model that has been used for the evaluation of appointment scheduling scenarios.

Four appointment scheduling policies for outpatient clinics were evaluated. The simulation results reflect that overbooking in general is beneficial for outpatients' clinics in terms of overall clinic utilization and total cost to the system. It can depends on the overbooking levels utilized. In this work it is being considered the case of two patients overbooking. Also, the results provide insights about the fact that using patients probabilities of attendance as a point of reference at the moment of assigning patients to appointment slots improve the performance of the schedule in terms of the metrics considered in the analysis. Balancing the patients assignment by overbooking using the patients' probabilities of

attendance, reduce the overflows which cause a decrease in the waiting cost and the overtime cost, and at the same time it reduces the number of empty slots that cause a non-utilization of the resources capacity, which involves a cost to the system. It is important to clarify that this is dependent on how good the constructed classification model is in terms of prediction accuracy.

Future work includes a deeper exploration of the two social costs considered in this research work by trying to identify other possible drivers to allocate and estimate them with more accuracy. Furthermore, explore the possibility of integrating other social costs that are more complex to estimate such as the “loss of life” and research how they directly and indirectly affect the system. Currently the cost model has been used only for the evaluation of scheduling scenarios. It may be beneficial to address the patients’ assignment to the appointment slot as an optimization problem with a cost minimization objective function. Also, it is important to highlight that the procedure developed to estimate the overflow probabilities is memoryless, the penalty is the same no matter how many times the patient has overflow from one slot to a consecutive one. As a future step, it could be beneficial to take in consideration the number of times a patient has overflow as a criteria to give a higher penalty or to establish an alternative step to re-schedule those patients with a higher risk of waiting more to receive services.

REFERENCES

1. NHE-Fact-Sheet [Internet]. Center for Medicare and Medicaid Services. 2014 [cited 2015 Feb 19]. Available from: <http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NHE-Fact-Sheet.html>
2. Muthuraman K, Lawley M. A stochastic overbooking model for outpatient clinical scheduling with no-shows. *IIE Transactions*. Taylor & Francis Group; 2008 Jul 21;40(9):820–37.
3. Bech M. The economics of non-attendance and the expected effect of charging a fine on non-attendees. *Health Policy*. 2005 Oct;74(2):181–91.
4. Norris JB, Kumar C, Chand S, Moskowitz H, Shade SA, Willis DR. An empirical investigation into factors affecting patient cancellations and no-shows at outpatient clinics. *Decision Support System*. 2014 Jan;57:428–43.
5. Lotfi V, Torres E. Improving an outpatient clinic utilization using decision analysis-based patient scheduling. *Socio-Economic Planning Sciences*. Elsevier; 2014;48(2):115–26.
6. Moore CG, Wilson-Witherspoon P, Probst JC. Time and money: effects of no-shows at a family practice residency clinic. *The Annals of Family Medicine*. 33(7):522–7.
7. Shaparin N, White R, Andreae M, Hall C, Kaufman A. A longitudinal linear model of patient characteristics to predict failure to attend an inner-city chronic pain clinic. *Journal of Pain*. 2014 Jul;15(7):704–11.
8. Lee VJ, Earnest A, Chen MI, Krishnan B. Predictors of failed attendances in a multi-specialty outpatient centre using electronic databases. *BMC Health Services Research*. 2005 Jan;5(1):51.
9. Parikh A, Gupta K, Wilson AC, Fields K, Cosgrove NM, Kostis JB. The effectiveness of outpatient appointment reminder systems in reducing no-show rates. *American Journal of Medicine*. 2010 Jun;123(6):542–8.
10. Oppenheim GL, Bergman JJ, English EC. Failed appointments: a review. *Journal of Family Practice*. 1979 Apr;8(4):789–96.
11. Hixon AL, Chapman RW, Nuovo J. Failure to keep clinic appointments: implications for residency education and productivity. *The Annals of Family Medicine*. 1999 Oct;31(9):627–30.
12. Stone CA, Palmer JH, Saxby PJ, Devaraj VS. Reducing non-attendance at outpatient clinics. *Journal of the Royal Society of Medicine*. 1999 Mar;92(3):114–8.

13. Rosario K. Appointment no-shows are costly in dollars, time [Internet]. The Official Homepage of the United States Army. 2013 [cited 2015 Feb 19]. Available from: http://www.army.mil/article/116502/Appointment_no_shows_are_costly_in_dollars_time/
14. LaGanga LR, Lawrence SR. Clinic Overbooking to Improve Patient Access and Increase Provider Productivity. *Decision Sciences*. 2007 May;38(2):251–76.
15. Lacy NL, Paulman A, Reuter MD, Lovejoy B. Why we don't come: patient perceptions on no-shows. *The Annals of Family Medicine*. 2(6):541–5.
16. Dove HG, Schneider KC. The usefulness of patients' individual characteristics in predicting no-shows in outpatient clinics. *Medical Care*. 1981 Jul;19(7):734–40.
17. Dravenstott R, Kirchner HL, Strömlad C. Applying Predictive Modeling to Identify Patients at Risk to No-Show. Center for Health Research. 2014.
18. Deyo RA, Inui TS. Dropouts and broken appointments. A literature review and agenda for future research. *Medical Care*. 1980 Nov;18(11):1146–57.
19. Daggy J, Lawley M, Willis D, Thayer D, Suelzer C, DeLaurentis P-C, et al. Using no-show modeling to improve clinic performance. *Journal Health Informatics*. 2010 Dec 1;16(4):246–59.
20. Barron WM. Failed appointments. Who misses them, why they are missed, and what can be done. *Primary Care*. 1980 Dec;7(4):563–74.
21. Frankel S, Farrow A, West R. Non-attendance or non-invitation? A case-control study of failed outpatient appointments. *BMJ*. 1989 May 20;298(6684):1343–5.
22. Giunta D, Briatore A, Baum A. Factors associated with nonattendance at clinical medicine scheduled outpatient appointments in a university general hospital. *Patient Preference and Adherence*. 2013;1163–70.
23. Moser SE. Effectiveness of post card appointment reminders. *The Journal of Family Practice*. 1994 Sep;14(3):281–8.
24. Sharp DJ, Hamilton W. Non-attendance at general practices and outpatient clinics. *BMJ*. 2001 Nov 10;323(7321):1081–2.
25. Gupta D, Denton B. Appointment scheduling in healthcare: Challenges and opportunities. *IIE Transactions*. Taylor & Francis Group; 2008 Jul 21;40(9):800–19.
26. Tsai P-FJ, Teng G-Y. A stochastic appointment scheduling system on multiple resources with dynamic call-in sequence and patient no-shows for an outpatient clinic. *European Journal Operations Research*. 2014 Dec;239(2):427–36.

27. Tang J, Yan C, Cao P. Appointment scheduling algorithm considering routine and urgent patients. *Expert Systems with Application*. 2014 Aug;41(10):4529–41.
28. Mbada CE, Nonvignon J, Ajayi O, Dada OO, Awotidebe TO, Johnson OE, et al. Impact of missed appointments for out-patient physiotherapy on cost, efficiency, and patients' recovery. *Hong Kong Physiotherapy Journal*; 2013.
29. Berg B. P., Murr M., Chermak D., Woodall J., Pignone M., Sandler R. S. and Denton B. T. Estimating the Cost of No-Shows and Evaluating the Effects of Mitigation Strategies. *Medical Decision Making*. 2013: 976-85.
30. Loh W-Y. Classification and regression trees. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2011 Jan 6;1(1):14–23.
31. Data Mining: Practical Machine Learning Tools and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems): 9780123748560: Computer Science Books @ Amazon.com
32. Gordon L. SAS Global Forum 2013 Data Mining and Text Analytics Using Classification and Regression Trees (CART) in SAS ® Enterprise Miner TM For Applications in Public Health . Leonard Gordon , University of Kentucky , Lexington , KY SAS Global Forum 2013 Data Mi. 2013;(Gordon 2010):1–8.
33. Shepard D, Hodgkin D, Anthony Y. Analysis of Hospital Costs: A Manual for Managers. Institute for Health Policy-Heller School Brandeis University. 1998 Sept.
34. Tan, Pang, Michael Steinbach, and Vipin Kumar. Classification:Basic Concepts, Decision Trees, and Model Evaluation. *Introduction to Data Mining*. Boston: Pearson Addison Wesley, 2005.
35. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. "Three Based Methods." *An Introduction to Statistical LEarning with Applications in R*. Springer, 2013. 303-330.
36. "What Is R?" R: The Foundation <<https://www.r-project.org/about.html>>.
37. Therneau, Terry M., and Elizabeth J. Atkinson. An Introduction to Recursive Patitioning Using the RPART Routines. Mayo Foundation, 29 June 2015. Web
38. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. 1999a Feb.
39. Ridgeway, Greg. Generalized boosted Models: A guide to the gbm package. 2007 Aug.

40. McLean, Robert A. Cost Behavior. Financial Management in Health Care Organizations. 2nd ed. Clifton Park, NY: Delmar Learning, 2003.
41. Hansen, Don R., Maryanne M. Mowen, and Liming Guan. Cost Management Accounting and Control. 6th ed. Mason, Ohio: Thomson/South-Western, 2007.
42. Park, Chan S. Contemporary Engineering Economics. 4th ed. Upper Saddle River, N.J.: Prentice Hall, 2005.
43. Schappert, S. and Rechsteiner, E. Ambulatory Medical Care Utilization Estimates for 2006. 2008 Aug. <<http://www.cdc.gov/nchs/data/nhsr/nhsr008.pdf#8>>.
44. Sección 16. Derechos del Empleado. Artículo I Constitución del Estado Libre Asociado de Puerto Rico.
45. Hayter, Anthony J. Random Variables & Counting Techniques. Probability and Statistics for Engineers and Scientists. 2nd ed. Boston: PWS Pub., 1996.

APPENDIX A

A. 1 Classification and Regression Trees Steps

A.1.1 Tree Growing

Tree growing deals with two main issues. The first one is related to the splitting process, which depends on the attributes types and the number of ways to split. In the case of CART, only binary splits are allowed. Also, it depends on the attribute test condition, which is the term used to refer to the determination of the attribute that provides the best split. This is assessed through the use of impurity measures as metrics of the homogeneity of the nodes. Three main impurity based criteria can be mentioned: Classification error, Gini index, and Entropy. The classification error simply measures the fraction in the region, that do not belong to the most common class, this is because in classification trees the purpose is to divide the observations in regions and assign a record in a region to the most commonly occurring class [35]. The Gini index is a measure of the total variance across the classes, taking a smaller value as the proportion of the training records in the node that are from a particular class is close to 0 or to 1. This implies that the observations in the node tends towards a particular class, something that is beneficial for classification matters. Entropy is another alternative, numerically similar to the Gini Index. Entropy and Gini Index are more sensitive to node purity, reason why are more used in tree growing than the classification error [35]. The following are the equations for the calculation of the three impurity based criterions:

$$\text{Classification Error} = 1 - \max_i(\hat{p}_{ik}) \quad [\text{Equation 26}]$$

$$\textbf{Gini Index} = \sum_{i=1}^i [\hat{p}_{ik} (1 - \hat{p}_{ik})] \quad [\text{Equation 27}]$$

$$\textbf{Entropy} = - \sum_{i=1}^i \hat{p}_{ik} \log \hat{p}_{ik} \quad [\text{Equation 28}]$$

Where \hat{p}_{ik} is the proportion of records that belong to the class i of the node k .

The second main issue in tree growing is related to the stopping criteria. A tree can be expanded completely until all the records belong to the same class or all the records have identical attribute values [34]. However, several stopping conditions can be followed that go from simple early stopping rules to more restrictive rules. This has its benefit since in general smaller trees with small prediction errors are preferred to avoid over-fitting. Model over-fitting is a phenomenon that occurs as the tree becomes too large and the test error rate (generalization error) increases while the training error increases [34]. This is due to the fact that a model can fit very well the training records, but has a poorer performance with unseen records (test records). A model with a good training error can present a poorer generalization error than a model with a higher training error [34].

A.1.2 Tree Pruning

There is two pruning strategies, pre-pruning and post-pruning, which differ in the stopping criteria. Pre-pruning is an early stopping rule, where the algorithm stops before the tree is fully grown. This is done by using a restrictive stopping condition as a gain in the impurity measure or when the estimated generalization error value is lower than a threshold value (α) established. Determining the α value may be a challenge since a high value can result in under-fitting, but a value too low may result in over-fitting. Post-pruning is a strategy where the tree is fully grown and then it is

trimmed back in order to obtain a subtree with a lower test error rate. This strategy usually tends to present better results than pre-pruning, due to the premature termination of the pre-pruning strategy.

A.1.3 Tree Performance Validation

CART applies a learning algorithm to identify a model that best fits the relationship between the response variable and the prediction attributes. Basically, the classification technique works by applying the learning algorithm to an initial data set known as **training set**, which contains records whose class labels are known, and constructing a classification model [34]. Then the classification model is applied to a different data set, known as the **test set**, in order to assess the model prediction power. Cross-validation is a method that helps when it is available a limited amount of data that difficult the extraction a considerable test set that could be used to estimate the testing error rate. It consists of holding out a subset of the training observations from the fitting process, and using that held out observations for testing purposes [35]. From the several cross-validation techniques, the K-Fold Cross Validation has been selected because offers the advantage of allowing to choose the size of each test set is and how many trials are average over. It involves dividing the data set in k groups of equal size, preferably. One fold is selected as the validation set for testing purposes, and the rest $k-1$ folds are used to fit the model. The procedure is repeated k times, alternating the folds so each time a different group is used as a validation set.

A classification model performance is assessed based on the number of records that were correctly and incorrectly predicted. One way to represent the results and tabulate the predictions is through the construction of a confusion matrix. Figure 11

presents an illustrative example of a confusion matrix for a classification model with a binary response {Show (0), No-Show (1)}.

		Condition	
		Show	No-Show
Test Outcome	Show	True Positive	False Positive (Type I Error)
	No-Show	False Negative (Type II Error)	True Negative

Figure 11. Illustrative Example of a Confusion Matrix

Each entry of the confusion matrix provides the number of records that were predicted as class i , but their true condition was class j . For example, the records on the True Positive entry were predicted as Show and their true condition was Show. In other words, those records were correctly predicted by the classification model. It would be convenient to express the performance of the model by using a single value. The following performance metrics (Equations 29-32) can be calculated from the confusion matrix entries in order to assess the power of the model.

$$\text{Sensitivity} = \frac{\# \text{ of records True Positive}}{\# \text{ of records Condition Show}} \quad [\text{Equation 29}]$$

$$\text{Specificity} = \frac{\# \text{ of records True Negative}}{\# \text{ of records Condition No-Show}} \quad [\text{Equation 30}]$$

$$\text{Show Predictive Value} = \frac{\# \text{ of records True Positive}}{\# \text{ of records Test Outcome Show}} \quad [\text{Equation 31}]$$

$$\text{No-Show Predictive Value} = \frac{\# \text{ of records True Negative}}{\# \text{ of records Test Outcome No-Show}} \quad [\text{Equation 32}]$$

Sensitivity (Equation 29) and Specificity (Equation 30) assess the portion of patients correctly classified as show and no-show, respectively. The Show Predictive Value

(Equation 30) calculates the proportion of records that were correctly classified as Show, from all the records that were classified with this class label. The No-Show Predictive Value (Equation 31) provides similar information but for the case of the records classified as No-Shows. These four performance metrics are used to assess the prediction power of the classification model for each individual response class. The overall prediction power of the model can be quantified by calculating the Error Rate shown in Equation 33.

$$\mathbf{Error\ Rate} = 1 - \mathbf{Accuracy} = 1 - \frac{\mathbf{True\ Positive} + \mathbf{True\ Negative}}{\mathbf{Total\ No.\ of\ Predictions}} \quad [\text{Equation 33}]$$

When a k-fold cross validation method is used, the error rate is the average of the errors obtained for each fold (Equation 34).

$$\mathbf{CV\ Error\ Rate} = \frac{1}{k} \sum_{i=1}^k \mathbf{Error\ Rate}_i \quad [\text{Equation 34}]$$

Since two data sets are involved in the process, training data and test data, two types of error rates can be calculated. Both performance metrics are calculated using Equation 34, changing the data used. The first one is the Resubstitution Error Rate (RE), which is known as an optimistic performance metric because it tests the performance of the model by using the same training data that was used to fit it. The second one is the Testing Error Rate (TE), considered a pessimistic measure because it tests the validity of the model with an independent data set that was not involved in the model fitting process. With the purpose of balancing and considering both types of error rates, the Generalized Error Rate (Equation 35) is calculated by performing a weighted average where, in the case of this research, the weight value w is assigned as the proportion of the data set that contains each set (training and testing).

$$\textbf{Generalized Error Rate} = w_{Testing} * RE + w_{Training} * TE \quad [\text{Equation 35}]$$

A.2 Gradient Boosted Trees

Boosting is a slow learning approach that improves the prediction power of a resulting decision tree. It works on producing a prediction model by growing sequential weak models, where each tree is grown using information from previously grown trees. Each tree is fit on a modified version of the original data set [35]. Gradient Boosting is an approach developed by Jerome H. Friedman, which consists of applying the boosting methodology by constructing each successive tree based on the prediction residuals of the previous constructed tree. The technique optimizes for an arbitrary differentiable *loss function*.

Gradient boosted trees require the specification of three essential tuning parameters:

- ***Number of trees to be generated successively:*** Choosing a value too large can result in over-fitting, reason why a value of one hundred trees is typically selected and then cross-validation is used to select the optimal value [35].
- ***Shrinkage parameter:*** Controls the rate at which the method learns. The typical values of this parameter are 0.01 or 0.001 [35].
- ***Number of splits on each tree generated:*** This parameter is considered as the interaction depth parameter because it controls the interaction order of the model, since the number of splits implies the number of variables involved.

Basically, the technique works by generating an initial tree by taking in consideration the tuning parameters and by selecting the best partitioning of the data according to the impurity measure selected to assess this. Then, the residuals of the resulting tree are calculated by computing the deviations of the predicted values from the mean values. The next tree is then fitted to the residuals of the previous tree. This procedure is repeated until the number of trees specified as a parameter is reached. On each iteration, the technique is searching for the best partition that will further decrease the error of the data. At the end, the model is a black box because it cannot be tree-based represented since it is composed of the aggregation of sequential trees.

The procedure just described is the general methodology of the technique and it applies for regression modeling. However, when dealing with a classification problem, there are several differences in the methodology. Friedman developed an alternate procedure for multi-class classification problems. Details about the algorithm developed by Friedman can be accessed in his publication “Greedy Function Approximation: A Gradient Boosting Machine” published in 1999 [38]. In general, the procedure generates a different boosting tree for each class of the categorical dependent variable. Also, for each one, it creates a vector of values $[0,1]$ to indicate if an observation in the data belongs or not to the respective class. The algorithm applies a logistic transformation to compute the residuals. The process continues for the number of trees indicated as parameters. At the end, in order to obtain the classification probabilities, a logistic transformation is applied to the prediction in the vector of prediction $[0, 1]$ of each class.

A.3 R packages for CART and Gradient Boosted Trees

A.3.1 R-part

A data mining technique as classification trees, which includes the analysis of an extensive amount of data, requires the use of a statistical software that facilitate the task. R is a computer language and environment for statistical computing and graphics, which is available as a Free Software in source code form [36]. It works through a simple programming language that facilitates data manipulation, calculations and the creation of illustrative outputs. It is an environment where statistical techniques are implemented; most of them are available via packages that can be downloaded through the CRAN family of internet sites. R-part, whose name comes from the phrase “recursive partitioning”, is an R package that consists of routines that implement the ideas of CART as presented by Breiman, Friedman, Olshen and Stone in books and programs [37]. The program builds classification and regression tree models with binary responses.

R-part works following the three steps explained in Appendix A, Section A.1. First, the variable which best splits the data according to the impurity based criterion used is selected; the Gini Index is the splitting index being used for this analysis. The data is partitioned, then the process is applied to each sub-group. This procedure goes on until a stopping criterion is reached or no further improvement can be achieved. The purpose of using an impurity based criterion is risk reduction. In other words, reducing the risk of classifying incorrectly the records of a particular partition. R-part also implements the Altered Prior Method for the calculation of the expected loss, which is the probability of misclassification, to assess the risk reduction criteria.

Altered Priors are node probabilities that accounts for the initial prior probabilities, which are equal to the observed class frequencies in the original sample; for the loss matrix when incorrectly classifying a record of class i as class j ; and for the proportion of the cases included in the node. The altered priors impact the choice of split by assisting the impurity rule in the process of choosing the splits that are likely to be good in terms of the misclassification risk [37]. As a second step, cross-validation is used to trim back the full tree by selecting the tree with the smallest possible number of nodes that reduces the misclassification risk. Finally, the validation step can be performed by generating the confusion matrix and calculating the performance metrics mentioned in Section A.1, Sub-section A.1.3.

A.3.2 Caret and gbm

Caret, which is an abbreviation for “classification and regression training”, is a package in R which contains functions that help in the process of model training complex regression and classification problems. This package is composed of several number of R packages, an approximate of twenty-six packages. From all the available packages, the *gbm* package is of interest because it fits a generalized boosted regression or classification model, based on Friedman’s gradient boosting methodology. Basically, the package applies the methodology described in Section A.2, based on the specification of several parameters. In this section the essential ones will be explained. The first parameter that should be specified is the *distribution*, which defines the methods for computing the associated deviance, initial value, the gradient, and the constants to predict in each terminal node [39]. The second parameter is *n.trees*, which is the number of trees that should be generated for the

additive expansion. The third and fourth parameters are the *interaction.depth* and the *shrinkage*. The package allows to construct a black box model and allows to plot it in order to select the optimal *n.trees* value to avoid overfitting. The predictions of the resulting model can be summarized as a confusion matrix and the results can be analyzed using the performance metrics in Section A.1, Sub-section A.1.3., as it is done for the CART methodology.

APPENDIX B

B.1 CART Results-Representative Example

Using the r-part package from R (refer to Appendix A, Section A.3, Sub-section 1.3.1), a model was adjusted for each data set of each example, as follows:

```
cfit<-rpart(p_att~ l_time+age+gender+race+m_status+p_language+  
distance+t_patient,data=df,method="class")
```

The code line expresses that the prediction model it is being constructed considering the patients' attendance (p_att) as the outcome and considering the eight attributes, mentioned earlier, as prediction variables. Since it is a classification model, the method is specified as "class".

A cross-validation approach was used, dividing the data sets (500 samples) in five folds (k=5) of 100 samples each. The analysis was held for each data set by taking k-1 folds for training the model and one fold for testing. The procedure was repeated k times, alternating the folds (identified as F_i) so each time a different group is used as a validation set (Figure 12).

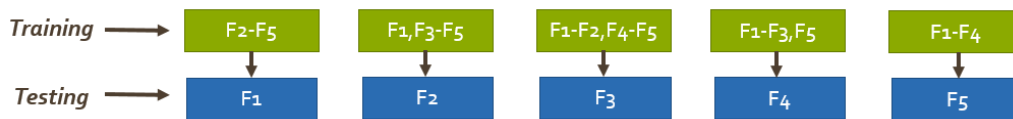


Figure 12. Representation of the Folding Procedure

On each fold, the model has been constructed using the training folds. The prediction power of the model has been evaluated using the training data and then using the testing data. A confusion matrix has been generated for each case, in order to calculate the performance

measures of interest for each folding. From all the performance measures mentioned in Appendix A, Section A.1, Sub-section A.1.3, the Generalized Error Rate is the one of interest, since it is desired to assess the prediction power of the model in relation to the two responses (Show, No-Show). The cross-validated resubstitution error has been calculated for each data set by averaging the resubstitution error of each fold (refer to Equation 9). The same procedure has been done for the cross-validated testing error rate, but using the testing error rate of each fold. Then for each data set, the cross-validated general error rate has been computed as a weighted average of the cross-validated resubstitution error and the cross-validated testing error, as presented in Equation 10 in Appendix A, Section A.1, Sub-section A.1.3. The weight parameter (w) has been selected as the proportion of the data that constitutes the training data and the testing data. In this case, on each fold, four fifths of the data is being used for training and one fifth for testing. Figures 13 and 14 present an example of the confusion matrix and the performance measure calculation for a data set, for the initial fully grown tree, in the case of the unknown pattern data example. This procedure was done for the model constructed using the initial fully grown trees and then for the model constructed after pruning it, with the purpose of evaluating how the error rates changes. The pruning process was done for each fold of each data set. The tree with the number of splits that provided the lower cross-validated error was selected. This was evaluated based on the Cross-validated Error vs. Tree Size plot that r-part provides as part of the analysis. Figure 15 presents an illustration of the plot.

Classification Error	Training F2-F5		Training F1, F3-F5		Training F1-F2,F4-F5		Training F1-F3, F5		Training F1-F4	
	Show	No-Show	Show	No-Show	Show	No-Show	Show	No-Show	Show	No-Show
Show	151	59	149	61	125	47	152	66	140	61
No-Show	54	136	53	137	70	158	56	126	70	129
Positive Predictive Value	0.72		0.71		0.73		0.70		0.70	
Negative Predictive Value	0.72		0.72		0.69		0.69		0.65	
Sensitivity	0.74		0.74		0.64		0.73		0.67	
Specificity	0.70		0.69		0.77		0.66		0.68	
Resubstitution Error	0.28		0.29		0.29		0.31		0.33	
Cross-validated R.Error	0.30									

Figure 13. Example-Confusion Matrix and Performance Measures using Training Data for the Unknown Pattern Data Example

Classification Error	Testing F1		Testing F2		Testing F3		Testing F4		Testing F5	
	Show	No-Show	Show	No-Show	Show	No-Show	Show	No-Show	Show	No-Show
Show	30	26	29	23	19	15	26	32	26	29
No-Show	20	24	24	24	41	25	21	21	19	26
Positive Predictive Value	0.54		0.56		0.56		0.45		0.47	
Negative Predictive Value	0.55		0.50		0.38		0.50		0.58	
Sensitivity	0.60		0.55		0.32		0.55		0.58	
Specificity	0.48		0.51		0.63		0.40		0.47	
Testing Error	0.46		0.47		0.56		0.53		0.48	
Cross-validated Testing Error	0.50									

Figure 14. Example-Confusion Matrix and Performance Measures using Testing Data for the Unknown Pattern Data Example

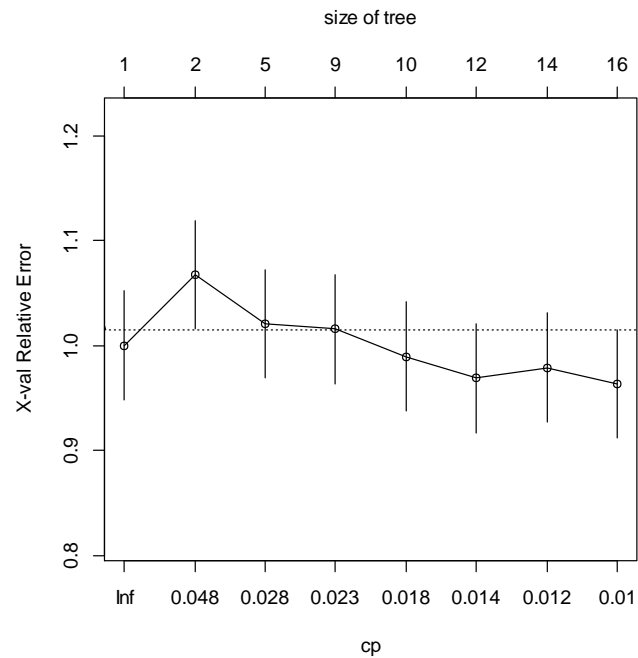


Figure 15. Example-Cross-Validated Error vs. Tree Size Plot

The results for each data set, for each example, before and after pruning have been summarized in a table format. Refer to Tables 9, 10, 11 and 12 for the summaries.

Table 9. Error Rates for Fully Grown Trees-Unknown Pattern Data Example

Sample	Cross-Validated Resubstitution Error Rate	Cross-Validated Testing Error Rate	Generalized Cross- Validated Error Rate
1	0.297	0.424	0.399
2	0.295	0.506	0.464
3	0.293	0.486	0.447
4	0.301	0.548	0.499
5	0.297	0.504	0.463
6	0.288	0.478	0.440
7	0.301	0.526	0.481
8	0.307	0.556	0.506
9	0.284	0.474	0.436
10	0.299	0.456	0.425
11	0.324	0.518	0.479
12	0.299	0.494	0.455
13	0.313	0.504	0.466
14	0.289	0.526	0.479
15	0.285	0.474	0.436
16	0.287	0.502	0.459
17	0.303	0.530	0.485
18	0.305	0.526	0.482
19	0.284	0.528	0.479
20	0.291	0.500	0.458
21	0.283	0.524	0.476
22	0.314	0.528	0.485
23	0.289	0.534	0.485
24	0.299	0.548	0.498
25	0.289	0.506	0.463
26	0.312	0.406	0.387
27	0.301	0.444	0.415
28	0.282	0.504	0.460
29	0.299	0.500	0.460
30	0.303	0.498	0.459

Table 10. Error Rates for Prune Trees-Unknown Pattern Data Example

Sample	Cross-Validated Resubstitution Error Rate	Cross-Validated Testing Error Rate	Generalized Cross- Validated Error Rate
1	0.329	0.492	0.459
2	0.392	0.506	0.483
3	0.423	0.470	0.461
4	0.421	0.510	0.492
5	0.429	0.476	0.467
6	0.419	0.422	0.421
7	0.385	0.528	0.499
8	0.397	0.496	0.476
9	0.384	0.444	0.432
10	0.447	0.496	0.486
11	0.408	0.506	0.486
12	0.335	0.530	0.491
13	0.417	0.454	0.447
14	0.393	0.522	0.496
15	0.361	0.468	0.447
16	0.401	0.512	0.490
17	0.357	0.497	0.469
18	0.401	0.510	0.488
19	0.405	0.474	0.460
20	0.340	0.500	0.468
21	0.327	0.554	0.509
22	0.431	0.506	0.491
23	0.355	0.500	0.471
24	0.466	0.466	0.466
25	0.378	0.424	0.415
26	0.379	0.424	0.415
27	0.398	0.460	0.448
28	0.393	0.502	0.480
29	0.360	0.536	0.501
30	0.365	0.482	0.459

The unknown pattern data example resulted in bigger fully grown trees that did not presented an evident patterned behavior in the attributes splitting. This is reflected in the high cross-validated testing error rates, which are near to or higher than 0.50, something that is expected for random data with no traceable pattern. The Generalized Cross-Validated Error Rate ranged from 0.387 to 0.506. These values are considerable higher than desired for the assessment of the predictive power of a classification model. However, samples 1

and 26 presents slightly lower cross-validated error rates, something that highlight a possible pattern detected by the technique. For example, checking the data from sample 1, it can be identified that patients of gender B and race B and C had a higher tendency of failing an appointment with a 63% being no shows. Also, 60% of the patients younger than forty-eight year ($\text{age} < 48$) that booked their appointments more than thirteen days before the appointment day ($\text{lead time} > 30$) were no shows. Patients with a marital status B and primary language A failed the appointments with a 52%. According to the data patients of type A that lives near to the clinic ($\text{distance} \leq 8$ miles) represent 73% of the no-shows. This is contrary to what the studies have presented, patients that live far from the clinic tend to have high no-show incidence, but it is important to remember that this is a hypothetical representative example. Those are example of the unknown pattern of the data that the technique could have captured and represented in the final prediction model.

It was expected to achieve lower values for the prune trees but the values where higher ranging from 0.415 to 0.509. The biggest percentage of difference is fifteen percent, which is a lower value considering that a smallest tree provides a prediction power very similar to a fully grown tree.

Table 11. Error Rates for Fully Grown Trees-Patterned Data Example

Sample	Cross-Validated Resubstitution Error Rate	Cross-Validated Testing Error Rate	Generalized Cross- Validated Error Rate
1	0.220	0.304	0.287
2	0.219	0.328	0.306
3	0.136	0.136	0.136
4	0.201	0.378	0.343
5	0.215	0.476	0.424
6	0.200	0.400	0.360
7	0.196	0.414	0.370
8	0.181	0.400	0.356
9	0.188	0.478	0.420
10	0.180	0.306	0.281
11	0.219	0.538	0.474
12	0.192	0.460	0.406
13	0.194	0.372	0.336
14	0.204	0.348	0.319
15	0.193	0.396	0.355
16	0.217	0.516	0.456
17	0.190	0.448	0.396
18	0.176	0.406	0.360
19	0.198	0.448	0.398
20	0.189	0.342	0.311
21	0.203	0.378	0.343
22	0.202	0.430	0.384
23	0.168	0.290	0.266
24	0.188	0.354	0.321
25	0.204	0.408	0.367
26	0.197	0.336	0.308
27	0.198	0.392	0.353
28	0.183	0.342	0.310
29	0.209	0.486	0.431
30	0.201	0.382	0.346

Table 12. Error Rates for Prune Trees-Patterned Data Example

Sample	Cross-Validated Resubstitution Error Rate	Cross-Validated Testing Error Rate	Generalized Cross- Validated Error Rate
1	0.256	0.256	0.256
2	0.249	0.300	0.290
3	0.136	0.136	0.136
4	0.254	0.254	0.254
5	0.282	0.282	0.282
6	0.260	0.260	0.260
7	0.245	0.252	0.251
8	0.244	0.306	0.294
9	0.264	0.264	0.264
10	0.218	0.218	0.218
11	0.244	0.560	0.497
12	0.227	0.368	0.340
13	0.226	0.324	0.304
14	0.238	0.238	0.238
15	0.235	0.302	0.289
16	0.263	0.466	0.425
17	0.266	0.266	0.266
18	0.251	0.268	0.265
19	0.243	0.358	0.335
20	0.216	0.298	0.282
21	0.258	0.260	0.260
22	0.257	0.322	0.309
23	0.216	0.216	0.216
24	0.234	0.246	0.244
25	0.271	0.262	0.264
26	0.234	0.234	0.234
27	0.252	0.252	0.252
28	0.238	0.238	0.238
29	0.268	0.342	0.327
30	0.253	0.262	0.260

In contrast to the unknown pattern data example, the technique captured the patterns integrated in the patterned data example. The data was managed creating a pattern of no-show for patients below thirty years old who live further than fourteen miles from the clinic. As a result, the fully grown trees were smaller and the attributes age and distance resulted with the biggest importance on the splitting process. Also, the prediction error of these examples was significantly smaller. The Generalized Cross-validated Error Rate ranged

from 0.136 to 0.474, for the fully grown trees, and ranged from 0.136 to 0.497 for the prune trees.

B.2 Gradient Boosted Trees Results-Representative Example

Using the *gbm* package included in the *caret* package, a black-box model has been constructed for each data set, maintaining the five-fold cross-validation approach. The results have been represented as a confusion matrix and used to compute the performance metrics of interest.

Four tuning parameters were specified in order to run the analysis. First of all is the *distribution*, and since in this research work it is being assessed is a binary classification problem, the *distribution* selected is Bernoulli. The second parameter specified is *n.trees*, which is the number of trees that should be generated for the additive expansion. For this first trial a value of *n.trees* equal to one-hundred has been selected. The *shrinkage* parameter has been set to 0.001, which is one of the typical values used [35]. A value smaller for the *shrinkage* could require a very large value of *n.trees* in order to achieve a good performance in the prediction model. The *interaction.depth* has been set to a value of 1, implying that multiple small trees will be generated with one split. In boosting, using smaller trees is sufficient due to the fact that the growth of successive trees is dependent of the results from the trees grown previously [35]. The final model is an additive model of the small trees generated.

Table 13. Comparison Error Rates - CART Prune Trees vs. GBM Trees

Generalized Cross-Validated Error Rate-Random Data Example			
Sample	CART-Prune Trees	GBM	% Reduction
1	0.459	0.445	3.05
2	0.483	0.447	7.45
3	0.461	0.457	0.87
4	0.492	0.478	2.85
5	0.467	0.441	5.57
6	0.421	0.434	-3.09
7	0.499	0.460	7.82
8	0.476	0.449	5.67
9	0.432	0.401	7.18
10	0.486	0.392	19.34
11	0.486	0.459	5.56
12	0.491	0.420	14.46
13	0.447	0.470	-5.15
14	0.496	0.430	13.31
15	0.447	0.432	3.36
16	0.490	0.461	5.92
17	0.469	0.460	1.92
18	0.488	0.429	12.09
19	0.460	0.471	-2.39
20	0.468	0.421	10.04
21	0.509	0.440	13.56
22	0.491	0.460	6.40
23	0.471	0.480	-1.91
24	0.466	0.470	-0.86
25	0.415	0.415	0.00
26	0.415	0.388	6.51
27	0.448	0.418	6.70
28	0.480	0.420	12.50
29	0.501	0.453	9.58
30	0.459	0.464	-1.09

Table 13 presents a summary of the results for the Generalized Cross-Validated Error for both analysis, CART and GBM. Both methods are compared in terms of the percent of reduction in the value of the performance metric. It can be seen that in twenty-five of the thirty data sets samples, a reduction in the error rate has been achieved, with a maximum of approximate twenty percent and a minimum of zero percent of reduction. Five of the thirty data sets resulted in an increase in the generalized cross-validated error rate, with a maximum increase of approximate 5%

and a minimum increase of one percent; which are small values taking in consideration that is an illustrative example with data generated with an unknown pattern. In general, it can be seen that a considerable increase in the accuracy of the prediction model can be achieved by applying an additive modelling classification trees methodology such as gradient boosted trees. A reduction of up to twenty percent in the error rate is significant.

APPENDIX C

Table 14. Replication Runs for Policy 1

Replicate	Performance Measures					
	Total Assigned Empty Slots	Total Assigned Overbooked Slots	Total Real Empty Slots	Total Overflowed Slots	Total Expected Cost	Total Actual Cost
1	245	0	245	0	26297.49	26054.02
2	245	0	245	0	26297.49	26054.02
3	245	0	245	0	26297.49	26054.02
4	245	0	245	0	26297.49	26054.02
5	245	0	245	0	26297.49	26054.02
6	245	0	245	0	26297.49	26054.02
7	245	0	245	0	26297.49	26054.02
8	245	0	245	0	26297.49	26054.02
9	245	0	245	0	26297.49	26054.02
10	245	0	245	0	26297.49	26054.02
11	245	0	245	0	26297.49	26054.02
12	245	0	245	0	26297.49	26054.02
13	245	0	245	0	26297.49	26054.02
14	245	0	245	0	26297.49	26054.02
15	245	0	245	0	26297.49	26054.02
16	245	0	245	0	26297.49	26054.02
17	245	0	245	0	26297.49	26054.02
18	245	0	245	0	26297.49	26054.02
19	245	0	245	0	26297.49	26054.02
20	245	0	245	0	26297.49	26054.02
21	245	0	245	0	26297.49	26054.02
22	245	0	245	0	26297.49	26054.02
23	245	0	245	0	26297.49	26054.02
24	245	0	245	0	26297.49	26054.02
25	245	0	245	0	26297.49	26054.02
26	245	0	245	0	26297.49	26054.02
27	245	0	245	0	26297.49	26054.02
28	245	0	245	0	26297.49	26054.02
29	245	0	245	0	26297.49	26054.02
30	245	0	245	0	26297.49	26054.02
31	245	0	245	0	26297.49	26054.02
32	245	0	245	0	26297.49	26054.02
33	245	0	245	0	26297.49	26054.02
34	245	0	245	0	26297.49	26054.02
35	245	0	245	0	26297.49	26054.02
36	245	0	245	0	26297.49	26054.02
37	245	0	245	0	26297.49	26054.02
38	245	0	245	0	26297.49	26054.02
39	245	0	245	0	26297.49	26054.02
40	245	0	245	0	26297.49	26054.02
41	245	0	245	0	26297.49	26054.02
42	245	0	245	0	26297.49	26054.02
43	245	0	245	0	26297.49	26054.02
44	245	0	245	0	26297.49	26054.02
45	245	0	245	0	26297.49	26054.02
46	245	0	245	0	26297.49	26054.02
47	245	0	245	0	26297.49	26054.02
48	245	0	245	0	26297.49	26054.02
49	245	0	245	0	26297.49	26054.02
50	245	0	245	0	26297.49	26054.02

Table 15. Statistics Results for the Replication Runs of Policy 1

Statistic	Performance Measures					
	Total Assigned Empty Slots	Total Assigned Overbooked Slots	Total Real Empty Slots	Total Overflowed Slots	Total Expected Cost	Total Actual Cost
<i>Average</i>	245	0	245	0	26297.49	26054.02
<i>Std. Dev.</i>	0	0	0	0	0	0
<i>Number of Replicates (n)</i>	0	0	0	0	0	0
<i>Confidence Interval (h)</i>	0	0	0	0	0	0

Table 16. Replication Runs for Policy 2

Replicate	Performance Measures					
	Total Assigned Empty Slots	Total Assigned Overbooked Slots	Total Real Empty Slots	Total Overflowed Slots	Total Expected Cost	Total Actual Cost
1	115	37	83	80	20795.68	20215.49
2	118	40	87	85	20811.88	21488.50
3	121	43	92	83	20851.53	21070.13
4	113	35	87	70	20811.46	19971.80
5	118	40	86	90	20767.57	20854.26
6	118	40	86	81	20771.14	19994.80
7	120	42	89	89	20843.86	21240.26
8	120	42	86	78	20742.40	19691.46
9	119	41	95	81	20708.42	20849.44
10	125	47	93	93	20793.45	21148.24
11	118	40	87	84	20810.20	20527.92
12	122	44	89	88	20777.69	20378.52
13	125	47	93	96	20748.92	21550.42
14	121	43	91	83	20815.66	21090.86
15	115	37	90	83	20787.32	20863.36
16	116	38	84	81	28147.56	19939.70
17	124	46	94	88	20828.01	21630.80
18	114	36	86	80	20744.43	20295.87
19	125	47	85	91	29359.11	20826.71
20	120	42	83	80	20761.35	19861.59
21	117	39	84	86	28796.94	20748.61
22	121	43	92	71	20772.83	19957.88
23	124	46	91	95	20815.19	22101.99
24	126	48	86	85	20769.32	20750.88
25	122	44	89	95	20772.30	21189.70
26	130	52	92	89	28296.85	21270.08
27	122	44	87	92	29938.01	20831.26
28	117	39	84	83	28030.85	20440.72
29	123	45	90	91	28137.91	21419.48
30	123	45	88	85	20682.15	21061.04
31	121	43	91	98	20790.27	21902.04
32	119	41	87	83	20750.00	20121.19
33	123	45	88	85	28795.30	20757.70
34	121	43	84	86	20721.95	20293.60
35	119	41	89	79	20873.57	20732.42
36	115	37	85	94	29292.99	21178.34
37	120	42	87	73	20797.28	20020.08
38	121	43	88	88	20817.68	20909.37
39	123	45	86	84	20771.34	20197.03
40	123	45	86	84	27927.03	20194.76
41	127	49	91	93	27998.31	21699.81
42	127	49	86	98	28177.15	21814.84
43	111	33	88	70	29648.52	19895.96
44	126	48	88	103	20739.40	22072.17
45	118	40	87	90	29400.88	20932.37
46	132	54	90	95	28786.33	21263.26
47	116	38	89	67	20718.02	20277.41
48	119	41	86	94	20792.92	21005.93
49	121	43	85	92	20796.50	21028.94
50	120	42	88	90	20819.16	21111.59

Table 17. Statistics Results for the Replication Runs of Policy 2

Statistic	Performance Measures					
	Total Assigned Empty Slots	Total Assigned Overbooked Slots	Total Real Empty Slots	Total Overflowed Slots	Total Expected Cost	Total Actual Cost
<i>Average</i>	120.68	42.68	87.96	86.04	23342.09	20813.41
<i>Std. Dev.</i>	4.23	4.23	2.93	7.79	3786.34	622.71
<i>Number of Replicates (n)</i>	2	16	2	13	43	1
<i>Confidence Interval (h)</i>	1.20	1.20	0.83	2.21	1076.83	177.10

Table 18. Replication Runs for Policy 3

Replicate	Performance Measures					
	Total Assigned Empty Slots	Total Assigned Overbooked Slots	Total Real Empty Slots	Total Overflowed Slots	Total Expected Cost	Total Actual Cost
1	62	67	35	145	21684.24	22979.31
2	57	62	27	146	21620.72	22993.42
3	58	63	25	157	21488.28	22415.55
4	64	69	35	130	21445.92	23232.09
5	61	66	30	145	21546.92	22849.10
6	64	69	27	136	21404.37	22942.86
7	61	66	32	145	21576.21	22415.84
8	57	62	39	131	21618.61	23376.70
9	63	68	27	139	21507.07	21780.06
10	61	66	38	130	21619.86	24402.24
11	60	65	31	137	21489.99	22683.02
12	64	69	35	130	21519.96	22625.41
13	54	59	30	142	21557.46	23759.12
14	61	66	35	149	21482.79	25355.47
15	67	72	32	147	21413.86	23780.87
16	56	61	33	131	21416.36	22199.21
17	68	73	33	143	21520.83	23311.45
18	63	68	37	137	21584.33	22546.05
19	62	67	29	151	21460.07	24632.98
20	60	65	32	132	21541.80	22719.18
21	62	67	33	129	21653.09	22654.22
22	61	66	29	142	21543.44	22711.83
23	60	65	28	139	21444.84	21310.65
24	64	69	35	141	21524.97	23282.65
25	60	65	32	137	21390.10	23376.41
26	60	65	33	140	21462.67	22199.21
27	60	65	31	144	21659.47	22430.24
28	54	59	25	135	21447.02	21758.31
29	58	63	29	126	21466.92	21144.57
30	54	59	28	141	21384.15	22422.89
31	60	65	33	125	21487.01	21895.87
32	64	69	27	133	21494.02	21527.28
33	63	68	37	130	21519.30	22950.50
34	60	65	33	138	21617.55	22300.32
35	61	66	30	130	21479.91	23607.45
36	57	62	30	133	21617.33	23506.33
37	57	62	30	121	21399.98	20119.04
38	59	64	27	128	21511.97	20162.25
39	60	65	31	155	21344.17	26272.55
40	62	67	31	155	21610.43	24199.72
41	57	62	31	128	21594.90	22025.79
42	58	63	33	139	21586.83	24069.80
43	63	68	30	150	21397.81	23354.66
44	51	56	28	127	21689.99	20451.18
45	58	63	34	121	21434.36	22184.80
46	58	63	27	134	21546.84	21982.29
47	62	67	31	135	21548.41	22581.91
48	57	62	29	129	21474.57	20841.23
49	65	70	36	130	21451.11	23824.37
50	63	68	35	144	21609.45	22928.75

Table 19. Statistics Results for the Replication Runs of Policy 3

Statistic	Performance Measures					
	Total Assigned Empty Slots	Total Assigned Overbooked Slots	Total Real Empty Slots	Total Overflowed Slots	Total Expected Cost	Total Actual Cost
<i>Average</i>	60.22	65.22	31.36	137.24	21517.85	22741.54
<i>Std. Dev.</i>	3.40	3.40	3.37	8.78	85.48	1196.84
<i>Number of Replicates (n)</i>	5	4	19	7	0	4
<i>Confidence Interval (h)</i>	0.97	0.97	0.96	2.50	24.31	340.38

Table 20. Replication Runs for Policy 4

Replicate	Performance Measures					
	Total Assigned Empty Slots	Total Assigned Overbooked Slots	Total Real Empty Slots	Total Overflowed Slots	Total Expected Cost	Total Actual Cost
1	61	66	36	134	21436.78	22408.78
2	65	70	28	143	21406.88	23585.70
3	59	64	24	137	21473.35	20862.69
4	64	69	32	132	21362.80	22971.96
5	54	59	27	139	21433.98	23600.10
6	60	65	39	129	21458.59	23932.83
7	58	63	34	138	21488.82	23094.82
8	60	65	38	137	21460.76	22531.64
9	53	58	26	138	21353.98	21288.90
10	56	61	32	120	21447.04	22769.73
11	59	64	32	133	21496.00	21303.59
12	59	64	37	129	21275.27	23405.51
13	53	58	25	128	21465.19	21050.52
14	55	60	31	134	21505.73	20761.87
15	55	60	28	149	21490.53	22877.90
16	52	57	28	137	21443.49	22625.12
17	52	57	26	133	21450.64	21541.68
18	56	61	35	121	21428.15	22827.64
19	63	68	25	142	21458.73	21101.07
20	58	63	28	132	21456.65	21917.33
21	62	67	28	139	21493.41	22069.00
22	59	64	29	131	21467.63	22408.49
23	59	64	28	131	21343.10	21613.99
24	65	70	36	134	21511.96	23723.25
25	55	60	30	131	21359.97	21332.40
26	53	58	31	141	21450.64	22480.80
27	61	66	30	137	21482.06	23101.88
28	61	66	29	137	21554.72	22206.26
29	54	59	36	136	21422.57	23521.03
30	56	61	30	128	21388.20	21332.40
31	64	69	34	132	21391.19	22134.25
32	58	63	37	119	21566.60	21939.37
33	57	62	35	138	21515.44	23029.86
34	61	66	25	153	21435.99	21909.98
35	58	63	30	135	21449.71	22899.65
36	59	64	35	133	21413.81	22675.97
37	53	58	31	114	21374.40	22834.69
38	63	68	31	141	21424.25	21115.77
39	61	66	30	141	21454.99	22950.21
40	57	62	35	135	21415.85	23939.88
41	55	60	39	108	21462.65	20949.98
42	58	63	36	116	21456.14	22762.68
43	57	62	38	120	21512.13	22329.42
44	59	64	26	133	21405.49	21339.45
45	53	58	28	123	21422.30	22321.78
46	57	62	27	151	21478.96	21780.06
47	60	65	34	122	21518.38	21780.35
48	54	59	31	136	21437.70	22329.13
49	59	64	31	135	21369.49	22683.02
50	58	63	28	136	21319.57	21967.88

Table 21. Statistics Results for the Replication Runs of Policy 4

Statistic	Performance Measures					
	Total Assigned Empty Slots	Total Assigned Overbooked Slots	Total Real Empty Slots	Total Overflowed Slots	Total Expected Cost	Total Actual Cost
Average	57.96	62.96	31.18	133.02	21441.85	22318.44
Std. Dev.	3.49	3.49	4.14	8.91	58.00	856.91
Number of Replicates (n)	6	5	29	7	0	2
Confidence Interval (h)	0.99	0.99	1.18	2.53	16.49	243.70