An Integrated Decision-Support Tool to Forecast and Schedule No-Show Appointments in Healthcare

A dissertation presented to
the faculty of
the Russ College of Engineering and Technology of Ohio University

In partial fulfillment
of the requirements for the degree
Doctor of Philosophy

Maria M. Rinder
August 2012
© 2012 Maria M. Rinder. All Rights Reserved.
This dissertation titled
An Integrated Decision-Support Tool to Forecast and Schedule No-Show Appointments
in Healthcare

by

MARIA M. RINDER

has been approved for
the Department of Industrial and Systems Engineering
and the Russ College of Engineering and Technology by

__________________________

Gary Weckman
Associate Professor of Industrial and Systems Engineering

__________________________

Dennis Irwin
Dean, Russ College of Engineering and Technology
ABSTRACT

RINDER, MARIA M., Ph.D., August 2012, Industrial and Systems Engineering

An Integrated Decision-Support Tool to Forecast and Schedule No-Show Appointments in Healthcare (227 pp.)

Director of Dissertation: Gary Weckman

The central purpose of this research is to demonstrate that there is a relationship between patient behavior and no-show rates that can be predicted and that a model is developed to simulate an appointment system that takes in consideration patient behavior. We hypothesize that the no-show rate is associated with the patient’s personal characteristics (i.e., age, gender, race, socioeconomic status, education). The technical question becomes: How can one design a model that can assess the relationship between no-show rates and patient behavior to maximize the quality of services rendered to patients? To answer this question, one should be able to (a) assess or diagnose the current performance level in fact the no-show rate, (b) assess or diagnose how patient behavior influences no-show rates, and (c) devise and implement strategies and interventions to reach the target level of the service quality. The long term goals of this research are to address the above issues with an emphasis on the role of patient behavior in the healthcare quality performance.

The study performed showed that the individual no-show probability can be determined and can be included in the scheduling system. It also proved that the use of
individual no-show probabilities together with different strategies in scheduling can improve considerably patients’ waiting time, doctor idle time and the number of patients seen every day, and indirectly the healthcare organization performance and patient satisfaction.

This research has demonstrated the feasibility of modeling a complex healthcare process with ANN, identifying process changes that greatly improve key performance indicators, and validating the process through discrete time event simulation. Additionally, healthcare professionals were surveyed regarding the usefulness of such findings, further validating the importance of this work.

Approved: _____________________________________________________________

Gary Weckman
Associate Professor of Industrial and Systems Engineering
ACKNOWLEDGMENTS

The doctoral journey is no easy journey. It could not be possible without the support of other people. What follows is my attempt to thank people who were with me in this journey.

To my advisor, Dr. Gary Weckman, thank you for your excellent guidance, patience, and for creating an excellent environment for doing research. You are the best advisor a student can have.

To my committee members, Dr. Schwerha, Dr. Park, Dr. Snow and Dr. Paschold, thank you for your input through all these years.

To my fellow students in the ISE Department, thank you for making our department a friendly place. To Mrs. Seelhorst, thank you for helping me with all the paperwork during years.

To Mr. Marcus at Central State University, thank you for being so understanding and supportive.

To Mrs. Janet Weckman, thank you for opening the doors to your house and making us feel like home.

To my parents, Elena and Constantin, and my sister, Ana, thank you for your unconditional love and support. You are far away, but so close. Talking to you every day gave me the strength to continue.
To my wonderful husband, Peter, I have no words to express the thanks you deserve. From the patience you had when I felt “dark”, to the time you spent listening to my questions and concerns, to the positive thinking, to taking me in the much needed vacations. Thank you for making my life easier. I love you ∞.
<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract ..............................................................................................................</td>
</tr>
<tr>
<td>Acknowledgments ...............................................................................................</td>
</tr>
<tr>
<td>List of Tables ...................................................................................................</td>
</tr>
<tr>
<td>List of Figures .................................................................................................</td>
</tr>
<tr>
<td>Chapter 1: Introduction ..................................................................................</td>
</tr>
<tr>
<td>1.1 Background .................................................................................................</td>
</tr>
<tr>
<td>1.2 Hypothesis .................................................................................................</td>
</tr>
<tr>
<td>1.3 Objectives .................................................................................................</td>
</tr>
<tr>
<td>1.4 Significance ...............................................................................................</td>
</tr>
<tr>
<td>Chapter 2: Literature Review .........................................................................</td>
</tr>
<tr>
<td>2.1 Scheduling Methods in Healthcare Industry ............................................</td>
</tr>
<tr>
<td>2.2 Use of Industrial Engineering in Healthcare ............................................</td>
</tr>
<tr>
<td>2.2.1 Work Scheduling ...................................................................................</td>
</tr>
<tr>
<td>2.2.2 Staff Scheduling ...................................................................................</td>
</tr>
<tr>
<td>2.3 Description of Patient Flow .....................................................................</td>
</tr>
</tbody>
</table>
Chapter 4: Methodology ................................................................................................... 78

4.1 Data Collection ....................................................................................................... 81

4.2 Data Description .................................................................................................... 82

4.3 Data Preprocessing .............................................................................................. 86

4.3.1 Instance Selection ........................................................................................... 87

4.3.2 Missing Data Handling ................................................................................... 88

4.3.3 Feature Selection ............................................................................................ 90

4.3.4 Feature Construction ..................................................................................... 90

4.3.5 Feature Coding ................................................................................................ 96

4.4 Preliminary Analysis ............................................................................................ 99

4.4.1 Analysis of Basic Feature Set ....................................................................... 100

4.4.2 Determination of No-Show Rates ................................................................. 107

4.5 Advanced Analysis ............................................................................................. 113

4.5.1 Application of Neural Network in No-show Analysis .................................. 115

4.6 Knowledge Extraction ......................................................................................... 119

4.6.1 Sensitivity Analysis ....................................................................................... 119

4.5.2 Decision Tree ............................................................................................... 124
4.6. Validation .................................................................................................................. 133

4.6.1 Simulation .............................................................................................................. 134

4.6.2. Survey Results ................................................................................................. 149

Chapter 5: Discussion .................................................................................................... 160

5.1. Benefits and Advantages of the Decision Support Tool ....................................... 161

5.2 Limitations of the Decision Support Tool .............................................................. 163

Chapter 6: Conclusions and Recommendations ........................................................... 167

6.1. Conclusions .......................................................................................................... 167

6.2. Further Improvements for the Decision Support Tool ........................................... 170

References ...................................................................................................................... 172

Appendix 1: Description of Evidence for Group 1 .................................................... 185

Appendix 2: Description of Evidence for group 2 ....................................................... 191

Appendix 3: VBA Module to populate missing data ..................................................... 194

Appendix 4: Rules Derived from Decision Trees for Primary Clinic 2 ....................... 195

Appendix 5: Survey and Introduction Letter ................................................................. 218

Appendix 6: IRB Approval Exemption .......................................................................... 227
LIST OF TABLES

Table 1. Summary of Predictors ................................................................. 50
Table 2. Activation function ..................................................................... 65
Table 3. Examples of Rules Created with TREPAN for Saginaw Bay Dataset (Weckman et al., 2009) ......................................................... 75
Table 4. Example of Database Features .................................................... 83
Table 5. Number of Missing Values for No-Show Appointments Before and After Populating the Fields ......................................................... 89
Table 6. ...................................................................................................... 92
Table 7. Weather Events Coding ............................................................... 93
Table 8. Appointment time transformation ................................................ 95
Table 9. Average Income Data Sample ....................................................... 95
Table 10. Data Coding ................................................................................ 97
Table 11. Code transformation using “Tag as symbol” function ................. 98
Table 12. Descriptive Statistics for Population Age .................................... 100
Table 13. Preliminary Analysis of No-Show ................................................. 109
Table 14. Comparison of No-Show Rates Before and After Preprocessing ...... 113
Table 15. Summary of Models for Primary Clinic 1 .................................. 116
<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>Summary of Models for Primary Clinic 2</td>
<td>117</td>
</tr>
<tr>
<td>17</td>
<td>Summary Models for Primary Clinic 3</td>
<td>118</td>
</tr>
<tr>
<td>18</td>
<td>Summary Models for Specialty Clinic</td>
<td>118</td>
</tr>
<tr>
<td>19</td>
<td>Total Weight Error for Primary Clinic 2</td>
<td>120</td>
</tr>
<tr>
<td>20</td>
<td>Sensitivity Analysis for Primary Clinic 2</td>
<td>121</td>
</tr>
<tr>
<td>21</td>
<td>Total Weight Error for Specialty Clinic</td>
<td>122</td>
</tr>
<tr>
<td>22</td>
<td>Sensitivity Analysis for Specialty Clinic</td>
<td>123</td>
</tr>
<tr>
<td>23</td>
<td>Decision tree specification for CHAID analysis for primary clinic 2</td>
<td>125</td>
</tr>
<tr>
<td>24</td>
<td>Arrival Status Classification Using CHAID Analysis</td>
<td>126</td>
</tr>
<tr>
<td>25</td>
<td>Decision Tree Rules for Primary Clinic 2 for Arrival Status 6 (no-show)</td>
<td>130</td>
</tr>
<tr>
<td>26</td>
<td>Steps for the Simulation Process</td>
<td>136</td>
</tr>
<tr>
<td>27</td>
<td>Results for Actual Status Simulation from 8-5 PM</td>
<td>140</td>
</tr>
<tr>
<td>28</td>
<td>Results for Actual Status Simulation from 8 AM to Last Patient Seen</td>
<td>141</td>
</tr>
<tr>
<td>29</td>
<td>Number of Patients Waiting and Waiting Times</td>
<td>142</td>
</tr>
<tr>
<td>30</td>
<td>Results for Simulation with Different No-Show Appointments Randomly Distributed</td>
<td>143</td>
</tr>
<tr>
<td>31</td>
<td>Results for Simulation with Different No-Show Appointments Grouped</td>
<td>145</td>
</tr>
<tr>
<td>32</td>
<td>Results for Simulation with Different No-show Appointments Grouped</td>
<td>145</td>
</tr>
<tr>
<td>33</td>
<td>Results for Simulation with Different No-show Appointments Grouped and with Deviation of Arrival</td>
<td>146</td>
</tr>
<tr>
<td>Table</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>------</td>
</tr>
<tr>
<td>34</td>
<td>Question 3 Results</td>
<td>150</td>
</tr>
<tr>
<td>35</td>
<td>Question 4 Results</td>
<td>150</td>
</tr>
<tr>
<td>36</td>
<td>Question 5 Answer Summary</td>
<td>152</td>
</tr>
<tr>
<td>37</td>
<td>Question 6 Answer Summary</td>
<td>153</td>
</tr>
<tr>
<td>38</td>
<td>Question 7 Answers’ Summary</td>
<td>153</td>
</tr>
<tr>
<td>39</td>
<td>Distribution of Scheduling Method, Level of Satisfaction and Level of Experience</td>
<td>158</td>
</tr>
<tr>
<td>40</td>
<td>No-show Rate and Level and Experience</td>
<td>159</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

Figure 1. Patient Flow in an Outpatient Facility ............................................................... 27

Figure 2. Flow Diagram of Articles Selected in the Process of Study Identification ...... 32

Figure 3. Flow Chart of the Six Steps Used in Genetic Algorithm (Chien et al, 2008) ... 39

Figure 4. Biological Neuron (Adopted from Optimal Trader, 2011) ............................. 60

Figure 5. Neural Network with Supervised Learning ....................................................... 62

Figure 6. Perceptron (Adopted from Matthews, 2004) .................................................... 64

Figure 7. Multi layer perceptron architecture ................................................................. 66

Figure 8. Procedural Flow for ANN Training (Young et al, 2008) .................................. 68

Figure 9. Neural Network Architecture and Input Variables ........................................ 70

Figure 10. a.) Two Inputs Varied; b.) One Input and Output Varied (Adopted from Weckman et al., 2010). ................................................................. 73

Figure 11. TREPAN Decision Tree for Saginaw Bay Dataset (Weckman et al., 2009). . 75

Figure 12. Flow Chart of the Research ........................................................................... 78

Figure 13. Detailed Flow-Chart of the Research ............................................................... 80

Figure 14. Flow Chart of Data Preprocessing ................................................................. 88

Figure 15. CDX technologies add-in for Microsoft Excel ................................................. 94

Figure 16. Preliminary Analysis Flow Chart ................................................................. 99

Figure 17. Mean, Median, Mode, and Standard Deviation of Population Age by Clinic. ....................................................................................................................... 101
Figure 18. Gender Distribution for Each Clinic, Overall Clinics and United States..... 102
Figure 19. Race Distribution for Each Clinic, Overall Clinics and United States........ 103
Figure 20. Insurance Type Distribution for Each Clinic, Overall Clinics and United States. .............................................................................................................................. 104
Figure 21. Provider Type Distribution for Each Clinic and Overall.............................. 105
Figure 22. Appointment Type Distribution for Each Clinic and Overall. ................. 106
Figure 23. Employment Type Distribution for Each Clinic and Overall..................... 107
Figure 24. No-show Rates by Appointment Type. ..................................................... 110
Figure 25. No-show Rates for Primary Care Clinic 1................................................. 111
Figure 26. No-show Rates for Primary Care clinic 2................................................. 112
Figure 27. Arrival Status for Primary Care Clinic 3................................................. 112
Figure 28. Arrival Status for Specialty Clinic. ............................................................ 113
Figure 29. Flow Chart of Advanced Analysis. ............................................................ 114
Figure 30. Sensitivity Analysis for Primary Clinic 2.................................................. 121
Figure 31. Sensitivity Analysis for Specialty Clinic.................................................. 124
Figure 32. Decision Tree for Primary Clinic 2......................................................... 128
Figure 33. Decision Tree for Arrival Status 6............................................................ 129
Figure 34. Flow Chart of The Simulation Process..................................................... 135
Figure 35. Flow Chart for Simulation Process......................................................... 136
Figure 36. Summary of Doctors’ Utilization Time for the Four Simulations. .......... 147
Figure 37. Summary of Average Patient Waiting Time to See a Doctor ....................... 147
Figure 38. Summary of Total Maximum and Average Patient Waiting Time. .............. 148
Figure 39. Total Number of Patients Seen in One Day. ................................................. 148
Figure 40. Level of Experience of Respondents. ............................................................ 154
Figure 41. Current Position for Respondents................................................................. 155
Figure 42. Number of Appointments per Month for Respondents’ Organization........ 156
Figure 43. Distribution of No-show Rates Among Respondents. ................................. 157
CHAPTER 1: INTRODUCTION

1.1 Background

The healthcare industry represents 17.4% of the gross domestic product of the United States. According to the Bureau of Labor Statistics, it was the largest industry in 2006 (Bureau of Labor Statistics). It generated expenditures of 2.5 trillion USD in 2009 (OECD, 2011). These costs are growing at such a rate that the amount of public money needed to finance healthcare, which currently stands at 45% of total healthcare expenditures, is expected to double by 2050 (Gupta and Denton, 2009). Efficient use of resources is critical in an era of rapidly escalating costs (Sweeney, 1996), so healthcare managers face increased pressure to reduce costs while still delivering high-quality care (Wright et al., 2006). A healthcare provider operates in a regulated industry where the quality of performance is predominantly evaluated by the services rendered to the patient and the effectiveness of the process. An important measure of process effectiveness is the number of patients who visit the practice every day.

Missed patient appointments (no-show) have an adverse effect on resource utilization in healthcare services. Missed appointments lead to loss of revenue, inefficient scheduling, and underutilization of personnel (Macharia et al., 1992). The no-shows are recognized as a drag on operating efficiency and optimal patient care. No-shows reduce provider productivity and efficiency, increase health care costs, and limit the ability of a
clinic to serve its client population by reducing its effective capacity (LaGanga and Lawrence, 2007). Missed appointments can influence patient safety and health, as well. The scheduling system can affect both direct waiting time and indirect waiting time. Whereas direct waiting is an inconvenience to the patient, excessive indirect waiting can pose a serious safety concern (Murray and Berwick, 2003). If patients cannot get timely service, their condition can worsen or become life threatening. In addition, by not showing for an appointment, a patient denies service to others who may have shown up. Long direct and indirect waiting times are a source for patient dissatisfaction.

Patient satisfaction is an important measure of quality of healthcare, and healthcare providers give it great importance. These are largely influenced by the performance of the scheduling system embedded in the healthcare process and by patient behavior. In this research, a model is developed to assess the relationship between patient characteristics and the no-show rate and assess simulated strategies (interventions) to determine their impact on the scheduling system. The model will be based on the latest in data mining and knowledge extraction techniques, both traditional (such as regression) and nontraditional (such as artificial neural networks and decision trees).

Patient behavior can be influenced by a number of factors: (a) relationship with the human operator (e.g., physician, registered nurse, technician, nursing aid), (b) technology (e.g., EKG machine, computer), (c) organization (e.g., staffing, scheduling of appointments), (d) physical environment (e.g., waiting room lighting, ventilation), and (f)
personal characteristics (e.g., demographics, education, socioeconomic status). The work result may be measured by quality of service. In healthcare, quality of service is measured by the success of the outcome, and patient satisfaction.

1.2. Hypothesis

In this research, it is hypothesized that the no-show rate is associated with the personal characteristics (i.e., age, gender, race, socioeconomic status, education) of the patient (i.e., work object). The hypotheses are as follows:

**Hypothesis 1:** *No show rates are statistically different from those groups identified by personal characteristics.*

The analysis will be performed between groups belonging to the same variables (i.e., different race, different age, and different insurance type).

**Hypothesis 2:** *No show rates are significantly correlated to personal characteristics of patients.*

The second hypothesis will further verify the correlation between patient characteristics and no-show rates. The results will be used to simulate various scenarios for patient scheduling to improve organization performance.

**Hypothesis 3:** *No show rates are significantly correlated to appointment characteristics.*
The third hypothesis will verify the correlation between appointment characteristics (i.e., day and time of appointment) and no-show rates. The relationship found between these two will have a greater importance because it is much easier to change appointment characteristics, while patient characteristics cannot be changed.

1.3. Objectives

The technical question becomes “How can one design a model that can assess the relationship between no-show rates and patient characteristics in order to maximize the quality of services rendered to patients?” To answer this question, one should be able to (a) assess or diagnose the current performance level of no-show rates, (b) assess or diagnose how patient characteristics influence no-show rates, and (c) devise and implement simulated strategies and interventions to reach the fullest possible scheduling. The long-term goals of this research are to address the above issues with an emphasis on the role of patient characteristics in the quality performance.

1.4. Significance

The significance of this research is in the method of forecasting patients who may not show up for appointments. While studies performed in the past accounted only for a single patient characteristic to forecast a no-show, this research accounts for multiple patient characteristics simultaneously and adds new characteristics to the already existing
ones. Some studies took into account the no-show rate when proposing a new scheduling algorithm, but failed to precisely forecast who may miss appointments.

A formal methodology that can forecast the relationship between patient characteristics and no-show could provide healthcare organizations the ability to make better decisions when scheduling patients and resources. These decisions can ultimately improve system performance, resource efficiency, revenue enhancement, and, patients’ satisfaction.

The broader impacts of the research offer the promise of improved competitiveness within the healthcare sector, through the reduction of incidences and costs of missed appointments. This approach has the potential to become a “standalone” business strategy capable of improving organizational performance. Although the proposed technology targets the healthcare sector, this project may be applicable to other service sectors such as transportation, retail, and distribution.
CHAPTER 2: LITERATURE REVIEW

2.1. Scheduling Methods in Healthcare Industry

Patients who schedule appointments face two types of waiting time (Gupta and Denton, 2008). Indirect waiting time is the period between the moment patients requested the appointment and the actual time of appointment. The direct waiting time is the difference between the appointment time and the time the patient is served by the provider. The scheduling system can affect both direct waiting time and indirect waiting time. Whereas direct waiting is an inconvenience to the patient, excessive indirect waiting can pose a serious safety concern (Murray and Berwick, 2003). If patients cannot get the healthcare service at the time needed, their condition can worsen, developing a threatening condition to their life. Long direct and indirect waiting times are a source for patient dissatisfaction.

In developing patient schedules, there are a number of external factors that cannot be controlled, such as no-show appointments, cancellations, emergency appointments, and walk-ins. These factors are directly determined by patients’ needs. It is a challenge for the practice to manage no-show or walk-ins appointment, since they can interfere with the success of the daily operation. A smooth patient flow often determines the success of the day. It can also determine the profitable use of provider and staff time. The scheduling process is the first contact between patient and provider, being the “basis of
the patient’s first impression of facility and provider” (Wojtys et al., 2009). It is a factor in determining if the patient will visit the practice again or not.

2.2 Use of Industrial Engineering in Healthcare

The first use of industrial engineering in healthcare happened in 1913 when Frank Gilbreth used industrial engineering tools, specifically motion study techniques, for methods’ improvement in a hospital. In 1940s, Lillian Gilbreth published articles that showed the benefits of using methods’ improvement techniques in hospitals and nursing. In the 1950s, American Hospital Association (AHA) started to offer workshops on methods’ improvement, and, in 1952, created the Committee on Methods Improvement. At the same time, universities across US started to offer courses on application of methods’ improvement in hospital administration. In 1961, the Hospital Management Systems Society (HMSS) was founded. These events marked the beginning of the use of industrial engineering methods in healthcare management. As this field grew, the Institute of Industrial Engineering (IIE) recognized the role of industrial engineering techniques and formed the hospital section in 1964 that later became the Health Services Division. In 1987, due to the growing use of information technology in healthcare, HMSS changed its name to Healthcare Information and Management Systems Society (HIMSS). In 1988, IIE approved the formation of the Society for Health Systems (SHS) that replaced Health Services Division (Kachhal, 2001).
In the last decades, many universities created industrial engineering programs with specialization in healthcare systems (Kachhal, 2001). Many industrial engineers work in healthcare system as employees or consultants and are typically known as management engineers or operations analysts (Kachhal, 2001). Almost all industrial engineering tools and techniques are used currently in the healthcare system: methods improvement and work simplification, staffing analysis, scheduling, queuing and simulation, optimization, decision support systems, quality improvement, and statistical analysis.

This review will discuss the use of industrial engineering techniques in improving scheduling in healthcare. There are two facets in the process: work (patient) scheduling and resource scheduling. In health systems, the patient is the work. The work/patient can be scheduled for those departments where there are a given number of slots available (e.g. outpatient clinic). For other departments, the work scheduling cannot be possible (e.g. emergency room). Work scheduling has to be accompanied by resource scheduling that includes staff scheduling, instrumentation scheduling, room scheduling. This is accomplished based on the conditions from each department, such as the number of slots available, time required for each slot, equitability between personnel (number of hours to be worked, days in a week, and holidays).
2.2.1 Work Scheduling.

Patient appointments are scheduled based on the number of appointment slots that are available in the practice. The number of appointments available is established based on the type of work needed to be completed, such as a regular visit, follow-up visit, tests and procedures, education sections, and the number of providers available. It is a current practice that new appointments are made over the phone. The follow-up appointments are made at the end of the actual visit or by phone. Appointments are made until all available slots are filled in. Scheduling can be performed using a manual system (e.g. appointment book) or a computerized system. Both methods have to search for an available spot based on patient requirements. The manual system requires that the scheduler has expert knowledge and experience on the job. The computerized system can do the same thing in a much shorter period of time.

2.2.2 Staff Scheduling.

Staff scheduling can be performed only after the staff requirements have been determined. Each staff member is assigned a pattern of working days and off days (Kachhal, 2001). There are two types of personnel scheduling: cyclical and non-cyclical. Cyclical scheduling assigns the same pattern after a certain number of days or weeks. The advantage of this type is that personnel know the schedule. The disadvantage is the lack of flexibility in accommodating demand and worker needs. The non-cyclical type generates new schedules for a short period of time (e.g. two weeks). It is based on
demand and available staff. Its advantage is that it can accommodate change in demand and staff needs. Its disadvantage is that it requires more planning time than cyclical schedules. Developing of both personnel schedules has involved heuristics, trial-and-error, or optimization techniques (Kachhal, 2001).

2.3 Description of Patient Flow

A flow map is shown in Figure 1. The process starts with the patient calling the healthcare provider to schedule an appointment with the physician. The day of the appointment the patient shows up at the provider. Upon arrival at the clinic, the patient is greeted by the person at the Patient Service Assistant desk (PSA). The PSA personnel check the appointment in the database, update patient personal information such as address, phone number, and check medical insurance information. The patient is directed to the waiting room.

Once the doctor finishes seeing the previous patient, the nurse from the nurse station pulls the next file from the record-queue and calls the corresponding patient. The doctor examines the patient. Based on the results of the examination, two options are available: (1) the patient receives treatment or (2) is directed to another physician/test. In the first case, the patient goes to the nurse station, then to the check-out office. If necessary, a new appointment is scheduled. After that, the patient leaves the practice. In the second case, the patient goes to the nurse station and then to the check-out station.
Here a new appointment is scheduled with the new physician, and the patient leaves the practice. Once the patient leaves the check-out desk, his/her file is sent to the billing office where a bill is issued, and then to the medical records room where all his/her information is updated.

Figure 1. Patient Flow in an Outpatient Facility
2.4. Earlier Scheduling Methods

In this section we review the earlier methods used in patient scheduling. The earlier articles (Lindley, 1952, and Bailey, 1952) are based on mathematical programming methods and queuing theory. These studies consider that the conditions in which work is performed are static, which may be far from the real work environment. Lindley (1952) uses the case of a single server, and random patients’ arrivals. He concluded that scheduling at regular intervals improves system performance. Following Lindley’s model, more mathematical models were developed where more dynamic conditions were added such that the proposed environment is as close as possible to the real environment. Some of the mathematical models are the ones by Jansson (1966) which considers an individual block and constant appointment time, Soriano (1966) which considers multiple-block/fixed intervals; and Mercer (1960) which considers individual-block and late arrivals. He also considers that the probability of patient no-show is greater than zero. Later mathematical models use dynamic programming, variable-block/fixed interval (Fries and Marathe, 1981), multi-server queuing models with nonhomogeneous arrivals (Brahimi and Worthington, 1991), queuing system with multiple doctors and random arrival time (Liu and Liu, 1998), and two-stage stochastic linear programming to determine optimal appointment intervals (Denton and Gupta, 2003).
Bailey (1952) used a simplified queuing technique. It was based on fairly static conditions such as appointments at regular intervals that were equal to the average consultation time, patients arriving punctually at their appointments, patients seen in the order they arrive, and the consultant seeing only one patient at the time. Some variables such as the number of patients attending the clinic, length of appointment interval, and number of patients waiting at any time, were varied for comparison reasons. The goal of the model was to reduce patient waiting and consultant idle time. Bailey’s model used a Monte Carlo simulation technique and is considered one of the first simulation models. It was further developed to include scheduling of groups of patients based on the appointment length, and resting time for consultants (Welch and Bailey, 1952). Welch (1954) concluded that based on the mathematical model used, the factors that may influence appointment scheduling are punctuality of medical staff and patients, and appointment interval.

Other simulation models were developed a few years after Bailey’s model. Fetter and Thompson (1966) varied some of the variables that were considered fixed in the previous models, such as patient’s punctuality, number of appointments, no-show rates, walk-ins rates, appointment intervals, and patient loads. Vissers and Wijngaard (1979) modelled the system based on five variables: mean consultation time, consultation time, patient’s punctuality, number of appointments, and “system earliness”. Ho and Lau (1992, 1995, and 1999) developed some of the most comprehensive models by including
50 appointment rules. They concluded that no-show, service time, and number of patients affect the system performance. Klassen and Rohleder (1996 and 2000) use simulation with different variance patients and fixed appointment intervals. They concluded that “low-variance” patients assigned at the beginning of the appointment session may perform better. They also included two slots for emergency calls, and scheduler error when classifying patients.

These techniques improved the way scheduling is completed; however, they still have limitations. It is expected that the new scheduling techniques should adapt to the highly dynamic environment that characterizes the present healthcare system. The presented scheduling techniques lack generality. The studies analyzed a specific clinic; they did account for a specific environment that may not apply to other clinics. A second limitation would be the use of very simple models, such single server, single-phase, and non randomization of arrival patterns. A third limitation is that none of these models accounts for the effects of walk-ins, no-show, or emergencies which are very common in real life.

2.5. Advanced Scheduling Techniques

This section focuses on analysis of more complex methods used in scheduling systems in healthcare that can improve scheduling methods through the inclusion of variability and complexity.
To retrieve relevant studies from peer-reviewed literature, two electronic databases were searched: Medline, and Pub Med. The search included articles published up to October 2009 using the following keywords: MODEL AND (DATA) AND (HEALTHCARE OR MEDICAL) and (SCHEDULING). In addition to the electronic search, bibliographies in relevant papers were reviewed to identify additional studies.

The studies that were included in the review met the following requirements: (a) the paper described application of data modeling technique to improve scheduling in healthcare. The modeling technique has to be a data mining technique; (b) the modeling technique was applied to a real database; and (c) the paper was a full report published in English in a peer-reviewed journal. After article titles were obtained from the databases mentioned above, the abstracts were reviewed for possible inclusion. For the abstracts that were selected as possible inclusion, or abstracts that provided insufficient information, full papers were retrieved. The same steps were applied to the articles that were selected from bibliographies. After reading the full article, the articles that met all the criteria mentioned above were selected. 2174 titles were obtained following the search. 2067 citations were not related to the search and were excluded. All 107 abstracts left were read and sixty three were excluded because they did not meet the selection criteria. Thirty five full papers were reviewed and twenty six papers were excluded. Nine articles were finally included in the review. The nine articles were grouped into two groups: first group consisted of six studies that describe techniques used in inpatient
scheduling (Podgorelec and Kokol, 1997; Isken and Rajagopalan, 2002; Kaandorp and Koole, 2007; Chien et al., 2008; Ogulata et al., 2008; and Chakraborty et al., 2010). The second group consisted of three studies that describe techniques used in nurse scheduling (Beaulieu et al., 2000; Aickelin and Dowsland, 2004; and DeGrano et al., 2009). Figure 2 shows the flow diagram of articles selected in the process of study identification.

Figure 2. Flow Diagram of Articles Selected in the Process of Study Identification
Description of evidence was summarized for all ten studies in terms of intervention, outcome, study design, and main results. The description of evidence for the five studies that reported techniques uses inpatient scheduling (Group 1) and is listed in Appendix 1. These studies have as outcomes measured patient waiting time, idle time for staff or devices, and overall duration. More information is presented as follows:


2. Two of the studies (Ogulata et al., 2008; and Chakraborty et al., 2010) introduced mathematical models to schedule patients and personnel in hospital services.

3. Kaandorp and Koole (2007) used a local search procedure to find the optimal schedule with a weighted average of expected waiting times for patients, idle time for doctors, and tardiness as objectives.

4. A data mining approach to support simulation modeling of patient flow was introduced by Isken and Rajagopalan (2002).

Appendix 2 summarizes the description of evidence for the three that reported techniques used to improve nurse scheduling (Group 2). These three studies had measured the reduction in time necessary for the scheduler to develop the schedule.
1. The three studies introduced three different approaches in staff scheduling: one introduces an auction-optimization method, one a mathematical model, and one an indirect genetic algorithm.

2. The general outcomes of these studies were reduced time allocated to staff scheduling.

3. All studies include different constraints in model development such as number of hours worked in a week, and number of nights worked in a week.

4. All studies were performed in the USA, Canada, and UK.

2.6. Description of Models


Mathematical modelling is based on the optimization concepts of linear programming. The objective of the mathematical modelling is minimization or maximization of the objective function. It has to take into account a number of constraints that are defined by the problem type. In general, optimization techniques include linear programming, integer programming (a linear programming technique that requires the variables to be integers), mixed integer programming, dynamic programming, constructive algorithms and much more. In this review, four studies that fall in this category were included: two studies that deal with physician and nurse scheduling (Beaulieu et al, 2000; and DeGrano et al., 2009), and two studies that deal with patient/staff scheduling (Ogulata et al. 2008; and Chakraborty et al., 2010). As the
per optimization methodology, all studies defined the variables that will be the input in the model, and the objective function. In the study by Beaulieu et al. (2000), the authors defined the model as a multi-objective integer model meaning that it has multiple objectives and all the solutions have to be integers. The objective function sought to minimize a weighted sum of all deviations. Ogulata et al. (2008) used three hierarchical mathematical models having the objective of maximization of number of patients seen in one week, obtaining a balanced distribution of patients among physicians, and minimization of patient waiting time. The mathematical models were applied in three different stages: weekly patient selection, assignment of physiotherapists, and patient scheduling. The study by Chakraborty et al. (2010) introduced sequential clinical scheduling which can be formulated using dynamic programming. The study is based on earlier work (Muthuraman and Lawley, 2008) that has as the objective function maximization of the expected revenue for patients seen minus the cost for patient waiting and staff revenue, and includes the no-show appointments in the scheduling model. The research in this study is further used in subsequent research (Zeng et al., 2010; and Turkcan et al., 2011) that uses the same objective function and constraints, but changes the distribution of the patients’ no-show. This study (Chakraborty et al., 2010) considers that the no-show patients’ distribution is homogeneous. DeGrano et al. (2009) uses a different approach in his optimization model. His first stage is an auction where nurses can bid the shifts they want to work, and the second stage awards the shifts and performs
the scheduling. The model objective for the award model is to maximize the point value of bids awarded to the candidate winners. The objective for the assignment model is to seek award bids that were not selected as candidate winners but can create a feasible assignment.

2.6.2. Genetic Algorithm Used Inpatient and Nurse/Physician Scheduling.

Genetic algorithms (GAs) are heuristic search methods that are used in solving complex search and optimization problems. GAs are based on the mechanism of natural selection and natural genetics. Very often they are able to find the optimal solution in most complex search spaces and “offer significant benefits over other search and optimization techniques” (Podgorelec and Kokol., 1997). There are three studies that use GA for scheduling – patients or nurse: Podgorelec and Kokol (1997); Chien et al. (2008); and Aickelin and Dowsland (2004). GAs typically use six steps: 1.) selection of initial population; 2.) reproduction; 3.) crossover; 4.) mutation; 5.) evaluation; and 6.) selection of solution. Figure 3 shows the flow chart for the six steps as followed in Chien et al. (2008). All of the authors used the same GA algorithm described below:

- The initial population can be randomly selected or through seeded population. Podgorelec and Kokol (1997) used the methods of seeded population by filling the table with therapies randomly chosen. Chien et al. (2008) used a local search heuristic to derive the sequence of patients and therapies that will represent the

- The reproduction step is explained only in two articles. Podgorelec and Kokol (1997) used negative points for each individual selected from the initial population. The fewer the negative points an individual has, the more chances to be selected for crossover. In Chien et al (2008), chromosomes were randomly selected.

- The crossover step was described in Podgorelec and Kokol (1997) and Chien et al. (2008). The first study used the method of cutting chromosomes into different parts and then randomly putting them together. The second study used a modified version of preserving order-based crossover. After individuals were selected from the initial population, all precedence-dependent therapies were identified. These therapies were copied into the offspring chromosome at the same position. The remaining therapies were copied from parent 2 into the offspring following the same order.

- The mutation step was described in Podgorelec and Kokol (1997) as mutation between two random activities. In Chien et al. (2008) mutation procedure was performed by randomly selecting a therapy from a parent chromosome. The leftmost and rightmost positions were searched to find an interval within which the selected therapy can mutate without violating the precedence constraint. The therapy was then inserted in the selected position.
- The evaluation method was different in each study, but all use a fitness score to evaluate the chromosomes. Podgorelec and Kokol (1997) and Chien et al. (2008) used a combination of maximum waiting time and the makespan as the evaluation function. Aickelin and Dowsland (2004) used the total preference cost of all nurses as the evaluation function.

- The selection of solution was made based on the evaluation function. In the two studies involving patient scheduling, the solution with minimum waiting time and lowest makespan was chosen. In the study involving nurse scheduling, the solution with the minimum total preference cost was chosen.

2.6.3. Local Search and Data Mining Models.

Two articles included in this research deal with the use of local search and data mining techniques for patient and nurse scheduling: Kaandorp and Koole (2007) and Isken and Rajagopalan (2002). The local search method used by Kaandorp and Koole (2007) started with a feasible solution and searched for a better solution in its neighbourhood until a local minimum was found. The authors stated that the solution found was not a global minimum, but with a well-chosen neighbourhood it was possible to find the global minimum. To find a feasible solution, the authors used a mathematical model type of scheduling where they calculated patient mean waiting time, physician idle time, and tardiness. The objective was to minimize the sum of the three variables. The number of solutions given by this model was huge, so a search algorithm was needed.
Figure 3. Flow Chart of the Six Steps Used in Genetic Algorithm (Chien et al, 2008)
Isken and Rajagopalan (2002) used a data mining technique, specifically clustering, such as K-means, “to help the development of patient type definitions for purpose of building (...) simulation or analytical models of patient flow in hospital”. The authors started with the idea that patients have different needs such as different types of treatments or sequence of treatments, so consequently, different resources need to be allocated to different patients. This would be too big of a problem to be solved, so the solution would be patient classification in groups that have the same needs. To classify patients in groups, Isken and Rajagopalan (2002) used the number of total hours spent in different categories of hospital units, and the path and associated lengths of stay as input variables. The authors also used Diagnostic Related Groups (DRG) and Clinical Classification Software (CCS) that gave them information about the diagnoses and procedures. The K-means clustering method was used to classify patients. K-means is an applied clustering algorithm in unsupervised classification problems to find k optimal clusters in a data set. The algorithm steps are as follows:

- Initialization - Choose the number of clusters, k.
- Find new clusters - Randomly generate k clusters and determine the cluster centers, or directly generate k random points as cluster centers.
- Assign each point to the nearest cluster center.
- Recompute the new cluster centers.
- Repeat the two previous steps until some convergence criterion is met.
2.7. Discussion

Healthcare is consuming an increasing percentage of our economic gross domestic product (Hall, 2006). “The rising cost can be attributed in part to aging populations and expense of new, advanced treatment. Just as importantly, it can be attributed to the inefficiencies in healthcare delivery” (Hall, 2006). Nolan et al. (1996) noted that reduced delays and increased access could also reduce costs. A key point in improving healthcare delivery is the improvement of process performance. Part of process performance is balancing the demand and resources. In the healthcare system, the demand is the number of patients to be seen every day and the resources are the nurses, physicians, rooms, and instrumentation that are available. Balancing demand and resources will direct us towards the problem of scheduling patients and resources. In this review, we focused exclusively on methods for patient and staff scheduling.

In the last few decades, scheduling methods have been developed from manual scheduling to computerized approaches. A scheduler takes a substantial amount of time and resources to do manual scheduling (Chien et al., 2008). All the methods developed in the last decades try to improve not only the quality of service, but to help schedulers do their work more efficiently (Chien et al., 2008). The attention of the new methods is not aimed only at staff scheduling, but equally at patient scheduling. These methods aim at
the reduction of patient waiting time (indirect and direct waiting time), and total time spent in hospital.

As Podgorelec and Kokol (1997) mentioned, regardless of the methods used, there are basic rules that must be followed “in order to construct a qualitative and effective automated scheduling system”, such as feasibility of all obtained solutions, and fulfillment of all constraints. Another important aspect is to find an adequate solution in a reasonable amount of time. One indispensable property of the scheduling technique has to be the capability of solving general and independent problems, meaning the capability of the technique to be used in all kinds of situations, especially in an unpredicted situation. For example, it will allow the system to search for solutions even when activities already scheduled are cancelled.

The articles included in this review present different methods for patient and nurse/physician scheduling, mathematical modeling/optimization, genetic algorithm, local search, and data mining. All these methods have common objectives:

- To decrease patients waiting time (Podgorelec and Kokol, 1997; Kaandorp and Koole, 2007; Chien et al., 2008; and Ogulata et al., 2008)
- To reduce physician idle times (Kaandorp and Koole, 2007)
- To increase equipment utilization (Podgorelec and Kokol, 1997; Kaandorp and Koole, 2007; Chien et al., 2008; and Ogulata et al., 2008)
- To select maximum number of patients (Ogulata et al., 2008)
• To fairly distribute patients among physicians (Ogulata et al., 2008), and
• To reduce time and effort to construct schedule (Beaulieu et al., 2000; Aickelin and Dowsland, 2004).

Achieving these objectives will result, indirectly, in improvement of healthcare quality and cost reduction, parameters that are essential to the healthcare system. These models take into account a number of variables:

• number of open slots per day
• number of days in the schedule
• length of appointment
• number of patients to be scheduled every day
• number of patients assigned to each staff member
• overall duration of activities, process starting time and completion time
• earliest starting time and latest completion time
• number of shifts available during the schedule period
• cost of using of resources.

In addition to the variables used in these models, variables that are related to patient behavior can be used as inputs in a model. All the models included in this research considered that patients and staffs were always on time. This situation is not true in real life. Many times physicians are not on time at their clinics (Welch and Bailey, 1952) or patients are late for their appointments. A high percent represents the missed
appointment, when patients do not show up and do not cancel in advance. Recent reviews in the scientific literature concluded that the no-show rates may be around 20% (Bennett and Baxley, 2009), vary in between 15-30% in general adult and pediatric clinics (Deyo and Inui, 1980), and 2-15% in private practices (Barron, 1980). Missed patient appointments (no-show) have an adverse effect on resource utilization in healthcare services. Missed appointments lead to loss in revenue, inefficient scheduling, and underutilization of personnel (Macharia et al., 1992).

Based on information collected from the Organization for Economic Co-operation and development (OECD, 2011), an estimation of the no-show cost in healthcare in the U.S. was calculated. There are 2.43 practicing physicians for 1000 people (head count) in the U.S. According to the U.S. Census Bureau (U.S. Census Bureau, 2011), there are about 310.5 million people in U.S. That means that there are about 754,000 practicing physicians in U.S. Each physician has scheduled, on average, about 1550 appointments each year, resulting in a total of appointments of 1,168.7 million. If we consider that average no-show rate is 20% (Bennett, 1980), that means there are about 233.74 million no-show appointments each year. At a cost of $150/appointment, no-show appointment cost rises to about $35 billion/year. This amount represents lost income for the healthcare provider, where there are costs associated with it, such as physician and staff salaries and other resources.
The no-shows are recognized as a drag on operating efficiency and optimal patient care. No-shows reduce provider productivity and efficiency, increase health care costs, and limit the ability of a clinic to serve its client population by reducing its effective capacity (LaGanga and Lawrence, 2007). These measures are largely influenced by the performance of the scheduling system embedded in the healthcare process and by the patient behavior. While previous efforts have identified some relevant elements of the systems, they fail to provide a holistic, quantitative approach combining the organization scheduling system and patient behaviour into a common framework.

All the models included in this research have accounted for a series of constraints. Beaulieu et al. (2000) grouped the constraints into four distinct categories, but these types of constraints are found in all other studies related to nurse scheduling. The four groups are as follows:

- Compulsory constraints: one physician/nurse must be assigned to each shift of the period; a physician/nurse cannot work more than a shift per day; and a physician/nurse assigned to a night shift cannot be assigned to a day shift of the day after.

- Ergonomic constraints: upper limits on the number of weekly hours of certain types of shifts; limited number of successive working days; after a working weekend, Monday should be off; if a physician/nurse works for three consecutive night shifts, three consecutive days off must be assigned.
- Distribution constraints: seniority is taken into account.
- Goal constraints: a physician/nurse should work a specified number of hours per week; certain types of shifts must be fairly distributed among physicians/nurses.

Similar with the constraints found in the nurse/physician scheduling studies, the patient scheduling studies had their own constraints as follows:

- No patient can perform more than one activity at a time
- Every resource (personnel or device/instrument) can be used by only one patient at a time
- Each activity has to be performed in only one continuous time interval
- Every activity can be performed only with a specific resource
- Some activities have to be performed in an exact time order
- Makespan must not exceed the service time of the healthcare facility.

All these constraints help us find the most feasible solution, but do not guarantee that it is the best solution to use. More criteria can be included in the evaluation function helping in selection of the fittest solution. These constraints may be related to patient characteristics or behaviour.

The most important advantage of using these new methods in scheduling in healthcare is that it takes much less time than traditional methods. Also, by adjusting the importance of each input that may affect the quality of the final solution makes these methods more likely to be included in scheduling tools. They can solve much more
complex problems than the traditional methods by incorporating more input variables and more constraints. These methods can account for all constraints simultaneously compared with traditional methods (manual method) that can account for constraints only one at a time.

There are some limitations in using these new methods. All the artificial intelligence (AI) techniques develop rules based on the input and constraints. Since these inputs and constraints are specific to each healthcare facility, one may ask the question, “Can the AI models be applied to a wide range of healthcare facilities? Are these models valid?” The challenge lies in developing procedures that can resolve the conflict that may arise from using different sets of rules developed by multiple experts (Sitompul and Randhawa, 1990). Other considerations that have to be taken into account are related to patients such as indirect waiting time, late cancellation and no-show, emergency walk-ins and patient preferences (Gupta and Denton, 2008). Indirect waiting time is the difference between the time that a patient requests an appointment and the time of that appointment. Direct waiting time is the time between the appointment time and the actual time when a patient is seen by a nurse/physician. The scheduling system can affect both direct waiting time and indirect waiting time. Whereas direct waiting is an inconvenience to the patient, an excessive indirect wait can pose a serious safety concern (Murray and Berwick, 2003). Some patients cannot get the healthcare service at the time they need it, and their condition can worsen, developing into a life threatening condition. Both no-show and
walk-ins influence the scheduling system. No-shows create a gap in the system that results in increased costs. Also, by not showing for an appointment, a patient denies an appointment to another patient who may show up, and not directly increases the indirect waiting time. Walk-ins occur randomly and may increase waiting time for patients and overload for nurse/physicians. Nevertheless, managing these two aspects, no-shows and walk-ins, may be a challenging task. At last, patient preferences are not taken into account when scheduling. Preferences are very different from one person to another, but one model can create groups of patients (clusters) with similar preferences and then use these groups in developing a schedule.

2.8. No-Show Appointments in Healthcare Industry

Recent reviews in the scientific literature concluded that the following:

(a) the no-show rates may be approximately 20% (Bennett et al., 2009) and vary between 15-30% in general adult and pediatric clinics (Deyo and Inui, 1980), and 2-15% in private practices (Barron, 1980)

(b) patient demographics (i.e., age, gender) are not an accurate predictor for no-show appointments

(c) there are other factors, such as transportation, insurance type, that may predict the no-show appointment and
(d) the conventional intervention methods, such as a reminder phone call/card may reduce the no-show rates.

In the reviewed literature, the predictors of missed appointments were, in general, related to the demographic characteristics of the patient, such as gender, race, age, and socio-economic status. Very few studies took into account other variables such as symptom duration, severity of illness, or long waiting periods. Other variables should be considered in predicting the failed appointments, such as marital status, employment status, insurance type, and distance to appointment facility. Table 1 is a summary of patient characteristics that may be considered as predictors of missed appointments.
## Table 1. Summary of Predictors

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Bean and Talaga, 1992</th>
<th>Garuda et al., 1998</th>
<th>Our Review</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Significant</td>
<td>Not Significant</td>
<td>Significant</td>
</tr>
<tr>
<td>Age</td>
<td>12</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td>Gender</td>
<td>3</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Marital Status</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Race</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Previous Appointment Keeping</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Health Beliefs and Attitudes</td>
<td>3</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Drug and Alcohol Abuse</td>
<td>3</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Seriousness of Problem</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Referral Source</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Day of Appointment</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Time of Appointment</td>
<td>3</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>Weather</td>
<td>1</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Waiting Time</td>
<td>6</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>Insurance Type</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Distance</td>
<td>1</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>Parking</td>
<td>2</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Transportation</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Socioeconomic Status/Education</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>Number of Visits</td>
<td>-</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>Change of Physician</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>

Notes: The numbers represent the number of articles that include a specific predictor
<table>
<thead>
<tr>
<th>Predictor</th>
<th>Bean and Talaga, 1992</th>
<th>Garuda et al., 1998</th>
<th>Our Review</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Significant</td>
<td>Not Significant</td>
<td>Significant</td>
</tr>
<tr>
<td>First Appointment</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Cell Phone Availability</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Department</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Parental Age</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Number of Siblings</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Residency Status</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Notes: The numbers represent the number of articles that include a specific predictor.
Bean and Talaga (1992) classified the predictors into three categories: (a) demographic characteristics such as age, gender, marital status, race and social class; (b) patient behavior and attitudes such as previous no-show history, health benefits and attitudes, and psychological problems, alcohol and drug abuse; and (c) situational characteristics such as day and time of appointments, referral source, and weather. As shown in their review, age, history of no-show, referral source, day and time of appointment, insurance type, and weather and socio-economic status may be predictors for missed appointments. Gender, marital status, and distance to travel do not have any significance to missed appointments. All studies included in the review used only statistical methods to analyze the correlation between patient characteristics and no-show. All articles included in the review by Bean and Talaga (1992) were published before 1992.

Garuda et al. (1998) summarized the results of 26 studies published mostly between 1992 and 1994. The results showed that factors such as waiting time, no-show history, referral source, number of visits, day and time of appointment, socio-economic status, and transportation may influence the missed appointments. Age, gender, and race have not been shown to influence no-show appointments. Garuda et al. (1998) stated that, “demographic factors, although frequently identified and analyzed, are not as consistently predictive of no-show behavior, especially relative to other, cause-related factors”.

Our review includes articles published after 1998 (Peeters and Bayers; Bean and Talaga, 1992; Lee et al., 2005; Lehmann et al., 2007; Bennett et al., 2009) and three
articles published between 1993 and 1998 that were not included in the review by Garuda et al. (1998) (Margolis et al., 1993; Irwin et al., 1993; and McClure et al., 1996;). The predictors that are considered in this review, but were not considered in the previous two, are month of birth, first appointment, cellular phone availability, parental age, number of siblings, and residency status. The results showed that month of birth, cellular phone availability, clinic, and parental age are significant predictors of no-show. The other characteristics did not show to be significant for prediction of no-show.

In summary, there is a need for a model which could diagnose (i.e., predict) the relationship between patient characteristics and the quality performance of a healthcare system. The quality performance of a healthcare organization can be measured by the number of patients that visit the practice every day. The organization has set a number of openings according to the number of physicians in the practice and it relies on filling all of them. Consider a practice that has 10,000 annual visits. A 20% percent for no-show appointments (percent reflected in the literature review) and average revenue of $150 per visit would result in the following:

\[
10,000 \times 20\% = 2,000 \text{ no-show appointments}
\]

\[
2,000 \times $150 = $300,000 \text{ in lost revenue}
\]

Assume that the percentage of no-show appointments could be reduced to half, that is, 10%, which results in the following:

\[
10,000 \times 10\% = 1,000 \text{ no-show appointments}
\]
1,000 x $150 = $150,000 annual improvement revenue

Improvement in revenues could result in job creation, equipment acquisition, and indirectly a better quality performance for the healthcare organization. If the performance of the healthcare organization improves, patient satisfaction should improve. The main objective of a healthcare organization is the safety and health of the patients.

2.8. No-Show Appointments in Other Industries

The issue of missed appointments (no-show) is seen not only in healthcare industry, but also in all service sectors. It is more visible in the travel industry, specifically in airlines and hospitality industries. The profitability of these industries is directly related to maximum usage of their resources, such as airplane seats or hotel rooms. Because the resources in the service industry cannot be stored (a seat on a flight today cannot be used for a customer tomorrow), the resources are considered perishable goods.

Maximizing capacity in the airline industry is controlled by segregated ticket prices through yield management. Each fare price group is based on a certain percentage of available seats remaining on the flight - controlled by a computer algorithm (Wells and Wensveen, 2004; and Sasaki, 2008). Yield management, or revenue management, is the process of understanding, anticipating and influencing consumer behavior in order to maximize revenue or profits from a fixed, perishable resource. Yield management is
based on overbooking and price discrimination, which is charging different customers different prices for the same goods/services. Yield management was first implemented in 1985 by American Airlines, which credited a $1 billion/year increase in revenue to yield management (Boyd, 1998). This technique was then adopted by a majority of airlines and spread to other industries such as hotels, car rentals, and multi-family housing. Overbooking is a common practice used in airline industry and consists of selling the same seat to two different customers. Overbooking is made only for a limited number of seats (i.e., 10% in airline industry) and is done because a certain number of passengers are expected not to show up.

It becomes very important to the management to be able to forecast the customers that will not show up. Few models are used to forecast cancellations and no-shows. Conventional forecasting models use time-series models based on historical booking and no-show rates (Garrow and Koppelman, 2004). Such models consider that no-show rates are caused by flight-specific attributes such as departure time, day of the week, month, capacity, and origin and destination of flight (Ruppenthal and Toh, 1983). Time series methods take into account ‘the seasonally-weighted moving average of no-shows for previous instances of the same flight’ (Lawrence et al., 2003). However, these models do not take into account passengers’ characteristics and can be used only early in the booking process.
Recently, new models, named passenger-based predictive models, were developed. These models use explanatory variables extracted from the airline’s databases that contain passenger information and itinerary. These models compute the no-show probability for each passenger. Kalka and Weber (2000) used induction trees to compute passenger-level no-show probabilities. Hueglin and Vannoti (2001) used classification trees and logistic regression to predict no-show and cancellations. Some of the attributes used in computation and are a part of passenger characteristics are booking time, booking class, origin and destination city, day of the week, month, purpose of the trip, and total time of trip. Models similar to the model used in the airline industry can be designed and implemented in the healthcare industry to forecast no-show. These models will be a great addition to the scheduling systems used in the healthcare industry.
CHAPTER 3: ARTIFICIAL NEURAL NETWORKS

3.1 Statistical Analysis versus Neural Networks

In the literature review, the methods used to find a correlation between patient characteristics and missed appointments were statistical methods. Univariate analysis with simple regression and multivariate analysis with multiple logistic regressions were used the most. There were some inconsistencies in the results of the studies over time, especially regarding variables such as gender, race, and distance to travel, transportation, and marital status. These inconsistencies demonstrate that the statistical methods used in analysis cannot completely assess the relationship between patient characteristics and no-show rate. More complex analytic methods are needed to understand the correlation between the two.

In statistical analysis, when the variables are continuous, it is preferred to use multiple regression or analysis of variance; when the variables are dichotomous, that is they have to be classified in groups, it is preferred to use discriminant analysis or logistic regression (Mundfrom and Whitcomb, 1998). The data that will be used in this research consists of continuous and discontinuous variables therefore statistical analysis will be very difficult to use.

Missing data is great concern in statistical analysis. It can reduce the number of cases or it can introduce bias in the analysis (Mundfrom and Whitcomb, 1998). In statistical analysis, missing data rates of less than 1% are considered irrelevant, 1-5%
manageable, 5-15% requiring complex methods, and more than 15% impacting any kind of interpretation (Acuna and Rodriguez, 2009). There are few methods used to treat missing data in statistical analysis (Acuna and Rodriguez, 2009); they can be divided in three groups:

a.) Case deletion which deletes all instances with missing values. It is the easiest and the most commonly used. The limitation of this method is that it reduces the number of cases included in the analysis.

b.) Parameter estimation uses all available data to estimate the missing values. A limitation of this method is that it uses a strict assumption of a model distribution for variables, such as multivariate normal model.

c.) Imputation technique replaces missing values with values estimated from the available data employing relationships identified in the available data. This method reduces the variability of data.

For these reasons we considered that it is more appropriate to use neural networks instead of statistical analysis. Neural networks and decision trees enable the efficiency and accuracy of the data imputation (Marivate et al., 2008).

3.2. The Architecture of Artificial Neural Network

Artificial neural networks (ANNs) are mathematical or computational models that are inspired from biological neural networks. ANNs use non-linear statistical data
modeling tools to model relationships between inputs and outputs. ANNs are adaptable, being sensitive to the information that flows through the system. They are used for a wide variety of problems such classification, prediction, pattern recognition, or optimization (Jain and Mao, 1996).

Artificial neural networks have a relatively simple structure as the biological neural system that is its inspiration. Biological neurons are “very complex self-organizing systems that evolved in the course of millions of years” (Rojas, 1996). A nerve cell consists of a cell body and two types of branches, axon and dendrites. Dendrites are signal receivers and axons are signal transmitters. The connections between axons and dendrites are called synapses that are the functional unit between two neurons (Jain and Mao, 1996). Synapses learn from the activities in which they participate, those being the structure at the base of human memory. The human cerebral cortex contains about $10^{11}$ neurons, a number approximately equal to the numbers of stars in the Milky Way (Brunak and Lautrup, 1990). Each neuron is connected to $10^3$ to $10^4$ other neurons, the human brain containing $10^{14}$ to $10^{15}$ interconnections. Figure 4 is a representation of a biological neuron.
Artificial neural networks imitate the human brain. The base of the network consists of neurons (nodes) that are connected one to another (Page et al., 1993). The inputs are sent to the nodes. All inputs have different weights attached. The node receives the inputs and integrates them usually by addition. A ‘primitive function’ is then evaluated (Rojas, 1996). There are three important elements in a neural network: structure, topology, and learning algorithm used to determine the weights. The structure of the network deals with the number of nodes or processing elements and the way they are arranged. This will determine the number of layers in the networks, and the number of processing elements in each layer. In general, a neural network has an input layer, few hidden layers, and an output layer. The topology of the neural network determines how the connections between nodes are made and how the signal travels through these connections.

There are two types of network topologies:
• Feed-forward neural networks, where the signal travels only forward from the input layer to the output layer. Feed-forward networks are static, producing only one set of output values (reference, Jain et al). The classic example of feed-forward neural networks is the perceptron. The most common networks are multi layer perceptron, general feed-forward, and radial basis network.

• Recurrent (feedback) networks, with loops in their structure that provide feedback connections. These are dynamic networks allowing weights to be modified based on the feedback paths. The most used recurrent networks are competitive networks (Anderson, 1977), Kohonen’s self organizing maps (Kohonen, 1977), and Hopfield networks (Hopfield, 1982).

Learning is the fundamental trait of artificial intelligence (Jain et al., 1996). ANNs do not follow rules specified by humans, but learn the rules (input-output relationships) that already exist in a given dataset. The learning algorithm is used to determine the weights of the inputs. The weights are an important part of the neural network because they influence the quality of the output. They can be set explicitly, using a priori knowledge (NeuroAI, 2011), or can be obtained through network training. Through training, teaching patterns are fed to the network that can change the weights according to the learning rule. Learning is considered the rate of change in the weights (Dagli, 1994). There are three different types of learning algorithms: supervised, unsupervised, and hybrid.
• Supervised learning or Associative learning (Figure 5) in which pairs of input-output vectors are provided by an “external teacher” (Dagli, 1994). These pairs are used to train the network. The weights are updated until the rate of change in weights is close to zero. This signals that the network error is approaching zero. The supervised learning or “learning with a teacher” is characterized by the known values of the outputs. The weights are determined so that the network will produce an answer as close as possible to the known output value.

![Figure 5. Neural Network with Supervised Learning](image)

• Unsupervised learning or Self-Organization in which there are no ‘external teachers’. The training is completed using internal control. In this paradigm, the weights are discovered and not known a priori. In the unsupervised learning or “learning without a
teacher”, the output values are unknown; the network explores and finds correlations in the data, and organizes the patterns in categories.

- The hybrid learning is a combination of supervised and unsupervised learning. There are a few learning rules that are most used in training: error-correction, Boltzmann, Hebbian, and competitive learning. The error-correction rule uses the error signal \( (d-y) \) to modify the weights to reduce the error, where “d” is the desired output and “y” is the output generated by the network.

The back-propagation learning algorithm is the most common algorithm based on error-correction rule (Patterson, 1996; Haykin, 1994; Fausett, 1994). It works in two phases: phase 1 - propagation, and phase 2 – weight update. In the propagation phase, the signal is sent forward to generate the output activation; then the output signal is sent backward through the network to generate the delta for outputs and hidden nodes. In the weight update phase, the gradient for the weight is calculated by multiplying output delta and input activation. Based on the gradient, a learning rate is calculated. The sign of the gradient indicates if and where the error is increasing. The two phases are repeated until the error is minimized and the network performs well. A disadvantage of back-propagation is that it can get trapped at a local minimum, mainly because of the random initialization of weights. To improve the training, optimization algorithm for weights, such as gradient descent, conjugate gradients, and quasi-Newton, are used. These methods minimize the quadratic error function.
The first neural network was developed in the 1940s by McCulloch and Pitts (McCulloch and Pitts, 1943) and is known as McCulloch-Pitts network. It is a very simple network consisting of two input neurons and one output neuron; therefore, it has many limitations. Nevertheless, it created the basis for future development. In 1960s, Rosenblatt (Rosenblatt, 1962) developed the perceptron convergence theorem and in 1969, Minsky and Papert proposed a binary threshold unit for perceptron. Perceptron is a simple feed-forward neural network. Figure 6 is a representation of a perceptron.

![Perceptron Diagram](Adopted from Matthews, 2004)

Perceptron consists of a number of inputs \( x_n \), weights \( w_n \), a bias \( b \), and an output. Every input \( x_n \) (including the bias \( b \)) has a corresponding weight. The weighted sum \( \sum_{i=1}^{n} x_i * w_i + b \) is fed through an activation function that will determine the final output of the perceptron. The activation function for the perceptron is a logical function. If the weighted sum is a negative value, the activation function will return the value of
“0” (zero). If the weighted sum is a positive value, the activation function will return the value of “1” (one). The perceptron is fired if the activation function returns the value of “1” (one). This first perceptron was developed further, and one of the ways it was improved was the change of the activation function. The most used activation functions are presented in Table 2.

<table>
<thead>
<tr>
<th>Activation Function</th>
<th>Formula</th>
<th>Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step Function</td>
<td>( f(x) = 0 ) if ( x \leq 0 ) ( f(x) = 1 ) if ( x &gt; 0 )</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Identity Function</td>
<td>( f(x) = x )</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Sigmoid (Logistic) Function</td>
<td>( f(x) = \frac{1}{1+e^{-ax}} )</td>
<td><img src="image" alt="Graph" /></td>
</tr>
<tr>
<td>Hyperbolic Tangent</td>
<td>( f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} )</td>
<td><img src="image" alt="Graph" /></td>
</tr>
</tbody>
</table>

The next step was the development of the multilayer perceptron (MLP) by Werbos (1974) and Rumelhart et al. (1986). The multi layer perceptron is a modification of the perceptron. It can be used for non linearly separable problems, while the perceptron is used only for linearly separable problems. MLPs consist of multiple layers.
of neurons that have unidirectional connections between them (Jain et al., 1996) as seen in Figure 7.

Figure 7. Multi layer perceptron architecture

MLPs can have more than one hidden layer that is fully connected. Some characteristics of MLPs are connections inside layers, no direct connection between input and output layers, and the number of hidden units per layer can be more or less than the number of inputs or outputs. The input signal is sent from the input layer to the hidden layer where it is transformed by the activation function and sent forward to the next hidden layer. This process is repeated for every hidden layer until the signal reaches the output layer.

In MLPs, a signal moves from left to right. Each input has a value assigned to it called the weight which represents the importance of that particular input to the neuron. The greater the weight, the more important the input is. Some inputs can have negative
weights which mean the input inhibits the neuron. The value of each input is multiplied by the weight to emphasize the importance of an input for a specific neuron. For negative weights, the effect of multiplying the value by the weight is to deemphasize the importance (Microsoft, 2011). Multilayer perceptron became one of the most used neural networks. However, in the last decades, more complex neural networks were developed.

3.3 Neural Network Analysis

The process of determining the best model is an iterative one and is based on experimentation of different artificial neural networks (Young et al, 2008). That means one may try different network topologies (i.e., multi linear perceptron, general feed forward, radial basis), number of process elements in each layer, numbers of layers, types of activation function, and types of learning. Figure 8 shows the steps in the neural network training. The first step is data preprocessing which consists of eliminating noise, normalizing data and data coding (translating non-numerical data in numerical data). Data preprocessing will be discussed in more detail in the following sections. After the data is preprocessed, three subsets are created: training, cross-validation, and testing. The training subset is used to determine the weights in the network. Cross-validation is used to determine the weights of a network that promote generalization (Young et al., 2008). Utilization of cross-validation prevents over-fitting or memorization. The final model is
tested with the validation set data to ensure that the results on the testing and training set are real (Statistica, 2011).

![Procedural Flow for ANN Training](image)

Figure 8. Procedural Flow for ANN Training (Young et al, 2008)

After data was preprocessed and arranged into a format that is compatible with NeuroSolutions software, artificial neural network (ANN) models were designed, trained, and tested. ANNs assisted in data synthesis and interpretation, and subsequently in
decision making. Figure 9 shows neural network structure and the input variables included in the analysis.

In order to determine the best model, different topologies were used (i.e., multi linear perceptron (MLP), radial basis networks (RBN), and support vector machine). A series of parameters can be varied inside each topology:

- Percent of training data (i.e. 50% of the data is used for training)
- Number of hidden layers (one or two)
- Transfer function (i.e., linear, sigmoid, tanh)
- Learning rule (i.e., momentum, conjugate gradient, Levenberg-Marquardt, quickprop, or Delta-Bar-Delta)
- Weight update (i.e., on-line or batch)
- Number of epochs
- Termination algorithm (i.e., MSE)

The results of each ANN model are reported as learning curves, confusion matrix, and mean squared error (MSE) of training data. The model that shows a decreasing learning curve, the smallest MSE and only the diagonal of the confusion matrix populated will be the model that will be chosen.
Figure 9. Neural Network Architecture and Input Variables
3.4. Knowledge Extraction

ANNs are great tools in analyzing and modeling relationships in interactive complex non-linear systems (Weckman et al., 2010). Very often ANNs are considered “black-boxes” because their estimates appear incomprehensible. To overcome the limitation of “black-box”, knowledge extraction techniques and algorithm can be used to interpret the results and better understand them. Knowledge extraction can yield governing rules and insights into complex processes, provide rules that are human comprehensible (i.e., If-then-Else), improve accuracy in a nonlinear world and possibly derive true principles of complex behaviors. Sensitivity analysis, input-output surfaces and decision trees will be used as knowledge extraction methods.

3.4.1 Sensitivity Analysis

Sensitivity analysis is the study of how the variation in the output of a model can be quantitatively assigned to the different variation of the inputs of a model. Sensitivity analysis determines which inputs are important to the solution. Sensitivity analysis can recognize which inputs are sensitive (i.e., if an input is varied by 1%, the output changes by 10%) or insensitive (i.e., if an input is varied by 10%, the output changes by 1%). Sensitivity analysis can also be used to reduce the number of inputs by eliminating the inputs that are shown to be very insensitive.
Sensitivity analysis is used to extract cause-effect relationships between input and output variables and provides feedback as to which input variables are most significant relative to other input variables (Weckman et al., 2010). Based on the sensitivity analysis, the less significant input variables can be removed from the neural network. This will reduce the size and complexity of the network, and will decrease the training time. On the other side, by removing input variables, the impact over the output will be removed also. The results of a sensitivity analysis are displayed in a graph similar with a histogram. In general, it is recommended that 2 standard deviations should be used in sensitivity analysis.

3.4.2. Input-output surfaces

Input-output surfaces are 3D representations of the relationships between inputs and outputs. To generate these surfaces, the two inputs are varied from their sample minimum to maximum, while other input parameters are set to their sample means (Young and Weckman, 2010). The input-output surfaces can be used, in general, for parameter selection. There are two types of surfaces that can be generated:

1) by varying two inputs and observing the variation in the output
2.) by varying one input and the output and observing the variation in a second output

Figure 10 shows these types of surfaces for the corrosion dataset (Weckman et al., 2010). Figure 10.a. shows the variation in %inhibition when input variables Ni and resins
are varied between their minimums and maximums. It is a more accurate representation of the relationship that the neural network uses to estimate the output. This is explained by the fact that inputs are correlated – when one input varies, it is likely that another input will change too (Weckman et al, 2010). Figure 10.b. shows the effect of varying a sensitive value, in this case resins, and the output %inhibition. This idea of using one input and the output to solve another input value often leads to system optimization (Weckman et al., 2010). If the desired output is known, then the optimum values for the inputs can be determined.

![3D surface for Ni, resins and %inhibition](image1)

![3D surface for %inhibition, resins and Ni](image2)

<table>
<thead>
<tr>
<th>a.) 3D surface for Ni, resins and %inhibition</th>
<th>b.) 3D surface for %inhibition, resins and Ni</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 10. a.) Two Inputs Varied; b.) One Input and Output Varied (Adopted from Weckman et al., 2010).</td>
<td></td>
</tr>
</tbody>
</table>

Input-output surfaces have benefits over the sensitivity analysis method because they allow users to have more flexibility and power to determine how the model is created. It uses the whole range of values for the inputs while sensitivity analysis holds
the input values to their sample average. However, input-output surfaces can be generated after sensitivity analysis determines which input variables are the most significant.

3.4.3. Decision trees

Decision trees are a special type of graph drawn in a branching structure. Decision trees consist of nodes that are associated with logical tests and possible consequences and are the most used symbolic learning algorithms. Decision trees classify data through recursive partitioning of the data set into mutually exclusive subsets which best explain the variation in the dependent variable under observation (Biggs, et al., 1991 and Liepins, et al., 1990). Decision trees are fast, simple to implement, and can convert the learned hypothesis into a set of easily interpreted rules to mimic human reasoning in a way that gives an insight into the decision process.

A decision tree classifies instances by sorting them down the tree from the root node to leaf nodes. A leaf node gives the classification of the instance. Each branch of the decision tree represents a possible scenario of decision and its outcome. Each node of a decision tree specifies a test of some attribute, and each branch that descends from the node corresponds to a possible value for this attribute (Weckman et al., 2009). Decision trees work with multi-class classification problems and regression problems. Figure 11 shows a TREPAN decision tree for Saginaw Bay dataset created by Weckman et al. (2009). An example of rules created by a decision tree is given in Table 2.
Figure 11. TREPAN Decision Tree for Saginaw Bay Dataset (Weckman et al., 2009).

Table 3. Examples of Rules Created with TREPAN for Saginaw Bay Dataset (Weckman et al., 2009).

<table>
<thead>
<tr>
<th>Rule No.</th>
<th>Rule Text</th>
<th>Class Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>If POC&lt;&lt;1.495 and PSIO2&gt;2.021 and NH4&gt;15.029</td>
<td>CL1</td>
</tr>
<tr>
<td>2</td>
<td>If POC&gt;1.495 and POC&lt;=2.31 and PSIO2&lt;=2.485 and TSS&lt;=13.2 and POC&lt;=1.9 and Secchi&lt;=1.225</td>
<td>CL2</td>
</tr>
</tbody>
</table>

The decision tree algorithm used in this research is based on CHAID, considered one of the oldest tree classification methods. It is based on THAID algorithm (TTheta...
Automatic Interaction Detector) developed by Morgan and Messenger (1973) and it was at first proposed by Kass (1980) (Ripley, 1996). The acronym CHAID stands for Chi-squared Automatic Interaction Detector. CHAID “builds” non-binary trees that are successfully applied to very large datasets. It uses a very simple algorithm that is the Chi-square test to determine the best split at each step. The classification algorithm used in CHAID is explained below.

- **Preparing Predictors.** CHAID uses categorical predictors. So the first step would be to change any continuous predictors into categorical predictors. The algorithm does this by using an approximately equal number of observations to be included in each category. The categorical predictors are not changed.

- **Merging Categories.** The second step is to go through the predictors to determine which pair of predictors has the least influence over the dependent variable. To determine the least influential variable, it calculates the Chi-square test for each predictor. If the test is not statistically significant, it will merge the predictors and repeat the step. If the test is statistically significant, an adjusted p-value for the predictor is computed.

- **Selecting the Split Variable.** The third step is to perform the split using the predictor with the smallest p-value. If the p-value is greater than an α-value, no further split will be performed. The node becomes a terminal node.

The algorithm will continue the process until no splits can be performed.
Exhaustive CHAID Algorithm is a modification of CHAID that performs the merging until only two categories are left for each predictor. The next steps are identical to the CHAID algorithm. A disadvantage of exhaustive CHAID can be that it requires a longer computation time for large databases.

Bach and Cosic (2008) described the use of CHAID algorithm in the healthcare sector. They retrieved 221 articles published with application of decision trees in healthcare. They found that decision trees are used in “drug development and research (12%), data modeling for health care applications – e.g. nursing (11%), and executive information systems for health care (10%)”, public health informatics applications (9% of articles), and e-governance structures in health care (8% of articles), and forecasting treatment costs and demand for resources (7% of articles). CHAID was used in healthcare insurance customer profiling based on historical customer data (SMRES, 2007). The algorithm analyzed characteristics within a dataset and provided information on which type of person is more likely to buy certain products based on purchase history, geo-demographics, and lifestyle attributes (SMRES, 2007).
CHAPTER 4: METHODOLOGY

The methodology developed in this research is based on the use of artificial neural networks to find patterns in a database and apply the results to improve the scheduling system. It consists of three big parts: data preprocessing, data analysis, and knowledge extraction. The flow chart in Figure 12 shows the steps that were followed. Figure 13 is a detailed flow chart of the research.

Figure 12. Flow Chart of the Research

- Collect data – refers to data acquisition from the hospital.
- Data preprocessing is modification of the database fields so that the final format is compatible with the software to be used, making it easy to use and analyze. It is expected that the database will have missing values, duplicate values, and illegal values. Appropriate techniques such as cleansing data, feature selection and construction, feature coding will be used to deal with these issues.
• Preliminary analysis refers to the basic analysis of the data. Descriptive analysis will be used to determine the demographics of the population (i.e., age, race, gender). The results will be shown as a descriptive statistics summary and histograms.
Figure 13. Detailed Flow-Chart of the Research.
Moreover, the overall no-show, per clinic no-show rates, and per demographics no-show rates will be calculated and presented as histograms.

- Advanced analysis refers to the in-depth analysis of the database and any relationships that can be determined between patient’s characteristics in order to predict the probability of showing or not for the appointment.

- The final step consists of simulating new scheduling models that take into account the results from the advanced analysis.

Each of these steps is presented in detail in the following sections.

4.1 Data Collection

The database was obtained from a children’s hospital in Midwestern Ohio. The hospital has specialized services for cancer, diabetes and endocrine disorders, digestive disorders, heart diseases, neonatal care, neurosurgery, orthopedics, respiratory diseases, and urology. The hospital is ranked as one of the nation’s ten largest children’s hospitals and pediatric research centers. The hospital has approximately 400-beds and serves about 15,000 inpatient and 700,000 outpatient visits every year. The scheduling system used in this medical facility uses the following steps:

- Template schedules of resources (equipment, rooms, staff, and physicians) are entered into the system.
• Patient calls are directed to the central scheduling system that provides alternative appointment dates. The patient has the option of choosing from the available appointment dates.

• The hospital sends reminder letters about two weeks before the appointment date. These letters repeat the appointment time, location and anything patients need to do before the appointment (i.e., fasting for 12 hours before a procedure).

• The hospital also calls the patients about 72 hours before the appointment. These are automated calls, and patients have the option to cancel the appointment.

4.2. Data Description

The database received from the hospital consists of 87,173 appointments and is structured in four parts: specialty clinic, and three primary care clinics. Specialty clinic database includes appointments made in different specialty clinics located at three different zip codes and included 14,705 appointments between September 2\textsuperscript{nd}, 2008, and August 31\textsuperscript{st}, 2009. The primary care database included appointments made in the primary care clinic in three different locations between September 2\textsuperscript{nd}, 2008, and August 31\textsuperscript{st}, 2009. Primary care included 10,985 appointments for the first location, 32,237 appointments for the second location, and 29,246 appointments for the third location. An example from the database is shown in Table 4.
Table 4. Example of Database Features.

<table>
<thead>
<tr>
<th>Clinic Zip Code</th>
<th>Arrival Status</th>
<th>Appointment Type</th>
<th>Appointment Date</th>
<th>Appointment Type</th>
<th>Provider Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary care</td>
<td>Arrived</td>
<td>Regular Visit</td>
<td>10-Sept-08</td>
<td>05:00:00 PM</td>
<td>Physician</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Provider Name</th>
<th>Birth Year</th>
<th>Sex</th>
<th>Race</th>
<th>City</th>
<th>County Name</th>
<th>State</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2006</td>
<td>Female</td>
<td>White</td>
<td>Columbus</td>
<td>Franklin</td>
<td>OH</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Father’s Employment</th>
<th>Mother’s Employment</th>
<th>Primary Payer</th>
<th>Check-in Time</th>
<th>Vitals time</th>
<th>AVS time</th>
<th>Encounter Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employed</td>
<td>Unemployed</td>
<td>Government</td>
<td>04:53:00 PM</td>
<td>05:11:00 PM</td>
<td>06:01:00 PM</td>
<td></td>
</tr>
</tbody>
</table>

Note: The fields “clinic zip code”, “Provider name”, “Patient MRN”, “Zip code”, and “Encounter Number” are not completed due to personal data security.

The following fields were extracted from the hospital database:

- Clinic Zip Code represents the location of the clinic. There are three locations for the specialty clinic, one in the main campus, and two more about 20 miles from the main campus. There are also three locations for the primary clinic, one in the main campus and two others located about 11 miles from the main campus in the opposite direction.

- Arrival Status represents patient status as recorded by the registration office. There are seven levels recorded for patient arrival status: arrived, cancelled, departed, left without being seen, no-show, and scheduled. “Scheduled” means that the patient made the appointment. The registration staff needed to change it to "arrived", "no-show", “cancelled”, "departed" or "left without being seen" according to what happens after
scheduling. If the patient showed for the appointment and the appointment went as scheduled it was recorded as “arrived”. If the patient cancelled in advance, it was recorded as “cancelled”. If the patient decided to leave the clinic before being assessed by a nurse, the patient is "departed". If the patient was seen by a nurse and left the clinic before being seen by a doctor, it was recorded as "left without being seen". If the patient did not show up for the appointment, it was recorded as “no-show”. If the registration nurse failed to change the status of the arrival, the appointment will stay as “scheduled”.

- **Appointment Type** consists of five types of appointments: consult, education, follow-up, new patient, shot/lab, regular visit, and same day sick. Consult appointment is when the patient is part of a study. Education appointment is for training and education of the patient. Follow-up visit is designated for patients who require follow-up care. Shot/lab visits are appointments where patients receive an injection or have a lab test done. Regular visits are a general term for appointments that do not have any tests, procedure performed. Same-day sick visits refer to visits that treat needs on the same day such as an ear ache due to middle ear infection.

- **Appointment Date** is the day, month, and the year of appointment.

- **Appointment Time** represents the time of the day when the appointment was scheduled.

- **Provider Type** represents the type of staff that is scheduled to see the patient: nurse, nurse practitioner, physician, and registered dietician.
• Provider Name represents the actual provider seen. For privacy purposes, the provider information was coded.

• Patient MRN is the unique identification number for the patient. For anonymity purposes, these are not real identification number for patients.

• Patient Birth Year

• Sex represented by Male, or Female

• Race: American-Indian, White, Asian, African-American, Spanish Surname/American, or Other

• Patient Zip Code represents the zip code of the declared residence

• Patient City, County and State

• Father’s Employment Status and Mother’s Employment Status are represented by five groups: employed, self-employed, unemployed, not reported, or unknown

• Type of Insurance is represented by four groups: commercial insurance, self-pay, government insurance (Medicare and Medicaid), or other

• Check-in Time is the time when the patient sees the registration desk. At this point, the registration clerk collects patient and insurance information

• Vitals Time is the time stamp for the period when the patient sees the nurse who will collect physiological data such as blood pressure, temperature, and weight.
• AVS is the discharge time. Check-in time, vitals time, and AVS time are approximate times because the nurses and physicians, at their convenience, input them into the system.

• Encounter Number is the unique visit number.

4.3. Data Preprocessing

Data preprocessing is the first step of the data analysis. It is performed in order to transform raw data into data ready for processing procedures. Raw data was transformed into formats that are easy to use and can be effectively analyzed (i.e. non-numerical data was transformed into numerical data). The result of data preprocessing is the training dataset. The following preprocessing tools were used (Kotsiantis et al., 2006):

• Instance Selection (cleansing data) refers to identification of values that are suspicious (i.e., illegal/inconsistent values, misspellings), duplicate instance identification, and elimination. Illegal/inconsistent values refer to those values that are not in the range or interval of the majority values, or are not in the same format. For example, a gender other than female or male is considered an illegal value. These types of data are mistakes completed in the process of data collection or data typing. Misspellings are the values that were introduced incorrectly when data was typed.

• Missing Feature Values refer to incomplete data that is inevitable in dealing with real world data. The most important factor in dealing with missing data is to know what the source of the ‘unknown’ is: (a) missing value because it was lost or forgotten; (b) it
does not apply to a given instance; and (c) ‘don’t care value’ – the value is not important for the feature. The missing value in this database can be included in the first group, that is, lost or forgotten missing values. The techniques used in dealing with missing values are presented in the next sections.

- Feature Selection refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting it. This preprocessing technique reduces the size of the data and helps the learning algorithm to operate faster and more effectively. This study concentrated on the data that were redundant, and considered that all fields may be relevant to the analysis.

- Feature Construction/Transformation refers to generating new features based on the basic feature. This technique can create more concise and accurate classifiers.

- Feature Coding refers to transformation of non-numerical values in discrete numerical values.

Figure 14 shows the flow chart of the data preprocessing.

4.3.1. Instance Selection

The basic feature set was verified for misspellings, illegal values, and duplicates. Besides the missing values, no values were identified that had to be eliminated. Handling of missing data is described in the next section.
4.3.2. Missing Data Handling

The database obtained from the hospital had missing values in patient demographic fields. These missing values were organized into two groups:

- The first group consisted of patients who had more than one appointment. A case of missing values in the dataset is one where demographics information was filled in for
the first visit, but was not filled in for the reminder of the appointments. The study considered that these demographics do not change overtime. A VBA module was used to populate this type of missing field. The code for the module is presented in Appendix 3.

- The second group consisted of patients who did not have any historic demographics reported. In general, these patients were in the category of no-show appointments. They do belong to the category of new patients or returning patients. All were counted for the preliminary analysis. A request was made to the hospital to try to get more demographics for the missing data. For some patients the demographics were obtained and included in the dataset using the same VBA module (Appendix 3). The patients who did not have any demographics were removed from the advanced analysis because demographic data cannot be generated using a missing data algorithm.

Table 5 shows the modification of the number of missing data after the missing data fields were populated.

<table>
<thead>
<tr>
<th>Specialty Clinic</th>
<th>Primary Clinic 1</th>
<th>Primary Clinic 2</th>
<th>Primary Clinic 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Data</td>
<td>3831</td>
<td>3561</td>
<td>7866</td>
</tr>
<tr>
<td>After Populating</td>
<td>1818</td>
<td>0</td>
<td>1828</td>
</tr>
<tr>
<td>Populated</td>
<td>39</td>
<td>3561</td>
<td>4830</td>
</tr>
</tbody>
</table>
4.3.3. Feature Selection

Almost all features provided by the hospital were used in the analysis. Some redundant features were found, such as patient zip code, county and city. In order to avoid the redundant feature, we decided to eliminate the county and city features from the database and to transform the zip code feature into ‘distance from clinic’ feature. The transformation will be discussed in the next section. The check-in time, vitals time, and AVS time were not included in the advanced analysis, but they will be used in the simulation analysis.

4.3.4. Feature Construction

In addition to the basic feature set supplied by the hospital, a few more features were added to the database or were constructed based on existing ones. The following features were added:

- Weather characteristics for every day between September 2\textsuperscript{nd}, 2008, and August 31\textsuperscript{st}, 2010, were obtained from Weather Underground (2010). The website has weather records arranged by month and year. The access to the website services is free. Weather data was downloaded for each month (by day) of the study period. Data obtained from this service include temperature, dew point, humidity, sea level pressure, visibility, wind, precipitation, and events. Only temperature, precipitation, and events were selected to be used in this research. Temperature was given as high,
average, or low for every day. The high temperatures were chosen under the assumption that they occur during day and that travel to the clinic happened during the day. The precipitation data was given in inches of rain/snow. The weather events were recorded as none, rain, fog, snow, and as a combination of these three. Table 6 shows the type of data available. The highlighted columns represent data that were selected for analysis.
Table 6. 
Weather Daily Observations

<table>
<thead>
<tr>
<th>2009</th>
<th>Temp (F)</th>
<th>Dew point (F)</th>
<th>Humidity (%)</th>
<th>Atmospheric Pressure (in)</th>
<th>Visibility (mi)</th>
<th>Wind (mph)</th>
<th>Precipitation (in)</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>H</td>
<td>A</td>
<td>L</td>
<td>H</td>
<td>A</td>
<td>L</td>
<td>H</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>24</td>
<td>18</td>
<td>12</td>
<td>10</td>
<td>8</td>
<td>71</td>
<td>54</td>
</tr>
<tr>
<td>2</td>
<td>41</td>
<td>30</td>
<td>19</td>
<td>24</td>
<td>16</td>
<td>13</td>
<td>78</td>
<td>58</td>
</tr>
</tbody>
</table>

After data was obtained and selected, they were coded for use with Neurosolution software. Temperature and precipitation are continuous attributes that have relationships between them, and it was not necessary to code them. Events are discrete attributes and it was necessary to code them. Table 7 shows the coding for weather events.

Table 7. Weather Events Coding

<table>
<thead>
<tr>
<th>Weather Event</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fog</td>
<td>1</td>
</tr>
<tr>
<td>Rain</td>
<td>2</td>
</tr>
<tr>
<td>Snow</td>
<td>3</td>
</tr>
<tr>
<td>Thunderstorm</td>
<td>4</td>
</tr>
<tr>
<td>Fog, Rain</td>
<td>12</td>
</tr>
<tr>
<td>Fog, Snow</td>
<td>13</td>
</tr>
<tr>
<td>Fog, Rain, Thunderstorm</td>
<td>124</td>
</tr>
<tr>
<td>Rain, Thunderstorm</td>
<td>24</td>
</tr>
<tr>
<td>Rain, Snow</td>
<td>23</td>
</tr>
<tr>
<td>None</td>
<td>5</td>
</tr>
</tbody>
</table>

- Distance from home to the facility where the appointment took place was calculated based on the provider and patient zip codes using a CDX Technologies add-in for Microsoft Excel. The add-in was purchased. The add-in calculates the distance between two zip codes. Distance can be calculated as driving distance or as a radius. The driving distance was chosen. Figure 15 shows the add-in and its options.
Figure 15. CDX technologies add-in for Microsoft Excel.

- Time Spent in the Clinic was calculated using the four time stamps: appointment time, check-in time, vitals time, and AVS time. It was assumed that the arrival time is the same as the appointment time. Two waiting times can be determined: nurse waiting time that can be calculated as the difference between the vitals time and the check-in time; and the doctor waiting time that can be calculated as the difference between AVS time and vitals time. Total waiting time can be calculated by adding the two waiting times.

- Patient Age was calculated by subtracting the birth year from the year of appointment.

- Date of Appointment was transformed from one feature to three different features: year, month, and day of month. The year was eliminated after it was used to calculate the patient age. The day of appointment was transformed from a date to the day of the week using Microsoft Excel function “WEEKDAY()” as follows: Sunday became “1”,...
Monday “2”, Tuesday “3”, Wednesday “4”, Thursday “5”, Friday “6”, and Saturday “7”.

- Time of Appointment was modified from a given time to a time interval. Table 8 shows the transformation of the time of appointment into codes. For appointment times before 8.00 AM a code of 1 was used.

<table>
<thead>
<tr>
<th>Time of Appointment</th>
<th>Time Interval</th>
<th>Time of Appointment</th>
<th>Time Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-10.00 AM</td>
<td>1</td>
<td>4.01-6.00 PM</td>
<td>5</td>
</tr>
<tr>
<td>10.01-12.00 PM</td>
<td>2</td>
<td>6.01-8.00 PM</td>
<td>6</td>
</tr>
<tr>
<td>12.01-2.00 PM</td>
<td>3</td>
<td>8.01-10.00 PM</td>
<td>7</td>
</tr>
<tr>
<td>2.01-4.00 PM</td>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Some socio-economic attributes were included: average income, poverty level, proficiency test, and education level.

- Average Income was obtained from Income Tax List (2010), a free online service. The number of returns, average income, federal tax, and effective tax rate can be obtained for each zip code. Only average income was included in this research. Table 9 shows data that was used to obtain average income by zip code.

<table>
<thead>
<tr>
<th>Zip Code</th>
<th>Number of Returns</th>
<th>Average Income</th>
<th>Federal Tax</th>
<th>Effective Tax Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Athens, OH 45701</td>
<td>9,710</td>
<td>$45,070</td>
<td>$7,490</td>
<td>12.0%</td>
</tr>
</tbody>
</table>
Poverty Level and Education Level were obtained from the U.S. Census Bureau (USCB, 2011b).

The field “Arrival Status” was removed from the input features and then added as an output feature.

### 4.3.5. Feature Coding

The basic feature set (obtained from the hospital) had two types of values: numerical and non-numerical. The fields that contained numerical values were used as is. The fields that contained non-numerical values were coded using numerical values. The coding is shown in Table 10 and described below:

- The types of Arrival Status were coded as follows: 1 – arrived, 0- no-show, 2 – cancelled, 3 – departed, 4- left without being seen, 5 – scheduled.
- Type of Appointment was coded as follows: 10 – consult; 11 – new patient; 12 – regular visit or follow-up (for specialty clinic; 13 – same day sick visit; 14 – shot/lab; 15 – education.
- Type of Provider was coded as: 411 – physician; 412 – nurse; 413- nurse practitioner; 414- pharmacist; 415- educator; and 416 – registered dietician.
- Employment Status was coded as follows: 2011/2012 – employed father/mother; 2021/2022 unemployed father/mother; 2031/2032 self-employed father/mother; and 2041/2042 unknown/not reported/do not use father/mother.
- Type of Insurance: 301 – government insurance; 302 – commercial insurance; 303 – other; and 304 – self pay.

- Gender: 91 – Female, and 92 – Male.

- Race: 101 – Asian; 102 – Black; 103 – White; 104 – Spanish surname/American; 105 – Other.

Table 10. Data Coding

<table>
<thead>
<tr>
<th>Appointment Type</th>
<th>Employment Status</th>
<th>Arrival Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consult</td>
<td>Employed</td>
<td>Arrived</td>
</tr>
<tr>
<td>New patient</td>
<td>Unemployed</td>
<td>Cancelled</td>
</tr>
<tr>
<td>Regular visit</td>
<td>Self-employed</td>
<td>Departed</td>
</tr>
<tr>
<td>Follow up</td>
<td>Not reported</td>
<td>Left without being seen</td>
</tr>
<tr>
<td>Same day sick visit</td>
<td>Unknown</td>
<td>Scheduled</td>
</tr>
<tr>
<td>Shot/lab</td>
<td>Do not use</td>
<td>No show</td>
</tr>
<tr>
<td>Education</td>
<td></td>
<td>Worklist</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Provider Type</th>
<th>Insurance Type</th>
<th>Race</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician</td>
<td>Government insurance</td>
<td>Asian</td>
</tr>
<tr>
<td>Nurse</td>
<td>Commercial insurance</td>
<td>Black</td>
</tr>
<tr>
<td>Nurse practitioner</td>
<td>Other</td>
<td>White</td>
</tr>
<tr>
<td>Pharmacist</td>
<td>Self-pay</td>
<td>Spanish surname/American</td>
</tr>
<tr>
<td>Educator</td>
<td></td>
<td>Other</td>
</tr>
<tr>
<td>Registered dietician</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
</tr>
<tr>
<td>Male</td>
</tr>
</tbody>
</table>

These codes were used only for preliminary analysis. For the advanced analysis, the coding was changed to accommodate neural network analysis requirements. The
function “Tag as symbol” from Neuro Solution add-in for Microsoft Excel was used to tag attributes that needed to be coded for analysis. It is an automatic feature of the add-in that transforms data into binary data. “Tag as symbol provides a quick method for tagging multiple columns of data. If a column of data contains any non-numeric values (except for labels), it will be tagged as symbol” (Neuro solution manual, 2010). After the columns are tagged as symbol, the “(S)” prefix will be added in front of the column’s label, and the column will be recognized as symbolic by the software. Symbolic columns are unary coded. This can be done before the training is done by using the “Translate Symbolic Columns” sub-menu item from the Preprocess Data menu item. There are two conditions under which a column can be tagged as symbolic: the column contains non-numerical data, or is used for a classification problem. For the second situation, the desired output column has to be tagged as symbolic, too. Unary code of n (natural number) is generally represented by a sequence of “n” 1 bits followed by a “0” bit. The alternative representation is a sequence of “0” bits followed by a “1” bit (Kak, 2010). Neuro Solutions software follows the alternative representation. An example is shown in Table 11.

Table 11. Code transformation using “Tag as symbol” function

<table>
<thead>
<tr>
<th>Before Coding</th>
<th>After Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Gender 91</td>
</tr>
<tr>
<td>91</td>
<td>1</td>
</tr>
<tr>
<td>92</td>
<td>0</td>
</tr>
<tr>
<td>93</td>
<td>0</td>
</tr>
</tbody>
</table>
Following the coding for neural network analysis, the number of features increased. Each subgroup of the data will become a new feature (i.e., gender feature includes female and male; after coding ‘female’ and ‘male’ will become independent features).

4.4. Preliminary Analysis

A preliminary analysis was performed in order to determine the characteristics of the population included in the study, the overall and per clinic no-show rate, and the no-show rate for different demographics (i.e. race, gender, insurance type, appointment type). The flow chart in Figure 16 shows the steps followed in the preliminary analysis.

![Figure 16. Preliminary Analysis Flow Chart.](image)
4.4.1. Analysis of Basic Feature Set

The basic feature set was analyzed to determine the basic characteristics of the population such as age groups, gender, race, and insurance type. The following information was obtained:

- The average age of the population is 5.18 years (SD=5.01) for primary clinic #1, 4.63 years (SD=4.93) for primary clinic #2, 4.77 years (SD=4.69) for primary clinic #3, and 11.30 years (SD=5.17) for specialty clinic. More descriptive statistics for age are summarized in Table 12.

<table>
<thead>
<tr>
<th>Descriptive Statistics</th>
<th>Primary Clinic 1</th>
<th>Primary Clinic 2</th>
<th>Primary Clinic 3</th>
<th>Specialty Clinic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5.18</td>
<td>4.63</td>
<td>4.77</td>
<td>11.30</td>
</tr>
<tr>
<td>Median</td>
<td>3.00</td>
<td>3.00</td>
<td>3.00</td>
<td>12.00</td>
</tr>
<tr>
<td>Mode</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>16.00</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>5.01</td>
<td>4.93</td>
<td>4.69</td>
<td>5.17</td>
</tr>
<tr>
<td>Range</td>
<td>21.00</td>
<td>22.00</td>
<td>31.00</td>
<td>48.00</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Maximum</td>
<td>21.00</td>
<td>22.00</td>
<td>31.00</td>
<td>48.00</td>
</tr>
<tr>
<td>Confidence Level (95.0%)</td>
<td>0.10</td>
<td>0.10</td>
<td>0.06</td>
<td>0.10</td>
</tr>
</tbody>
</table>

Figure 17 shows the mean age, standard deviation, and median and mode for the four clinics. The age “0” means a patient who has the age in the range 0 to 12 months, an age of “1” means a patient between 12 and 24 months old, an age of “2” means a patient between 2 and 3 years old, and so on. The specialty clinic has a much higher mean for
age. This can be explained by a different reason: the diseases treated by the specialty clinic have the first occurrence at an older age, or patients are followed for a long period of time.

Figure 17. Mean, Median, Mode, and Standard Deviation of Population Age by Clinic.

- The gender distribution is almost the same across all clinics with a small difference for specialty clinic. Figure 18 shows the gender distribution. The gender distribution of the population included in this research is switched in comparison to the US gender distribution. It can be seen from Figure 18 that there are more male than female patients while the US gender distribution shows that there are more females than
males. Only the population from the specialty clinic follows the national trend for gender distribution.

Figure 18. Gender Distribution for Each Clinic, Overall Clinics and United States.

- Race Distribution varies from one clinic to another. For the primary clinics 1 and 2, the majority population is African-American (64.76% and 59.87% respectively); for the specialty clinic, the majority population is White (78.43%). For primary clinic #3, the population is spread somewhat equally among African-American, White, Spanish surname/American and Other. Only the specialty clinic follows the national race distribution. Race distribution is shown in Figure 19.
Insurance Type varies between clinics, with the majority being government insurance except in the specialty clinic where the majority of insurance type is commercial insurance. Figure 20 shows the distribution for Insurance Type. The healthcare coverage distribution in the population included in this research is different from the national healthcare coverage distribution. The national insurance type distribution is uniformly distributed between government insurance and commercial insurance (45% and 49%, respectively), with a very small percentage of self-paid (5%) insurance, while for the three primary clinics the majority of insurance is government insurance.
Figure 20. Insurance Type Distribution for Each Clinic, Overall Clinics and United States.

- Provider Type consists of physicians, nurses, nurse practitioners, pharmacists, educators, and registered dieticians. The majority of appointments are scheduled for physicians for all clinics. Specific to the specialty clinic is the registered dietician with a percentage of 13.74%. Figure 21 shows the Provider Type Distribution.
Appointment Type distribution was determined for each clinic. The results are shown as histograms in Figures 22. The majority of appointment type for each clinic is regular visit/follow up (range between 71.65% for primary clinic 3, and 83.63% for specialty clinic). The other types of appointments have lower prevalence, in general (lower than 17%).
Employment Type distribution was determined for each clinic. The results are shown as histograms in Figure 23. The employment type distribution varies from one clinic to another. Primary clinics 2 and 3 have a higher percentage of unemployed patients (60.39% and 72.45%, respectively). Specialty clinic has a higher percentage of employed patients (59.33%), while primary clinic 1 has a high percentage of “not reported” employment status (65.43%).
4.4.2. Determination of No-Show Rates

A preliminary analysis was performed using Microsoft Excel to determine the number of no-show appointments and the no-show rates. The following steps were followed:

- Filters for “Arrival Status” were identified as arrived, cancelled, departed, and left without being seen, no-show and scheduled.
- Data was filtered so that only no-show appointments are shown. The results are recorded in Table 14.
- Filters for “Appointment Type” were identified as consult follow-up, education, new patient, shot/lab, and same day sick.
• The no-show rates were calculated based on the formula:

\[
No - show\ rate = \frac{Number\ of\ no - show\ in\ appointment\ group}{Total\ number\ of\ appointments\ in\ group}
\]

• The overall no-show rate was calculated using the following formula:

\[
Overall\ no - show\ rate = \frac{Total\ number\ of\ no - show}{Total\ number\ of\ appointments}
\]

• The results are recorded in Table 13. Figure 24 is a histogram of no-show rates by appointment type.

The preliminary analysis results of the database showed that the following:

• The overall no-show rate is 19.59%.

• The rate of no-show appointments is 12.7% (1867 no-shows) in the specialty clinic, 28.34% in primary care #1, 19.81% primary care #2, and 19.54% primary care #3.
<table>
<thead>
<tr>
<th>Specialty</th>
<th>No.</th>
<th>Rate %</th>
<th>No.</th>
<th>Rate %</th>
<th>No.</th>
<th>Rate %</th>
<th>No.</th>
<th>Rate %</th>
<th>No.</th>
<th>Rate %</th>
<th>No.</th>
<th>Rate %</th>
<th>No.</th>
<th>Rate %</th>
<th>Total No-show</th>
<th>No. Appointments</th>
<th>No-show Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specialty</td>
<td>286</td>
<td>15.32</td>
<td>1553</td>
<td>83.18</td>
<td>2</td>
<td>0.11</td>
<td>26</td>
<td>1.39</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1867</td>
<td>14705</td>
<td>12.70</td>
</tr>
<tr>
<td>Clinic 1</td>
<td>261</td>
<td>8.38</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>136</td>
<td>4.37</td>
<td>-</td>
<td>-</td>
<td>2526</td>
<td>81.14</td>
<td>190</td>
<td>6.10</td>
<td>3113</td>
<td>10985</td>
<td>28.43</td>
</tr>
<tr>
<td>Clinic 2</td>
<td>399</td>
<td>6.25</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>173</td>
<td>2.71</td>
<td>46</td>
<td>0.72</td>
<td>5243</td>
<td>82.10</td>
<td>525</td>
<td>8.22</td>
<td>6386</td>
<td>32237</td>
<td>19.81</td>
</tr>
<tr>
<td>Clinic 3</td>
<td>319</td>
<td>5.58</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>232</td>
<td>4.06</td>
<td>62</td>
<td>1.08</td>
<td>4639</td>
<td>81.17</td>
<td>463</td>
<td>8.10</td>
<td>5715</td>
<td>29246</td>
<td>19.54</td>
</tr>
</tbody>
</table>

Overall No-show Rate 19.59%
In the no-show appointment group, follow-up and regular visits had the highest no-show rate 81.89% (average of the four groups) for the no-show group.

New Patients Group had the highest no-show rate for the specialty clinic that is 15.32%. For the primary care clinics new patient no-show rates vary from 5.58% to 8.38%.

The lowest no-show rates are for the Education Appointments and range between 0.08% and 0.66%.

Figure 24. No-show Rates by Appointment Type.

The no-show rates by clinic are presented in Figures 25, 26, 27, and 28: primary care clinic 1, primary care clinic 2, primary care clinic 3, and specialty clinic
respectively. The no-show rates were calculated after preprocessing of the basic data set. The results show that the no-show rates are 23.25% for primary clinic 1, 16.68% for primary care clinic 2, 0.73% for primary care clinic 3, and 0.45% for specialty clinic. These rates are different from those calculated before preprocessing of the data. Table 14 shows both rates for each clinic, before and after data preprocessing.

Figure 25. No-show Rates for Primary Care Clinic 1.
Figure 26. No-show Rates for Primary Care clinic 2.

Figure 27. Arrival Status for Primary Care Clinic 3.
Figure 28. Arrival Status for Specialty Clinic.

Table 14. Comparison of No-Show Rates Before and After Preprocessing

<table>
<thead>
<tr>
<th>No-show rates</th>
<th>Before Preprocessing</th>
<th>After Preprocessing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Clinic 1</td>
<td>28.34%</td>
<td>23.25%</td>
</tr>
<tr>
<td>Primary Clinic 2</td>
<td>19.81%</td>
<td>16.68%</td>
</tr>
<tr>
<td>Primary Clinic 3</td>
<td>19.54%</td>
<td>0.73%</td>
</tr>
<tr>
<td>Specialty Clinic</td>
<td>12.70%</td>
<td>0.45%</td>
</tr>
</tbody>
</table>

4.5. Advanced Analysis

Advanced analysis as part of this project consists of in-depth analysis of the relationship between patient characteristics (described in section 3.1.) and no-show appointments. Advanced analysis is shown in Figure 29. It consists of three steps: set-up
and training of neural networks, selection of best neural network, and knowledge extraction from the best model. A short presentation of the history and characteristics of ANNs was presented in section 1.5.

Figure 29. Flow Chart of Advanced Analysis.
4.5.1. Application of Neural Network in No-show Analysis

Different network models were deployed to determine the best model. Patient characteristics were used as inputs to the network. The inputs included in the analysis are structured into three groups:

- Related to the patient such as race, gender, age, employment status, and insurance type.
- Related to the environment such as provider type, appointment type, and time and day of appointment.
- Determined based on information from databases and include weather, distance to travel, proficiency scores, education level, and average income.

The first two types of inputs were recorded at the time of registration, and the last one was determined based on information available in the database. There is no specific order for the inputs to be included in the neural network analysis. They are not grouped into groups as they were presented above.

Neural Network analysis was performed for all four clinics as described in the previous section. Different parameters of neural networks (i.e., number of layers, numbers of epochs, and activation function) were used so that different models were developed for each clinic. After the models were developed, the best model was chosen to be included in the next step of advanced analysis, that is, knowledge extraction. The results are presented in the following paragraphs. The table shows the number of layers, the learning algorithm, and the number of epochs as parameters for the neural network.
model and r-square, mean square error, and correct classification as results of the analysis.

4.5.1.1 Primary Clinic 1

Primary Clinic 1 consists of 10,985. The models developed for primary clinic 1 are summarized in Table 15.

Table 15. Summary of Models for Primary Clinic 1

<table>
<thead>
<tr>
<th>Hidden Layers</th>
<th>Learning Algorithm</th>
<th>Number of Epochs</th>
<th>r-square</th>
<th>MSE</th>
<th>Correct Classification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Arrival</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>32,16,10*</td>
<td>Conjugate Gradient</td>
<td>1000</td>
<td>0.0078</td>
<td>0.0050</td>
<td>0.2346</td>
</tr>
<tr>
<td>32,16,10*</td>
<td>Conjugate Gradient</td>
<td>1000</td>
<td>0.0047</td>
<td>0.0104</td>
<td>0.2449</td>
</tr>
<tr>
<td>32,16,10*</td>
<td>Conjugate Gradient</td>
<td>3000</td>
<td>0.0070</td>
<td>0.0083</td>
<td>0.2377</td>
</tr>
<tr>
<td>32,16,10*</td>
<td>Conjugate Gradient</td>
<td>10000</td>
<td>0.0047</td>
<td>0.0007</td>
<td>0.3637</td>
</tr>
<tr>
<td>50,30,10*</td>
<td>Conjugate Gradient</td>
<td>1000</td>
<td>0.0115</td>
<td>0.0122</td>
<td>0.1983</td>
</tr>
<tr>
<td>50,30,10*</td>
<td>Conjugate Gradient</td>
<td>10000</td>
<td>0.0104</td>
<td>0.0096</td>
<td>0.2330</td>
</tr>
<tr>
<td>17,8,6</td>
<td>Conjugate Gradient</td>
<td>1000</td>
<td>0.0315</td>
<td>0.0209</td>
<td>0.2502</td>
</tr>
<tr>
<td>17,8,6</td>
<td>Conjugate Gradient</td>
<td>10000</td>
<td>0.0197</td>
<td>0.0217</td>
<td>0.2543</td>
</tr>
<tr>
<td>17,8,5</td>
<td>Conjugate Gradient</td>
<td>10000</td>
<td>0.0216</td>
<td>0.0132</td>
<td>0.2561</td>
</tr>
<tr>
<td>4,4,4,</td>
<td>Conjugate Gradient</td>
<td>1000</td>
<td>0.0278</td>
<td>0.0230</td>
<td>0.2476</td>
</tr>
</tbody>
</table>

Note: * - models that include all arrival status (1,2,3,4,5,6)
4.5.1.2. Primary Clinic 2

Primary clinic 2 consists of 32,237 initial appointments from which 1828 were removed due to missing information. The models developed for primary clinic 3 are summarized in Table 16.

Table 16. Summary of Models for Primary Clinic 2

<table>
<thead>
<tr>
<th>Hidden Layers</th>
<th>Learning Algorithm</th>
<th>Number of Epochs</th>
<th>r-square Arrivals</th>
<th>MSE Arrivals</th>
<th>Correct Classification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50,45,30</td>
<td>Conjugate Gradient</td>
<td>1000</td>
<td>0.6871</td>
<td>0.0304</td>
<td>99.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.8116</td>
<td>0.0432</td>
<td>92.96</td>
</tr>
<tr>
<td>50,45,30</td>
<td>Conjugate Gradient</td>
<td>1000</td>
<td>0.8420</td>
<td>0.0301</td>
<td>99.58</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.6922</td>
<td>0.0409</td>
<td>94.35</td>
</tr>
<tr>
<td>50,48,28</td>
<td>Conjugate Gradient</td>
<td>1000</td>
<td>0.9322</td>
<td>0.0106</td>
<td>99.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.9306</td>
<td>0.0106</td>
<td>96.19</td>
</tr>
<tr>
<td>28,14,9</td>
<td>Conjugate Gradient</td>
<td>3000</td>
<td>0.9397</td>
<td>0.0087</td>
<td>99.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.9390</td>
<td>0.0087</td>
<td>96.39</td>
</tr>
</tbody>
</table>

Note: * - models that include all arrival status (1,2,3,4,5,6)

4.5.1.3. Primary Clinic 3

Primary Clinic 3 consists of 29,246 initial appointments from which 6831 were removed due to missing information. The models developed for primary clinic 3 are summarized in Table 17.
Table 17. Summary Models for Primary Clinic 3

<table>
<thead>
<tr>
<th>Hidden Layers</th>
<th>Learning Algorithm</th>
<th>Number of Epochs</th>
<th>r-square</th>
<th>MSE</th>
<th>Correct Classification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Arrival</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>50,33,22</td>
<td>Conjugate Gradient</td>
<td>3000</td>
<td>0.0018</td>
<td>0.0007</td>
<td>0.0331</td>
</tr>
<tr>
<td>50,33,22</td>
<td>Conjugate Gradient</td>
<td>10000</td>
<td>0.0003</td>
<td>0.0008</td>
<td>0.0290</td>
</tr>
</tbody>
</table>

4.5.1.4. Specialty clinic

Specialty Clinic consists of 14,705 initial appointments from which 3,644 were removed due to missing information. 11,061 appointments were used in the advanced analysis. The models developed for specialty clinic are summarized in Table 18.

Table 18. Summary Models for Specialty Clinic

<table>
<thead>
<tr>
<th>Hidden Layers</th>
<th>Learning Algorithm</th>
<th>Number of Epochs</th>
<th>r-square</th>
<th>MSE</th>
<th>Correct Classification (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Arrival</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>32,16,10*</td>
<td>Conjugate Gradient</td>
<td></td>
<td>0.6096</td>
<td>0.1407</td>
<td>0.0372</td>
</tr>
<tr>
<td>30,15,10</td>
<td>Conjugate Gradient</td>
<td>1000</td>
<td>0.8457</td>
<td>0.8466</td>
<td>0.0008</td>
</tr>
<tr>
<td>30,15,10</td>
<td>Conjugate Gradient</td>
<td>3000</td>
<td>0.8709</td>
<td>0.8640</td>
<td>0.0065</td>
</tr>
<tr>
<td>30,20,10</td>
<td>Conjugate Gradient</td>
<td>3000</td>
<td>0.8619</td>
<td>0.8551</td>
<td>0.0069</td>
</tr>
</tbody>
</table>

Note: * - models that include all arrival status (1,2,3,4,5, and 6)
4.6. Knowledge Extraction

4.6.1. Sensitivity Analysis

Sensitivity Analysis was performed for all models included in the research (PC2 and SC). It was performed using Neuro Solution software. Variables were ranked based on the scores from sensitivity analysis. A comparison between different sensitivity analyses was performed using the square weight error that was calculated in the following steps:

- Variables were sorted alphabetically.
- The mean weight for each variable for all models for one clinic was calculated.
- The weight value for each model was subtracted from the mean value and squared.
- The squared weight errors were added to obtained the total squared weight error for each model.
- The model with the lowest error was chosen for development of decision trees that will be discussed in the next section.

Table 19 shows the results for Sensitivity Analysis for primary clinic 2. The model that has the lowest squared weight error is the one with three hidden layers, and 45-30-03 processing elements. The squared weight error is 0.7069.
Table 19. Total Weight Error for Primary Clinic 2

<table>
<thead>
<tr>
<th>Deviation Squared</th>
<th>28_14_9_02</th>
<th>50_43_28_02_10</th>
<th>50_45_30_03_6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.000197887</td>
<td>7.00802E-05</td>
<td>0.000503491</td>
</tr>
<tr>
<td>Appointment Day</td>
<td>0.140213236</td>
<td>0.505313676</td>
<td>0.057658651</td>
</tr>
<tr>
<td>Appointment Time</td>
<td>0.060402681</td>
<td>0.210988767</td>
<td>0.03887725</td>
</tr>
<tr>
<td>Appointment Type</td>
<td>9.59105E-05</td>
<td>0.210989477</td>
<td>0.057188352</td>
</tr>
<tr>
<td>Average Income</td>
<td>0.000373196</td>
<td>0.00317659</td>
<td>3.20395E-05</td>
</tr>
<tr>
<td>Distance</td>
<td>0.00019615</td>
<td>6.82233E-05</td>
<td>0.000379397</td>
</tr>
<tr>
<td>Education Level</td>
<td>0.00011279</td>
<td>4.69415E-06</td>
<td>3.961E-06</td>
</tr>
<tr>
<td>Events</td>
<td>0.000117695</td>
<td>6.64499E-06</td>
<td>0.000184927</td>
</tr>
<tr>
<td>Father's Employment</td>
<td>0.094902651</td>
<td>0.135676511</td>
<td>0.006911799</td>
</tr>
<tr>
<td>Mother's Employment</td>
<td>0.074999938</td>
<td>0.259998021</td>
<td>0.010396321</td>
</tr>
<tr>
<td>Poverty Level</td>
<td>0.00028132</td>
<td>0.00175317</td>
<td>0.000141284</td>
</tr>
<tr>
<td>Precipitation</td>
<td>0.000339297</td>
<td>0.00262638</td>
<td>0.000295998</td>
</tr>
<tr>
<td>Primary Payor</td>
<td>0.35738886</td>
<td>1.194706</td>
<td>0.455520693</td>
</tr>
<tr>
<td>Proficiency Scores</td>
<td>0.000104426</td>
<td>2.08552E-06</td>
<td>0.024060316</td>
</tr>
<tr>
<td>Provider Name</td>
<td>0.017738748</td>
<td>0.065887865</td>
<td>0.00457417</td>
</tr>
<tr>
<td>Provider Type</td>
<td>0.004604918</td>
<td>0.017348049</td>
<td>0.03188481</td>
</tr>
<tr>
<td>Race Description</td>
<td>0.011425127</td>
<td>0.043827005</td>
<td>0.00586796</td>
</tr>
<tr>
<td>Sex Description</td>
<td>0.005003964</td>
<td>0.020816595</td>
<td>0.012887837</td>
</tr>
<tr>
<td>Temperature</td>
<td>0.00010408</td>
<td>1.99833E-06</td>
<td>0.000238746</td>
</tr>
<tr>
<td><strong>Total Weight Error</strong></td>
<td><strong>0.768062872</strong></td>
<td><strong>2.666461307</strong></td>
<td><strong>0.706911672</strong></td>
</tr>
</tbody>
</table>

Table 20 shows the weights for each variable in primary clinic 2. As can be seen in the table, the variables with the larger weights are primary payer (0.6961), appointment day (0.2613), appointment type (0.2603), appointment time (0.2184), provider type (0.1978), gender (0.1347), mother’s and father’s employment (0.1232 and 0.1043, respectively) and proficiency scores (0.1763). The rest of the variables have a minimal
influence on arrival status and were removed from the analysis. Figure 30 shows the
weight for all variables included in the analysis.

Table 20. Sensitivity Analysis for Primary Clinic 2

<table>
<thead>
<tr>
<th>Variable</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Payor</td>
<td>0.696110</td>
</tr>
<tr>
<td>Appointment Day</td>
<td>0.261310</td>
</tr>
<tr>
<td>Appointment Type</td>
<td>0.260328</td>
</tr>
<tr>
<td>Appointment Time</td>
<td>0.218361</td>
</tr>
<tr>
<td>Provider Type</td>
<td>0.197790</td>
</tr>
<tr>
<td>Proficiency Scores</td>
<td>0.176301</td>
</tr>
<tr>
<td>Sex Description</td>
<td>0.134712</td>
</tr>
<tr>
<td>Mother's Employment</td>
<td>0.123150</td>
</tr>
<tr>
<td>Father's Employment</td>
<td>0.104325</td>
</tr>
</tbody>
</table>

Figure 30. Sensitivity Analysis for Primary Clinic 2.
Table 21 shows the results for sensitivity analysis for specialty clinic. The model that has the lowest squared weight error is the one with three hidden layers, and 30-15-10 processing elements. The squared weight error is 0.003.

<table>
<thead>
<tr>
<th>Deviation Squared</th>
<th>30_15_10_02_25</th>
<th>3HL</th>
<th>30_15_10_02_10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>3.11454E-07</td>
<td>9.6E-09</td>
<td>2.11707E-07</td>
</tr>
<tr>
<td>Appointment Date(2)</td>
<td>0.000101963</td>
<td>0.000373</td>
<td>0.050730903</td>
</tr>
<tr>
<td>Appointment Time(1)</td>
<td>0.000240323</td>
<td>0.000822</td>
<td>0.005511566</td>
</tr>
<tr>
<td>Appointment Type(10)</td>
<td>1.43725E-05</td>
<td>5.99E-05</td>
<td>0.006327902</td>
</tr>
<tr>
<td>Average Income</td>
<td>1.70783E-07</td>
<td>3.68E-07</td>
<td>1.13354E-07</td>
</tr>
<tr>
<td>Clinic Zip code(43017)</td>
<td>0.000156045</td>
<td>0.000565</td>
<td>0.014428419</td>
</tr>
<tr>
<td>Distance</td>
<td>2.25093E-06</td>
<td>2.2E-06</td>
<td>9.65396E-06</td>
</tr>
<tr>
<td>Father's Employment(2011)</td>
<td>0.000136062</td>
<td>0.000519</td>
<td>0.018732094</td>
</tr>
<tr>
<td>Higher Education percentage</td>
<td>8.90543E-07</td>
<td>2.58E-07</td>
<td>3.20762E-08</td>
</tr>
<tr>
<td>Mother's Employment(2011)</td>
<td>0.001666051</td>
<td>0.005837</td>
<td>0.04454213</td>
</tr>
<tr>
<td>Poverty Percentage</td>
<td>1.80367E-06</td>
<td>1.46E-06</td>
<td>6.31522E-06</td>
</tr>
<tr>
<td>Precipitations</td>
<td>5.75827E-08</td>
<td>5.13E-07</td>
<td>9.68026E-07</td>
</tr>
<tr>
<td>Primary Payor(301)</td>
<td>0.000316453</td>
<td>0.00095</td>
<td>0.005802357</td>
</tr>
<tr>
<td>Proficiency Test</td>
<td>1.01997E-06</td>
<td>2.72E-07</td>
<td>1.4575E-05</td>
</tr>
<tr>
<td>Provider Name(10)</td>
<td>0.000587018</td>
<td>0.001924</td>
<td>0.085745049</td>
</tr>
<tr>
<td>Provider Type(411)</td>
<td>8.48562E-05</td>
<td>0.00309</td>
<td>0.018431413</td>
</tr>
<tr>
<td>Race Description(101)</td>
<td>0.000172215</td>
<td>0.000587</td>
<td>0.006518864</td>
</tr>
<tr>
<td>Sex Description(91)</td>
<td>1.09594E-05</td>
<td>4.71E-05</td>
<td>0.017224637</td>
</tr>
<tr>
<td>State(1)</td>
<td>1.64471E-06</td>
<td>1.71E-05</td>
<td>0.000932781</td>
</tr>
<tr>
<td>Temperature</td>
<td>2.21451E-07</td>
<td>8.78E-08</td>
<td>1.85698E-06</td>
</tr>
<tr>
<td>Weather events</td>
<td>6.78419E-07</td>
<td>4.17E-08</td>
<td>3.55336E-06</td>
</tr>
<tr>
<td><strong>Total Weight Error</strong></td>
<td><strong>0.003495367</strong></td>
<td><strong>0.012015</strong></td>
<td><strong>0.274965395</strong></td>
</tr>
</tbody>
</table>
Table 22 shows the weights for each variable in specialty clinic. As noticed from the table, the variables with the highest impact are mother’s and father’s employment (0.0423 and 0.0132, respectively), provider name (0.02574), primary payer (0.0193), appointment time (0.01701), race (0.0146), appointment date (0.0116), and provider type (0.0107). Figure 31 shows the weight for all variables included in the analysis.

Table 22. Sensitivity Analysis for Specialty Clinic

<table>
<thead>
<tr>
<th>Variable</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mother's Employment</td>
<td>0.04233</td>
</tr>
<tr>
<td>Provider Name</td>
<td>0.02574</td>
</tr>
<tr>
<td>Primary Payer</td>
<td>0.01930</td>
</tr>
<tr>
<td>Appointment Time</td>
<td>0.01701</td>
</tr>
<tr>
<td>Race Description</td>
<td>0.01464</td>
</tr>
<tr>
<td>Clinic Zip Code</td>
<td>0.01400</td>
</tr>
<tr>
<td>Father's Employment</td>
<td>0.01318</td>
</tr>
<tr>
<td>Appointment Date</td>
<td>0.01161</td>
</tr>
<tr>
<td>Provider Type</td>
<td>0.01072</td>
</tr>
</tbody>
</table>
4.5.2 Decision Tree

Decision trees were created for Primary Clinic 2 and Specialty Clinic. These decision trees were used to validate the results from the sensitivity analysis and to derive rules that can be used in the scheduling system. The decision trees were created using SPSS software. The type of data included in the analysis consists of categorical and continuous data. As mentioned in section 3.6, the CHAID algorithm uses only categorical data. The continuous data was transformed into categorical data by creating categories.

Table 23 shows the decision tree analysis specification for Primary Clinic 2.
Table 23. Decision tree specification for CHAID analysis for primary clinic 2

<table>
<thead>
<tr>
<th>Model Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specifications</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Validation</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Results</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
The maximum tree, which represents the maximum numbers of levels in the tree, was set up for 10. The minimum number of cases in the parent node was set up for 100, and the minimum number of cases in child node was set up for 50. The dependent variable was Arrival Status and the independent variables were Appointment Type, Appointment Day, Appointment Time, Temperature, Precipitation, Events, Provider Type, Provider Name, Education Level, Poverty Level, Proficiency Scores, Age, Sex Description, Race Description, Distance, Average income, Father's Employment, Mother's Employment, and Primary payer. The tree had a total of 116 nodes on 10 levels from which 67 nodes are terminal nodes. The terminal nodes give the probabilities for arrival 1 (patient arrived for the appointment) and arrival 6 (patient did not arrive for the appointment or no-show). Table 24 shows the classification performed by the CHAID algorithm. The algorithm classified correctly 99.6% of arrival 1 and 96.7% of arrival 6.

Table 24. Arrival Status Classification Using CHAID Analysis

<table>
<thead>
<tr>
<th>Classification</th>
<th>Predicted</th>
<th>Percent Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>23712</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>168</td>
<td>0</td>
</tr>
<tr>
<td>Overall Percentage</td>
<td>79.9%</td>
<td>.0%</td>
</tr>
</tbody>
</table>

Growing Method: CHAID
Dependent Variable: (S)Arrival Status
Figure 32 shows the entire decision tree. Figure 33 shows only the right part of the tree that corresponds to the arrival status 6; that is, patient not showing for the appointment or no-show.
Figure 32. Decision Tree for Primary Clinic 2.
Figure 33. Decision Tree for Arrival Status 6.
Based on this decision tree there are 35 rules for arrival status 6. These rules give the probability of patient no-show based on the patient, appointment, and environmental characteristics. The rules are presented in Table 25. It can be seen that the characteristics that influence patient no-show are primary payer (insurance type), type of weather events occurring on the day of appointment, appointment day, race, temperature, distance to travel to clinic, provider type to be seen, and age. There are four rules with probabilities between 0.4-0.5, six rules with probabilities between 0.6-0.7, nine between 0.7-0.8 and the rest are higher than 0.80.

Table 25. Decision Tree Rules for Primary Clinic 2 for Arrival Status 6 (no-show)

<table>
<thead>
<tr>
<th>Node number</th>
<th>Node Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node 6</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name) ≤ 63), P(6)= 0.486842</td>
</tr>
<tr>
<td>Node 62</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature ≤ 37) AND (Provider Type ≤ 411) AND (Poverty level ≤ 12.8), THEN P(6)= 0.615385.</td>
</tr>
<tr>
<td>Node 86</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature ≤ 37) AND (Provider Type ≤ 411) AND (Poverty level &gt; 12.8) AND (Provider Name ≤ 67), THEN P(6)= 0.815018</td>
</tr>
<tr>
<td>Node 87</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 67) AND (Provider Name ≤ 69) AND (Temperature ≤ 37) AND (Provider Type ≤ 411) AND (Poverty level &gt; 12.8), THEN P(6)= 0.864023</td>
</tr>
<tr>
<td>Node 88</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 69) AND (Provider Name ≤ 71) AND (Temperature ≤ 37) AND (Provider Type ≤ 411) AND (Poverty level &gt; 12.8), THEN P(6)= 0.727273</td>
</tr>
<tr>
<td>Node 42</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature ≤ 37) AND (Provider Type &gt; 411), THEN P(6)= 0.661765</td>
</tr>
<tr>
<td>Node 43</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 37) AND (Temperature ≤ 50) AND (Events ≠ 3) AND (Events ≠ 1), THEN P(6)= 0.858871</td>
</tr>
</tbody>
</table>
Table 25. Continued

<table>
<thead>
<tr>
<th>Node number</th>
<th>Node Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node 44</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature) &gt; 37 AND (Temperature) ≤ 50 AND (Events = 30 OR (Events = 1), THEN P(6)=0.563636</td>
</tr>
<tr>
<td>Node 64</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature) &gt; 50 AND (Temperature ≤ 55) AND (Appointment Day ≠ 2) AND (Provider Name ≤ 68), THEN P(6)=0.812298</td>
</tr>
<tr>
<td>Node 65</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature) &gt; 50 AND (Temperature ≤ 55) AND (Appointment Day ≠ 2) AND (Provider Name &gt; 68), THEN P(6)=0.622951</td>
</tr>
<tr>
<td>Node 66</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature) &gt; 50 AND (Temperature ≤ 55) AND (Appointment Day = 2) AND (Precipitation ≤ 0), THEN P(6)=0.540541</td>
</tr>
<tr>
<td>Node 67</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature) &gt; 50 AND (Temperature ≤ 55) AND (Appointment Day = 2) AND (Precipitation &gt; 0), THEN P(6)=0.826923</td>
</tr>
<tr>
<td>Node 104</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature) &gt; 55 AND (Temperature ≤ 77) AND (Appointment Type) ≤ 12) AND (Race description ≤ 102) AND (Precipitation ≤ 0) AND (Events ≠ 2) AND (Events ≠ 1) THEN P(6)=0.898551</td>
</tr>
<tr>
<td>Node 105</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature) &gt; 55 AND (Temperature ≤ 77) AND (Appointment Type ≤ 12) AND (Race description) ≤ 102) (Precipitation ≤ 0) AND (Events = 2) OR (Events = 1), THEN P(6)=0.784000</td>
</tr>
<tr>
<td>Node 90</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature) &gt; 55 AND (Temperature ≤ 77) AND (Appointment Type ≤ 12) AND (Race description ≤ 102) AND (Precipitation &gt; 0.070) AND (Precipitation ≤ 0.070), THEN P(6)=0.898551</td>
</tr>
<tr>
<td>Node 91</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature) &gt; 55 AND (Temperature ≤ 77) AND (Appointment Type ≤ 12) AND (Race description ≤ 102) AND (Precipitation &gt; 0.070), THEN P(6)=0.805556</td>
</tr>
<tr>
<td>Node 69</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature) &gt; 55 AND (Temperature ≤ 77) AND (Appointment Type ≤ 12) AND (Race description &gt; 102) AND (Race description &gt; 103), THEN P(6)=0.782456</td>
</tr>
<tr>
<td>Node number</td>
<td>Node Rule</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------</td>
</tr>
<tr>
<td>Node 70</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 55) AND (Temperature ≤ 77) AND (Appointment Type) ≤ 12) AND (Race description &gt; 103), THEN P(6)= 0.786096</td>
</tr>
<tr>
<td>Node 48</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 55) AND (Temperature ≤ 77) AND (Appointment Type) ≤ 12), THEN P(6)= 0.636364</td>
</tr>
<tr>
<td>Node 71</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 77) AND (Temperature ≤ 80) AND (Events ≠ 2) AND (Events ≠ 24) AND (Events ≠ 4) AND (Appointment Day = 6) OR (Appointment Day = 4), THEN P(6)= 0.806202</td>
</tr>
<tr>
<td>Node 72</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 77) AND (Temperature ≤ 80) AND (Events ≠ 2) AND (Events ≠ 24) AND (Events ≠ 4) AND (Appointment Day ≠ 6) AND (Appointment Day ≠ 4), THEN P(6)= 0.936047</td>
</tr>
<tr>
<td>Node 73</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 77) AND (Temperature ≤ 80) AND (Events = 2) OR (Events = 24) OR (Events = 4) AND (Appointment Day ≠ 6) AND (Sex Description ≠ 91), THEN P(6)= 0.580645</td>
</tr>
<tr>
<td>Node 92</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 77) AND (Temperature ≤ 80) AND (Events = 2) OR (Events = 24) OR (Events = 4) AND (Appointment Day ≠ 6) AND (Sex Description ≠ 91), THEN P(6)= 0.827586</td>
</tr>
<tr>
<td>Node 93</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 77) AND (Temperature ≤ 80) AND (Events = 2) OR (Events = 24) OR (Events = 4) AND (Appointment Day ≠ 6) AND (Sex Description = 91), THEN P(6)= 0.651515</td>
</tr>
<tr>
<td>Node 106</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 80) AND (Appointment Day = 6) OR (Appointment Day = 5) AND (Sex Description ≠ 91) AND (Distance ≤ 4.3955) AND (Fathers Employment ≤ 2011), THEN P(6)= 0.803922</td>
</tr>
<tr>
<td>Node 107</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 80) AND (Appointment Day = 6) OR (Appointment Day = 5) AND (Sex Description ≠ 91) AND (Distance ≤ 4.3955) AND (Fathers Employment &gt; 2011), THEN P(6)= 0.944444</td>
</tr>
</tbody>
</table>
Table 25. Continued

<table>
<thead>
<tr>
<th>Node number</th>
<th>Node Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node 95</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 80) AND (Appointment Day = 6) OR (Appointment Day = 5) AND (Sex Description) OR (Sex Description ≠ 91) AND (Distance &gt; 4.3955), THEN P(6)= 0.746032</td>
</tr>
<tr>
<td>Node 76</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 80) AND (Appointment Day = 6) OR (Appointment Day = 5) AND (Sex Description) = 91, THEN P(6)= 0.766839</td>
</tr>
<tr>
<td>Node 96</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 80) AND (Appointment Day ≠ 6) AND (Appointment Day ≠ 4) AND (Appointment Day ≠ 5) AND (Events ≠ 2) AND (Events ≠ 1) AND (Race description ≤ 103), THEN P(6)= 0.875949</td>
</tr>
<tr>
<td>Node 97</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 80) AND (Appointment Day ≠ 6) AND (Appointment Day ≠ 4) AND (Appointment Day ≠ 5) AND (Events ≠ 2) AND (Events ≠ 1) AND (Race description &gt; 103), THEN P(6)= 0.865672</td>
</tr>
<tr>
<td>Node 78</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 80) AND (Appointment Day ≠ 6) AND (Appointment Day ≠ 4) AND (Appointment Day ≠ 5) AND (Events = 2) OR (Events = 1), THEN P(6)= 0.780000</td>
</tr>
<tr>
<td>Node 53</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 63) AND (Provider Name ≤ 71) AND (Temperature &gt; 80) AND (Appointment Day = 4), THEN P(6)= 0.728507</td>
</tr>
<tr>
<td>Node 27</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 71) AND (Events ≠ 3) AND (Events ≠ 13), THEN P(6)= 0.739583</td>
</tr>
<tr>
<td>Node 28</td>
<td>IF (Primary Payor &gt; 304) AND (Provider Name &gt; 71) AND (Events = 3) OR (Events = 13), THEN P(6)= 0.637500</td>
</tr>
</tbody>
</table>

4.6. Validation

The results of the methodology produced a list of rules that presented the relationships between patient characteristics and no-show appointments accurately. It is expected that the rules resulting from the analysis were designed to be used in decisions
during the patient scheduling process. It is also expected that the results from the knowledge extraction are accurate (generalization), consistent (repeatable), and comprehensive (easy to understand). These results were used in the simulation step, which is the last step in this research and is presented in the next section. The validation of results was done in two different ways:

- A simulation study was performed. The rules extracted from the neural networks analysis were implemented in the scheduling system. The influence of the implemented rules on the scheduling system was recorded.

- A survey was sent to the experts in the field from different healthcare organizations (i.e., hospitals, outpatient clinics, private practices). The survey included questions about the importance of such a tool, outcomes desired by healthcare organizations, and potential users. Appendix 5 shows the survey and the accompanying introduction letters. Appendix 6 shows the IRB approval letter for the use of the survey.

4.6.1 Simulation

Simulation is the imitation of a real-world situation with a mathematical model that does not affect operations (Render et al., 2006). The model is then experimented with to estimate the effects of various actions and decisions (Render et al., 2006). Figure 34 shows the steps that will be followed in the simulation process. In the first stage, patients’ data will randomly be collected from the database and introduced in the scheduling software. A schedule for 1 day will be generated. The results will be recorded. In the
second stage, patients’ data will be collected and the rules generated in the advanced analysis will be introduced in the scheduling software. The use of rules generated in advanced analysis will influence the scheduling algorithm. It is believed that these rules will influence the scheduling algorithm, consequently creating new schedules. These schedules will be compared with the schedules obtained in the first stage of the simulation process.

Figure 34. Flow Chart of The Simulation Process.

Figure 35 shows the steps of the simulation process.
Table 26 shows the simulation steps of the patient scheduling based on the outcome from the advanced analysis.

Table 26. Steps for the Simulation Process

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Define Problem</td>
<td>The reduction in no-show rate will improve organizational performance. The organization needs to evaluate new scheduling techniques based on forecasting of no-shows.</td>
</tr>
<tr>
<td>Introduce Important Variables</td>
<td>Define variables associated with the defined problem that will be included in the mathematical model</td>
</tr>
<tr>
<td>Construct Simulation Model</td>
<td>The model is built based on the results from the analysis of the relationship between patient characteristics and no-show rates. The effects of different scheduling techniques of the organization performance will be analyzed.</td>
</tr>
<tr>
<td>Specify Values of Variables to Be Tested</td>
<td>The rules extracted from advanced analysis will be used.</td>
</tr>
<tr>
<td>Conduct Simulation</td>
<td>Run the experiment after variables were specified.</td>
</tr>
<tr>
<td>Examine Results</td>
<td>The effects of different scheduling techniques will be analyzed.</td>
</tr>
<tr>
<td>Select Best Course of Action</td>
<td>The scenario with the best results (i.e., most patients’ scheduled, minimum idle time for staff and doctors, shortest span) will be accepted.</td>
</tr>
</tbody>
</table>
The problem that will be simulated is the improvement of the organizational performance through reduction of no-show rates. The forecasting of no-show appointments should be used in scheduling to optimally use the possible available time slots. The problem variables are patient characteristics, appointment characteristics, and environmental characteristics as described in previous sections. The simulation model is built on the basis of the rules extracted from the advanced analysis. The outcomes of the simulation model will be the number of patients scheduled every day, idle time for staff and doctors, total length of time of appointments, and waiting time. The simulation software should be able to apply the rules as input and change the schedule each time a new patient who fits in the rule description is entered in the system, so that the desired outcomes would be obtained.

The validation of the simulation would be the comparison of the model to the real system to evaluate its accuracy. The assumptions of the model should be analyzed to decide if the appropriate probability distribution is being used. An analysis of inputs and outputs should be made to see that the results are reasonable. If it is known what actual outputs result from a specific set of inputs, those inputs could be used in the computer model to see that the outputs of the simulation are consistent with the real world system.

A day of appointments from primary clinic 2 was selected. There were 142 appointments scheduled for a 12 hours period, from which 102 arrived and 40 were no-show appointments. The clinic has 5 providers, 4 doctors and 1 nurse. The doctors have different schedules: three of them work a nine hour schedule, while 2 of them work only
a 3 hour schedule. For the doctors that work a full shift (nine hours schedule) and the nurse, it is assumed that they have a one hour lunch break. The check-in times used in the simulation are the times provided by the database. We assumed that the check-in processing time is 5 minutes, vitals processing time is 10 minutes, and doctor processing time is 15 minutes. The numbers of patients scheduled for the 8 hours shift is 106 patients. From these patients, 21 are no-shows, making the no-show rate equal to about 20%.

The schedule was recreated to determine provider’s idle time, patient’s waiting time, the number of patients seen, and the total time of the day. The results of the simulation are shown in Table 25. The simulation of the actual day was performed using simulation software. The actual arriving time was used to create entities. Each entity travelled through the system in the following order:

- Entity was created by the source based on the actual check-in time.
- Each entity went to check-in station. The processing time in this station is considered to be 5 minutes. The model had 2 check-in stations. After check-in is finished, entities wait in the outer buffer to be called to the next station. The entities are allocated to each station based on “first in, first out” rule.
- The vitals’ station has a processing time of 10 minutes. There are 2 vital stations in the system. After the vitals were taken, entities wait in the outer buffer to be called to the last station. The entities are allocated to each station based on “first in, first out” rule.
• Doctor’s station has a processing time of 15 minutes. There are 3 doctors in the system working from 8-5 PM. The entities are allocated to each station based on “first in, first out” rule.

• After work is performed at final station, the entities are destroyed.

Model properties are detailed as follows:

• The model runs first from 8-5 PM which is the regular working time, and then from 8 AM until the last patient is seen.

• Each station can serve only one entity at the time.

• There is an arrival pattern for entities that is used to create entities. The arrival pattern cannot be repeated.

• Each entity has assigned a probability of no-show of “0” and “1”. A probability of no-show of “1” means that the entity will not arrive and subsequently will not be created.

• A total of 106 entities is considered, from which 21 have a probability of no-show of “0” and will not be created, leaving only 85 entities to enter the system.

• Entities are destroyed after they go through all stations. They cannot go back in the system.

The results of the first simulation (8-5 PM) are shown in Table 27. It can be seen that 85 patients arrived at the clinic, from which 45 were checked in at first station and 39 checked in at the second station. At the end of the shift (5.00PM) there were 2 patients waiting to be seen by the nurse. At the vitals station, 38 patients were seen by one nurse and 42 were seen by the second nurse. At the end of the shift 7 patients were waiting to
be seen by the doctor. The simulation also shows that the utilization time for doctors is at 74.79% overall for an eight hour shift from 8-5:00 PM. The lunch break of 60 minutes is accounted for in the calculation of utilization time. The low percent of utilization time is the result of the no-show rate that is about 20% overall for the clinic. The simulation validates the no-show rate and the influence of the no-shows on the utilization time.

Table 27. Results for Actual Status Simulation from 8-5 PM

<table>
<thead>
<tr>
<th>Number processed</th>
<th>Waiting Time (hours)</th>
<th>Schedule Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Entered</td>
<td>Exit</td>
</tr>
<tr>
<td>Source</td>
<td>85</td>
<td>85</td>
</tr>
<tr>
<td>Checkin1</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Checkin2</td>
<td>39</td>
<td>37</td>
</tr>
<tr>
<td><strong>Total check in</strong></td>
<td><strong>84</strong></td>
<td><strong>82</strong></td>
</tr>
<tr>
<td>Nurse1</td>
<td>38</td>
<td>34</td>
</tr>
<tr>
<td>Nurse2</td>
<td>42</td>
<td>39</td>
</tr>
<tr>
<td><strong>Total nurse</strong></td>
<td><strong>80</strong></td>
<td><strong>73</strong></td>
</tr>
<tr>
<td>Doctor1</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Doctor2</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td>Doctor3</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td><strong>Total doctors</strong></td>
<td><strong>71</strong></td>
<td><strong>71</strong></td>
</tr>
</tbody>
</table>

Table 28 shows the results for the simulation ran from 8 AM to the time the last patient is seen. The simulation ran from 8 AM-6.55 PM, making the total time of the
day equal to 10.92 hours. Considering that there is a 1 hour lunch break, the staff worked for 9.92 hours. This is the time considered when percent utilization time is calculated. The overall percent utilization time is 71.40% for the three doctors. This is consistent with the rate of no-show. The average waiting time for patients to see the doctor is 0.46 hours or 27.6 minutes, and the maximum patient waiting time is 1.19 hours or 71.4 minutes. The total average patient waiting time is 0.78 hours or 46.8 minutes, and the total maximum patient waiting time is 2.08 hours or 124.8 minutes.

Table 28. Results for Actual Status Simulation from 8 AM to Last Patient Seen

<table>
<thead>
<tr>
<th></th>
<th>Number processed</th>
<th>Waiting Time (hours)</th>
<th>Schedule Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Entered</td>
<td>Exit</td>
<td>Min</td>
</tr>
<tr>
<td>Source</td>
<td>85</td>
<td>85</td>
<td>-</td>
</tr>
<tr>
<td>Checkin1</td>
<td>45</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>Checkin2</td>
<td>40</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>Total check in time</td>
<td>85</td>
<td>85</td>
<td>0</td>
</tr>
<tr>
<td>Nurse1</td>
<td>40</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>Nurse2</td>
<td>45</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>Total nurse</td>
<td>85</td>
<td>85</td>
<td>0</td>
</tr>
<tr>
<td>Doctor1</td>
<td>28</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>Doctor2</td>
<td>40</td>
<td>40</td>
<td>0</td>
</tr>
<tr>
<td>Doctor3</td>
<td>17</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>Total doctors</td>
<td>85</td>
<td>85</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 29 shows the minimum, maximum, and average number of patients waiting to see a provider, and the minimum, maximum, and average waiting time. It can be seen that doctor 2 has the greatest waiting time (a maximum of 1.4 hours = 84 minutes) and the largest number of patients waiting (a maximum of 7) which is consistent with the utilization time. It was followed by doctor 1 with the maximum waiting time of 1.05 hour and a maximum number of patients waiting equal to five. The waiting time for vitals is on average 30 minutes and for check in is on average between 1.2-3.6 minutes.

<table>
<thead>
<tr>
<th></th>
<th>Number of patients</th>
<th>Waiting Time (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Minimum</td>
<td>Maximum</td>
</tr>
<tr>
<td><strong>Doctor 1</strong></td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td><strong>Doctor 2</strong></td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td><strong>Doctor 3</strong></td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>Nurse 1</strong></td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td><strong>Nurse 2</strong></td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td><strong>Check in 1</strong></td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td><strong>Check in 2</strong></td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

The first scenario was developed to include different probabilities of no-show. These were created based on the decision-tree probability of no-show. The first scenario was developed based on the assumptions that the appointments are scheduled 15 minutes apart. The probability of no-show is determined for each patient. We set a threshold of .75 for the no-show probability, so we included as no-show those instances where the
probability is greater than “0.75”. For all other instances the probability of no-show is considered “0”. The same assumptions for processing times as for the initial simulation were considered.

The results for the simulation are shown in Table 30. The simulation ran from 8AM – 6.05 PM, making the total working time to be 9.08 hours with lunch break excluded. The total utilization time increased to 83.52% (about 17% more), and the average patient waiting time to see a doctor decreased to 0.30 hours or 18 minutes from an average of 0.46 hours, a decrease of about 35%. The total average patient waiting time increased to 0.81 hours or 48.6 minutes from an average of 0.78 hours (only 3.8%). The total number of patients seen in a day increased from 85 to 91, an increase of 7%.

Table 30. Results for Simulation with Different No-Show Appointments Randomly Distributed

<table>
<thead>
<tr>
<th>Number processed</th>
<th>Waiting Time (hours)</th>
<th>Schedule Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Entered</td>
<td>Exit</td>
</tr>
<tr>
<td>Checkin1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entered</td>
<td>52</td>
<td>52</td>
</tr>
<tr>
<td>Exit</td>
<td>39</td>
<td>39</td>
</tr>
<tr>
<td>Checkin2</td>
<td>46</td>
<td>46</td>
</tr>
<tr>
<td>Total check in</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>Nurse1</td>
<td>46</td>
<td>46</td>
</tr>
<tr>
<td>Nurse2</td>
<td>45</td>
<td>45</td>
</tr>
<tr>
<td>Total nurse</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>Doctor1</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>Doctor2</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Doctor3</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>Total doctors</td>
<td>91</td>
<td>91</td>
</tr>
<tr>
<td>Total</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
The second, third and fourth scenarios were developed using the same assumptions as for the second scenario, except that the no-show appointments were not randomly distributed during the day, but they were grouped and assigned to the beginning of each hour and before lunch, and towards the end of the day, respectively. For the fourth scenario a deviation of 3.61 minutes was included for arrival time. The deviation was calculated from the actual arrival times used in the first simulation. Tables 31, 32, and 33 show the results of the simulation. Doctors’ utilization time increased to 85.25%, 87.34%, and 90.40% respectively. The average patient waiting time decreased to 0.25 hours or 15 minutes and 0.22 hours or 13.2 minutes for second and third scenario respectively. The average waiting time for doctors increased to 0.34 hours for the fourth scenario. The number of patients seen in one day increased to 98 and 101 respectively. Scenario two and three ran from 8AM – 6.35 PM or for 9.25 hours excluding the lunch time, and scenario four ran for 9.31 hours excluding the lunch break.
Table 31. Results for Simulation with Different No-Show Appointments Grouped

<table>
<thead>
<tr>
<th></th>
<th>Number processed</th>
<th>Waiting Time (hours)</th>
<th>Schedule Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Entered</td>
<td>Exit</td>
<td>Min</td>
</tr>
<tr>
<td>Checkin1</td>
<td>52</td>
<td>52</td>
<td>0</td>
</tr>
<tr>
<td>Checkin2</td>
<td>46</td>
<td>46</td>
<td>0</td>
</tr>
<tr>
<td>Total check in</td>
<td>98</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>Nurse1</td>
<td>53</td>
<td>53</td>
<td>0</td>
</tr>
<tr>
<td>Nurse2</td>
<td>45</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>Total nurse</td>
<td>98</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>Doctor1</td>
<td>35</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>Doctor2</td>
<td>27</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>Doctor3</td>
<td>36</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>Total doctors</td>
<td>98</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 32. Results for Simulation with Different No-show Appointments Grouped

<table>
<thead>
<tr>
<th></th>
<th>Number processed</th>
<th>Waiting Time (hours)</th>
<th>Schedule Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Entered</td>
<td>Exit</td>
<td>Min</td>
</tr>
<tr>
<td>Checkin1</td>
<td>59</td>
<td>59</td>
<td>0</td>
</tr>
<tr>
<td>Checkin2</td>
<td>39</td>
<td>39</td>
<td>0</td>
</tr>
<tr>
<td>Total check in</td>
<td>98</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>Nurse1</td>
<td>50</td>
<td>50</td>
<td>0</td>
</tr>
<tr>
<td>Nurse2</td>
<td>48</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>Total nurse</td>
<td>98</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>Doctor1</td>
<td>35</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>Doctor2</td>
<td>27</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>Doctor3</td>
<td>36</td>
<td>36</td>
<td>0</td>
</tr>
<tr>
<td>Total doctors</td>
<td>98</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 33. Results for Simulation with Different No-show Appointments Grouped and with Deviation of Arrival

<table>
<thead>
<tr>
<th>Source</th>
<th>Number processed</th>
<th>Waiting Time (hours)</th>
<th>Schedule Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Entered</td>
<td>Min</td>
<td>Max</td>
</tr>
<tr>
<td>Checkin1</td>
<td>59</td>
<td>0</td>
<td>0.33</td>
</tr>
<tr>
<td>Checkin2</td>
<td>42</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>Total check in</td>
<td>101</td>
<td>0</td>
<td>0.29</td>
</tr>
<tr>
<td>Nurse1</td>
<td>48</td>
<td>0</td>
<td>1.08</td>
</tr>
<tr>
<td>Nurse2</td>
<td>53</td>
<td>0</td>
<td>1.58</td>
</tr>
<tr>
<td>Total nurse</td>
<td>101</td>
<td>0</td>
<td>1.29</td>
</tr>
<tr>
<td>Doctor1</td>
<td>38</td>
<td>0</td>
<td>1.25</td>
</tr>
<tr>
<td>Doctor2</td>
<td>29</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Doctor3</td>
<td>34</td>
<td>0</td>
<td>0.83</td>
</tr>
<tr>
<td>Total doctors</td>
<td>101</td>
<td>0</td>
<td>1.00</td>
</tr>
<tr>
<td>Total</td>
<td>-</td>
<td>-</td>
<td>2.58</td>
</tr>
</tbody>
</table>

Figures 36, 37, 38 and 39 show a summary of doctors’ utilization time, average patient waiting time, and total number of patients seen in one day. As mentioned above, doctors’ utilization time increased from 71.40% to 90.40%, average patient waiting time to see a doctor decreased from 0.46 hours to 0.22 hours for scenario 3 and then increased for scenario 4, and the number of patients seen in one day increased from 85 to 101. However, the total maximum waiting time, that includes check in waiting time, vitals’ waiting time and doctor waiting time, increased from 2.08 hours to 2.58 hours (24%), while total average waiting time increased from 0.78 hours to 0.89 hours (14%).
Figure 36. Summary of Doctors’ Utilization Time for the Four Simulations.

Figure 37. Summary of Average Patient Waiting Time to See a Doctor
Figure 38. Summary of Total Maximum and Average Patient Waiting Time.

Figure 39. Total Number of Patients Seen in One Day.
4.6.2. *Survey Results*

The survey was sent to members of the Healthcare Division of the American Society of Quality. The division has approximately 4,000 members. From all invited members, we received 159 responses which made the response rate to be about 3.85%. The respondents had to give their consent to be able to access the survey. From the 159 respondents, three (1.9% of respondents) did not give their consent and could not continue taking the survey. The second question of the survey filtered the respondents by industry. They could choose as an answer “Healthcare industry” or “Other”. Thirty one respondents (21.2% of total respondents) answered “Other” and they were disqualified because they could not be considered experts in the field of the healthcare industry. For the rest of the survey, there were between 85-104 respondents that skipped the questions, so the analysis of the responses from the survey will focus only on a number between 55-71 respondents.

The third question of the survey asked, “Who/what decides whether and when to schedule a patient?” The responses are summarized in Table 34. We can see that an overwhelming percentage of 77.1% from the answers consist of “Human Scheduler” as a method for patients’ scheduling. It seems that the majority of healthcare providers do not use any algorithm in scheduling patients. The scheduling is done by a scheduler who has to look for an available spot available. In this case, the old manual book is replaced by the computer “spreadsheets” without other improvements.
Table 34. Question 3 Results

**Who/what decides whether and when to schedule a patient?**

<table>
<thead>
<tr>
<th>Answer Options</th>
<th>Response Percent</th>
<th>Response Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Scheduler</td>
<td>77.1%</td>
<td>54</td>
</tr>
<tr>
<td>Algorithm/Model</td>
<td>0.0%</td>
<td>0</td>
</tr>
<tr>
<td>Human Scheduler Assisted by Algorithm/Model</td>
<td>25.7%</td>
<td>18</td>
</tr>
</tbody>
</table>

Answered Question 70

The next question in the survey, question 4, is very important because it can give some insight into how satisfied the respondents are with the method they used. The average score of the sixty nine respondents was 2.91 out of 5 which shows that the scheduling methods currently used may not perform well. Only nineteen out of the sixty nine respondents (27.5%) answered that the methodology they used is somehow good, leaving 72.5% answering that the methodology they use is less than satisfying. The results are shown in Table 35.

Table 35. Question 4 Results

| On a scale from 0 to 5 , how good is the methodology/tool you are currently using? |
|----------------------------------|------------------|----------------|
| Answer Options                   | Unsatisfied | Very Satisfied | Rating Average | Response Count |
| Current Methodology/Tool         | 8            | 13             | 29             | 15             | 4              | 2.91           | 69             |

Answered Question 69
Question 5 in the survey asked, “What information do you or others at your place of employment use when scheduling patients?” This is an open-ended answer question; only fifty five respondents answered it. The information that is considered the most when scheduling patients is physician availability, patient’s preference of day and time, urgency of medical issue, reason for visit, patient type (regular or new patient), insurance type, diagnosis, and appointment type (new, follow-up, consult). Demographic information is somewhat requested. From the demographic information, the most used are name, age, and gender. The answers are summarized in Table 36.
Table 36. Question 5 Answer Summary

<table>
<thead>
<tr>
<th>Information Used When Scheduling Patients</th>
<th>Number of Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physician/Slots Availability</td>
<td>21</td>
</tr>
<tr>
<td>Patient's Preference</td>
<td>14</td>
</tr>
<tr>
<td>Urgency of Medical Issue</td>
<td>12</td>
</tr>
<tr>
<td>Reason for Visit</td>
<td>9</td>
</tr>
<tr>
<td>Previous Patient</td>
<td>8</td>
</tr>
<tr>
<td>Insurance</td>
<td>7</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>5</td>
</tr>
<tr>
<td>Appointment Type</td>
<td>5</td>
</tr>
<tr>
<td>Procedure Type</td>
<td>5</td>
</tr>
<tr>
<td>Demographic Information</td>
<td>4</td>
</tr>
<tr>
<td>Age, Date of Birth</td>
<td>6</td>
</tr>
<tr>
<td>Name</td>
<td>4</td>
</tr>
<tr>
<td>Time/Distance to Office</td>
<td>4</td>
</tr>
<tr>
<td>Referral</td>
<td>4</td>
</tr>
<tr>
<td>Gender</td>
<td>2</td>
</tr>
<tr>
<td>Medication, Other Therapies</td>
<td>2</td>
</tr>
<tr>
<td>Phone Number</td>
<td>1</td>
</tr>
<tr>
<td>Date Last Seen</td>
<td>1</td>
</tr>
<tr>
<td>Availability of Transportation</td>
<td>1</td>
</tr>
<tr>
<td>Ability to Come on Short Notice</td>
<td>1</td>
</tr>
<tr>
<td>Special Needs of the Patient</td>
<td>1</td>
</tr>
<tr>
<td>Patient Preparation</td>
<td>1</td>
</tr>
<tr>
<td>Medical Records</td>
<td>1</td>
</tr>
<tr>
<td>Pharmacy</td>
<td>1</td>
</tr>
<tr>
<td>Primary Care Provider</td>
<td>1</td>
</tr>
<tr>
<td>Health Risk Assessment Score</td>
<td>1</td>
</tr>
</tbody>
</table>

The next questions, number 6 and 7, asked, “How useful do you believe a decision support tool that predicts patient no-show would be in healthcare patient scheduling?” and “Gauge your interest level in investigating the possibility of adopting
such a decision support tool?”. Sixty nine responses were recorded for question six, from which an overwhelming 76.81% (53 responses) indicated that a decision support tool that predicts patient no-show can be very useful in healthcare patient scheduling. Sixty seven responses were recorded for question 7, from which 71.64% showed interest in adopting a tool to forecast no-show patients. The answers to these questions validate once again that this research is a very important step in improving patients’ scheduling, and that the healthcare industry is in need of such a tool. Table 37 shows the summary of the answers for question 6, and Table 38 shows the summary for question 7.

Table 37. Question 6 Answer Summary

<table>
<thead>
<tr>
<th>Answer Options</th>
<th>Not useful</th>
<th>Very Useful</th>
<th>Rating Average</th>
<th>Response Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Support Tool</td>
<td>4</td>
<td>5</td>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>4.00</td>
<td></td>
<td>69</td>
</tr>
</tbody>
</table>

Table 38. Question 7 Answers’ Summary

<table>
<thead>
<tr>
<th>Answer Options</th>
<th>Little Interest</th>
<th>Very Interested</th>
<th>Rating Average</th>
<th>Response Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interest</td>
<td>3</td>
<td>8</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>23</td>
<td>3.85</td>
<td>67</td>
</tr>
</tbody>
</table>
Questions eight and nine require information on the level of experience in healthcare and the current position in organization. Sixty two percent of respondents have more than fifteen years of experience in healthcare, about twenty three percent have between 6-14 years experience in healthcare, and about fifteen percent have 1-5 years experience in healthcare. From a total of sixty four respondents, about sixty six percent hold positions in the process improvement and about twenty three percent hold an administrative position. Figures 40 and 41 show the distribution of level of experience and the current position.

![Figure 40. Level of Experience of Respondents.](image-url)
Question 11 asked the respondents to estimate the number of patients seen every month by the providers in their organization. From the total number of respondents, 28 (47.5%) answered that less than 1,000 patients are seen by their organization every month, and 39 (52.5%) answered that more than 1,000 patients are seen every month. Two organizations see between 10,000-25,000 patients every month, ten see 5,000-10,000 patients per month, and thirteen see between 1,000-5,000 patients each month. The average number of patients seen every month is about 10,000, and the median is 1,200. Figure 42 shows the answers.
The last question in the survey asked the respondents to estimate current no-show rates in their organization. About seven percent responded that the no-show rate is more than 25%, about 26.1% have a no-show rate between 15-25%, while 39.1% have a no-show rate between 6-15%. About 23% of the respondents answered that they don’t know what the no-show rate in their organization is. Figure 43 shows the distribution of no-show rates among the respondents.
An in depth analysis of the survey answers was performed. The results are presented in the following table. Table 39 shows the distribution of “satisfied”/”unsatisfied” for the human scheduler method and human scheduler with algorithm based on the level of experience in healthcare. It can be seen that the majority of respondents (61.4%) have more than 15 years of experience in healthcare from which 35 (81.4%) use a human scheduler and only eight (18.6%) use a human scheduler with algorithm. In the group of human scheduler, 10 answered that they are not satisfied with the current scheduling method, 11 answered they are satisfied, and 14 had a neutral answer. In the group of human scheduler with algorithm, 5 (62.5%) answered that they
are very satisfied with the current scheduling method while 3 answered that they are unsatisfied or neutral.

Table 39. Distribution of Scheduling Method, Level of Satisfaction and Level of Experience

<table>
<thead>
<tr>
<th>Level of Experience</th>
<th>How satisfied are you with current method of scheduling?</th>
<th>Human Scheduler</th>
<th>Human scheduler and Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td></td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>1-5.</td>
<td></td>
<td>-</td>
<td>2</td>
</tr>
<tr>
<td>6-9.</td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>10-14.</td>
<td></td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&gt;15</td>
<td></td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 40 shows the distribution of no-show rates based on the level of experience in healthcare. The value of “0” on the column of no-show rate is allocated to “Don’t know” answer, while the “-“ symbol represents no answer for the specific group. It can be seen that the majority of the respondents have more than 15 years of experience in healthcare. Nineteen “Don’t know” answers were recorded in the groups of “no experience” (1 answer), “1-5” (4 answers), 6-9” (4 answers), and “More than 15” (9 answers) years experience in healthcare. In the group with the longest experience, 5 respondents work in the clinical area or other areas of the organization, while 4 respondents work in process improvement.
Table 40. No-show Rate and Level and Experience

<table>
<thead>
<tr>
<th>Level of Experience</th>
<th>0</th>
<th>0-5</th>
<th>6-10.</th>
<th>11-15.</th>
<th>16-20</th>
<th>21-25</th>
<th>&gt;25</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1-5.</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>6-9.</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>-</td>
<td>1</td>
</tr>
<tr>
<td>10-14.</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>&gt;15</td>
<td>9</td>
<td>8</td>
<td>9</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
CHAPTER 5: DISCUSSION

This research developed a decision-support tool to assist healthcare organizations in forecasting no-show appointments. It is based on data mining techniques used to classify no-show appointments using historical data, and to create forecasting no-show rules that can be used in scheduling systems. In this chapter we will discuss the benefits of the methodology, its challenges and limitations.

The following hypotheses were considered:

**Hypothesis 1:** *No show rates are statistically different from those groups identified by personal characteristics.* The analysis was performed between groups belonging to the same variables (i.e., different race, different age, and different insurance type).

**Hypothesis 2:** *No show rates are significantly correlated to personal characteristics of patients.* The second hypothesis verified the correlation between patient characteristics and no-show rates.

**Hypothesis 3:** *No show rates are significantly correlated to appointment characteristics.* The third hypothesis verified the correlation between appointment characteristics (i.e., day and time of appointment) and no-show rates.

The pre-processing step of the data showed that there are a large number of missing data. In most statistical analysis, the missing data is excluded and only complete data is used, resulting in not so accurate results. Statistical analysis does not work very well with missing data, so we decided not to perform any statistical analysis (i.e., ANOVA, logistic regression), but to use neural networks to analyze the data.
5.1. Benefits and Advantages of the Decision Support Tool

This study developed an integrated methodology of forecasting no-show appointments in the healthcare system. It collected and analyzed a large database from a large hospital in the Midwest. This research develops one of the first methodologies that integrates data mining techniques (neural networks), forecasting and patients’ scheduling. Data mining is the process of discovering hidden information in data (Feyen and Huglin, 2010). It is based on using historical data to create decision rules that give a probability of no-show for an individual patient. These probabilities can be used to schedule each patient in such a way that considers patient preference and develops the most successful schedule.

Some of the advantages of the decision-support tool based on data mining techniques are as follows:

2. *Better Exploitation of Information* – by forecasting/predicting, and recognition of patterns, better decisions, and more targeting can be made.

3. *Works with Very Large Data Sets* – Healthcare organizations can generate large amounts of data about resources, patients, treatments, diseases, and more. These large amounts of data are the key source for knowledge extraction. The methodology presented in this research is capable of working with huge databases. The database included in this study consisted of more than 80,000 appointments over a one-year period. The data mining techniques included in this research, neural networks and...
decision trees, were used to classify no-show appointments and to discover patterns in the no-show appointments. Neural networks classified the no-show appointment with more than 90% accuracy. The decision trees discovered patterns (rules) in the database that are used to assign probabilities to individual subjects based on historical data. The rules were used in the scheduling system with the final goal of improving patients’ waiting time, number of patients seen every day, and doctors’ idle time.

4. *Uses Structured Techniques* – The decision-support tool, through the neural networks and decision trees, can take a huge amount of data, extract knowledge and create structures that can be easily understood by the human experts. In this study, the structures are represented by the decision tree rules. Rules were developed for appointments in primary clinic 2. The dataset of appointments for primary clinic 2 consists of about 32,000 appointments. Using decision trees, 35 rules were created that are used to assign individual probabilities.

5. *Updates with New Data* - The neural networks can learn from data and create patterns. They have the capability of updating every time new data is included in the dataset. The update has to be performed by the data mining experts; neural networks do not update automatically when new data is included.

6. *Includes a Large Number of Variables* - The decision-support tool based on neural networks and decision trees works very well with a large number of variables. Neural networks can combine many predictors without over-fitting, while decision
trees work very well on selecting the variables that are most important when making
predictions. For the database used in this research, each appointment had about 22
variables. For each appointment, 6 more variables were added, making a total of 28
variables for each appointment.

7. *Uses Non-Linear Algorithms* – Both neural networks and decision trees are able to
detect nonlinear relationships automatically.

5.2 Limitations of the Decision Support Tool

1. *Missing Data* - Missing values, noise, and outliers can post a great challenge in
analyzing a dataset. Since the majority of data is collected from patients and
providers, it is likely that the information can be misplaced, incorrectly copied,
omitted, or not collected. In this research, we had to deal with missing data for all
clinics. For two clinics, some missing data could be replaced by using previous
recorded data for a specific patient. For two clinics, missing data was such a large
percent of the dataset that we had to remove the information and not include it in the
analysis.

2. *Heterogeneous Sources* – Patient information can be collected from different
sources. Information can collected from the patient himself (name, age, gender, race,
etc), from provider (diagnosis, treatment, therapy, etc), pharmacy (drug), or outside
sources (insurance). This information can be collected in different formats and most
importantly, it can come in non-mathematical form. It is very important that the
information is collected into a standardized form. It is a characteristic of medical

data to be expressed in words or images that cannot be written in equations or

formulas (Cios and Moore, 2002). The majority of data used for this study came in

non-mathematical form. Since the software used in analysis do not work with non-

mathematical forms, all data had to be coded or transformed into a form that was

compatible with the software required form. For neural network analysis, data were

modified into a binary form that includes only “0” and “1” and creates unique

combinations for each variable. For SPSS analysis, data was coded by assigning an

individual code to each variable.

3. Communication and Cooperation with Patient – It is imperative for data mining to

collect as much information as possible from all sources. One source, the patient, can

be of issue. Some patients choose to fill out all forms completely, while others

choose to fill out only the basic information such as name, address, date of birth, and

social security. Communicating to patient the importance of giving all the

information necessary is very important. When the information is collected

electronically, this problem may be easier to solve by making all fields mandatory to

complete. This way there is some assurance that patients will fill out the form. But

when the information is collected through hardcopies, the healthcare staff has to

manually check and ask the patient for the missing information. The whole

procedure can be tedious and time consuming, and likely to not have the desired

results.
4. **Ethical, Legal and Social Constraints** – One of the biggest obstacles in data mining is the privacy of records (Canlas, 2009). Patients’ data are real, private data. They are collected on human subjects, and there is a huge ethical and legal practice developed to prevent the abuse and misuse of patients’ data (Cios and Moore, 2002). The major concerns with the use of patients’ data are the privacy and security of human data, fear of lawsuits, data ownership, expected benefits, and administrative issues (Cios and Moore, 2002). Patients go to the doctor with full confidence that their discussion and the results of it are completely confidential. It is very important that this relationship maintains the same status. Patients have to feel comfortable in discussing their condition, whatever that may be, with their provider. The data used in this study were 100% anonymous. For security purposes, the patients’ identification number was replaced with randomly generated numbers before the data was disseminated. The data did not include names, social security numbers, or any other identification information. However, an IRB approval was obtained from the hospital from where the data were obtained, and from the Ohio University Office of Research.

5. **Healthcare Experts and Data Mining Specialists Work Together** – Successful application of data mining in the healthcare industry requires knowledge of the industry as well as knowledge of data mining methodologies and techniques (Koh and Tan, 2005). It is very important that the data mining experts work very closely with the healthcare experts. It is recommended that the healthcare organization
should develop their data mining applications that may require some considerable investment of resources.

6. The results of this research are for one specific children hospital’s four clinics. The extent that these results are applicable to other hospitals is not known.

7. The majority of patients have government insurance. It is unknown if the hospital is focused on serving patients that are covered only by government insurance or not.

8. The survey used voluntary responses with 3.5% response rate. Voluntary responses have the limitation that only respondents with possible interests may answer the questions.
6.1. Conclusions

An accurate decision support tool to forecast no-show appointments in healthcare was developed. Doctors’ utilization time reached 90.4%, exceeding the anticipated improving rate of 10%. Average patients’ waiting time to see a doctor reduced from the initial average waiting time of 0.46 hours (27.6 minutes) to 0.34 hours (20.4 minutes). The number of patients seen in one day increased from 85 to 101, an increase of about 18.82%. However, the total maximum waiting time increased from 2.08 hours to 2.58 hours, and the total average waiting time increased from 0.78 hours to 0.89 hours.

Major contributions were made to the understanding of no-show forecast.

- A decision support tool using data mining techniques was developed. It is offered as a candidate benchmark model to forecast individual no-show appointments and include them in the scheduling systems.
- The simulation demonstrated that improvements in the scheduling system can be made by using individual no-show appointment probabilities.

A total of 87,000 appointments was collected and included in the study, one of the largest sets of data in the literature. Conducting very comprehensive experiments revealed new knowledge in patient behavior. Trends collected from the experiments provide guidance in determining the variables that influence no-show appointments in healthcare. Based on the variables included in the analysis, neural networks could classify no-show
appointment with more than 90% accuracy. These results are considered to be very 
accurate, since in the world of probabilities the range of 0.9-1.00 is considered to be 
“very likely”. The decision trees developed rules that are used to assign individual no-
show probabilities to each patient. This is a great improvement since the methodologies 
available in the literature offer only general no-show probabilities based on the no-show 
rate.

The decision trees selected only the variables that have a great influence on no-
show behavior. The variables that have the greatest occurrence in the rules developed are 
variables related to patient such as gender, race, poverty level, insurance type, and 
father’s employment; appointment such as appointment day and type, provider type and 
name; and environment such as temperature, weather events and precipitation. A higher 
probability of no-show is seen in patients without insurance or have a combination of two 
or more types of insurances. A temperature lower than 37F or greater than 77F is 
correlated to higher probabilities of no-show. Father’s employment status other than 
employed is also correlated to higher no-show probabilities. A poverty level less than 
12.8% is correlated to lower probabilities of no-show, while a poverty level greater than 
12.8% is correlated to higher probabilities. Weather events such rain, snow, and fog are 
correlated to relatively higher probabilities.

The simulations performed in this study showed that by using individual no-show 
probabilities, the scheduling system can be improved tremendously. The number of 
patients seen every day increased from 85 in 9.92 hours (0.12 hours per patient) to 98 in
9.58 hours (0.10 hours per patient). If this rate can be sustained every day or improved, a healthcare organization of similar size can see an increase of 3,380 patients/year or an increase in revenue of about $500,000/year. In addition to the increase in the number of patients seen every day, doctor’s utilization time increased from 71.40% to 87%. This represents an improvement of about 22% in the idle time. The utilization time was considered the time doctors spend with patients. From the total numbers of hours worked every day, about 1 hour should be left for writing reports, charts, or other administrative tasks. However, it is the decision of each healthcare organization on what parameters they would like to improve through scheduling. One may have to pay attention to how many hours doctors and staff work overtime, or if staff schedule can be developed in such a way that with a minimum number required we can achieve a maximum utilization time (i.e., one nurse may work from 8-5.00 PM, the second may work from 9-6.00 PM).

The survey sent to the experts in the field of healthcare helped to better understand the status quo of the healthcare scheduling system. From the responders, 77.1% answered that the scheduling is performed by a human scheduler without any help from an algorithm, while 25.7% use a human scheduler helped by an algorithm. An overwhelmingly large percent do not use technology to better fit the patient to the scheduling system. This is backed up by the variables that are taken into account when performing scheduling. The most common variables taken into account are physician/slot availability, patient preference, and urgency of medical issue. It can be seen that there is no analysis of possible no-show appointments. When asked “On a scale from 0 to 5, how
good is the methodology/tool you are currently using?”, only 19 respondents (27.5%) answered that they are satisfied or very satisfied with the scheduling methodology they use. The majority of respondents (72.5%) answered that they are less than satisfied. While a human scheduler methodology may work for small practices, the more complex healthcare organization may have a great challenge in performing proper scheduling.

The survey also asked the healthcare experts, “How useful do you believe a decision support tool that predicts patient no-show would be in healthcare patient scheduling?” and “Gauge your interest level in investigating the possibility of adopting such a decision support tool.” A large majority (76.8%) answered that the decision support tool that predicts patient no-show can be very useful or useful, and 71.6% answered that they would be interested in adopting such a tool. These answers prove that the decision support tool developed in this research is a needed tool in healthcare scheduling systems, and can be a very powerful tool in improving the efficiency of healthcare organizations. About 62% of the respondents have more than 15 years of experience in the healthcare industry and about 66% work in process improvement departments. One may believe that the respondents’ years of experience is once again a reason to consider the necessity of the decision support tool developed here.

6.2. Further Improvements for the Decision Support Tool

Using the limitations described in the previous section some recommendations for improvement can be made:
- The decision support tool was developed and tested using the database from one hospital only and it was analyzed for individual clinics. When applied to practice, the tool can be improved by combining the data from all clinics of the same hospital in such a way that more general rules can be created to be applied to all clinics in a network. This will reduce the analysis cost and improve the response time.

- Ensure that data coming from different sources is in standard formats. By implementing a standard format for the data collected, the pre-processing time is reduced dramatically. It is preferred that the standard form is a mathematical form that is compatible with the software used by the decision support tool.

- The decision support tool can be further improved by inclusion of text mining. The data currently used is quantitative data. But a large part of data in the healthcare industry is text (physician’s notes) or image (i.e., X-ray, ultrasound results).

Although the results are valid for one hospital, this research demonstrated the feasibility of large process improvements through innovative modelling techniques. It can be a candidate benchmark model to forecast individual no-show appointments and include them in the scheduling systems.
REFERENCES


   [http://www.generation5.org/content/1999/perceptron.asp](http://www.generation5.org/content/1999/perceptron.asp).


74. NeuroAI – Artificial Neural Network 2011. 
   Http://www.learnartificialneuralnetworks.com/


76. Organization for Economic Co-operation and Development. 
   http://www.oecd.org/home/0,3305,en_2649_201185_1_1_1_1_1,00.html


98. U.S. Census Bureau 2011b. [http://www.census.gov/support/USACdataDownloads.html#EDU](http://www.census.gov/support/USACdataDownloads.html#EDU)


# APPENDIX 1: DESCRIPTION OF EVIDENCE FOR GROUP 1

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chakraborty et al. (2010)</td>
<td><strong>Intervention</strong>&lt;br&gt;Use of dynamic programming for scheduling including patient no-shows and general service time distributions.&lt;br&gt;Maximize the expected revenue for patients seen minus costs for patient waiting and staff overtime.&lt;br&gt;<strong>Outcome</strong>&lt;br&gt;Intervention&lt;br&gt;Hospital patients&lt;br&gt;It includes arbitrary service time distributions.&lt;br&gt;<strong>Study design</strong>&lt;br&gt;The objective function is unimodal for general service time and the unimodality is <em>independent</em> of the type of the service time distribution.&lt;br&gt;<strong>Study population</strong>&lt;br&gt;<strong>Methods</strong>&lt;br&gt;It develops a special case of gamma service times which requires significantly less computation.&lt;br&gt;It shows how the computational needs can be reduced significantly when service times are approximated by a gamma distribution.&lt;br&gt;<strong>Main results</strong>&lt;br&gt;The value of the maximum profit decreases with increasing variance of the service time.</td>
</tr>
<tr>
<td>Source</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Chien et al., 2008</td>
<td><strong>Intervention</strong></td>
</tr>
<tr>
<td></td>
<td>Use of a genetic algorithm (GA) to solve a hybrid shop scheduling problem to increase service quality.</td>
</tr>
<tr>
<td></td>
<td>Mixed integer programming model for validation purposes.</td>
</tr>
<tr>
<td></td>
<td>Reduce patient waiting time and improve operation efficiency.</td>
</tr>
<tr>
<td></td>
<td><strong>Intervention</strong></td>
</tr>
<tr>
<td></td>
<td>Patients in a physical therapy facility</td>
</tr>
<tr>
<td></td>
<td>Patients are considered jobs and medical resources machines.</td>
</tr>
<tr>
<td></td>
<td>Considered medical resources as identical and set in parallel.</td>
</tr>
<tr>
<td></td>
<td>GA is employed for sequencing module to establish new search points in historical information based on following steps:</td>
</tr>
<tr>
<td></td>
<td>• A set of initial solutions is encoded as a set of chromosomes called population.</td>
</tr>
<tr>
<td></td>
<td>• Crossover and mutation are applied to the population to generate offspring from parents.</td>
</tr>
<tr>
<td></td>
<td>• An evaluation and scaling function is used to assign fitness a value to each chromosome.</td>
</tr>
<tr>
<td></td>
<td>[ f(k) = w_w \cdot W_{max}(k) + w_c \cdot C_{max}(k) + P(k) ]</td>
</tr>
<tr>
<td></td>
<td>Where: ( W_{max}(k) ) is maximum waiting time for chromosome K</td>
</tr>
<tr>
<td></td>
<td>( C_{max}(k) ) is the makespan of chromosome k</td>
</tr>
<tr>
<td></td>
<td>( P(k) ) is a penalty function of chromosome k</td>
</tr>
<tr>
<td></td>
<td>• Selection mechanism is employed to improve the solution quality.</td>
</tr>
<tr>
<td></td>
<td>• The roulette wheel approach is used for population selection.</td>
</tr>
<tr>
<td></td>
<td>• To validate the solution, a mixed integer programming is used to provide the optimal solution.</td>
</tr>
<tr>
<td></td>
<td>• Objectives: minimize maximum waiting time and minimize the makespan.</td>
</tr>
<tr>
<td></td>
<td>[ \text{Minimize} \quad w_w \cdot W_{max} + w_c \cdot C_{max} ]</td>
</tr>
<tr>
<td></td>
<td>• Both methods reveal identical waiting time in both problems.</td>
</tr>
<tr>
<td></td>
<td>• Average computation time for MIP is:</td>
</tr>
<tr>
<td></td>
<td>• 1.405 s (StDev=1.916s) for 10 patients problem</td>
</tr>
<tr>
<td></td>
<td>• 4496.505s (StDev=889.102s) for 20 patients problem</td>
</tr>
<tr>
<td></td>
<td>• Average computation time for GA is less than 10s in both problems.</td>
</tr>
<tr>
<td>Source</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| Ogulata et al., 2008 | - Development and testing of a hierarchical mathematical model for patient and staff scheduling  
- Maximize number of patients seen in a day, decrease waiting time for patients, and achieve fair distribution of patients among physiotherapist.  
- Intervention  
- Four therapists are available.  
- The problem was broken into three stages:  
  - 1st stage: weekly patient selection is based on priority (high, medium, and low) and availability of physiotherapist.  
  - 2nd stage: assignment to physiotherapist.  
  - 3rd stage: patient scheduling. Each patient is assigned to one time intervals in a working day.  
- Using stage 1 model, 54 patients were selected based on their priority. From these, 44% have high priority, 49% medium, and 7% low.  
- In stage two, the patients are assigned to a physiotherapist as follows: 14 patients to 1st therapist, 14 to 2nd, 13 to 3rd, and 13 to the last one.  
- The proposed algorithm scheduled 6 more patients than the scheduling method used in the hospital. The proposed algorithm scheduled 24 patients with high priority compared with 15 scheduled with the existing method.  
- The proposed algorithm decreased patient waiting time.  
- Limitations: the algorithm does not take into account patient preference for day of the week. |
<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaandorp and Koole, 2007</td>
<td><strong>Intervention</strong>&lt;br&gt;A local search procedure in patient scheduling&lt;br&gt;<strong>Outcome</strong>&lt;br&gt;Weighted average of expected waiting times of patients;&lt;br&gt;Idle time of the doctor;&lt;br&gt;Tardiness (lateness)&lt;br&gt;<strong>Study design</strong>&lt;br&gt;N/A&lt;br&gt;<strong>Study population</strong>&lt;br&gt;N/A&lt;br&gt;<strong>Methods</strong>&lt;br&gt;The formulas to calculate mean waiting time $W(x)$, idle time $I(x)$, and tardiness $L(x)$ are developed. The objective function becomes:&lt;br&gt;$$MIN \quad \alpha_W W(x) + \alpha_I I(x) + \alpha_L L(x)$$&lt;br&gt;Where: $\alpha_W$, $\alpha_I$, and $\alpha_L$ are the weights for mean waiting time, idle time, and tardiness.&lt;br&gt;<strong>Main Results</strong>&lt;br&gt;Since the number of solution is too big, a neighborhood is defined.&lt;br&gt;A local search is started to find the most feasible solution in the neighborhood. A local minimum is found.&lt;br&gt;The authors prove that the local minimum is the global minimum.&lt;br&gt;The authors include an equal no-show probability for each patient.&lt;br&gt;The results of the proposed model give a better solution than the two existing schedules.&lt;br&gt;The variation in no-show probability influences the scheduling. For a higher probability, the mean waiting time, idle time and tardiness become larger.</td>
</tr>
<tr>
<td>Source</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
</tbody>
</table>
| Isken and Rajagopalan, 2002 | **Intervention**  
**Outcome**  
**Study design**  
**Study population**  
**Methods**  

- Use of data-mining technique – K-means cluster analysis – to support a process of building computer simulation models.  
- Effectiveness of data preprocessing and clustering methods.  
- N/A  
- 8261 patient visit records from several months in 1996.  
- A master data table was created. It involves selection of patient based on the importance of certain variables and cases.  
- The clustering problem was defined: to partition these pattern vectors into a number of clusters, such that each of them has at least one cluster member, are pair by pair disjoint and are mutually exhaustive. Euclidean distance function was used for two vectors.  
- The most important variable is Diagnosis Related Groups (DRG).  
- K-means algorithm is used to find k optimal clusters in the data set. The function used is:  
\[
MIN \ j = \sum_{k=1}^{K} (\sum_{j \in P_k} |x_j - m_j|)
\]

Where \(x_j\) is a dimensional pattern vector.

- For cluster selection, a measure named cluster purity was used. It uses the number of distinct DRGs appearing in a cluster.

- Data preparation affects positively the quality of solutions.

- The K-means algorithm completed the task in 3-5s (when implemented in ClustanGraphics 5)
<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Podgorelec and Kokol, 1997</td>
<td><strong>Intervention</strong>&lt;br&gt;• Scheduling algorithm based on genetic algorithm and machine learning.&lt;br&gt;• Optimal schedule that can result in finishing all the activities scheduled as soon as possible, with the least patient waiting time and maximum device utilization (lowest overall duration).&lt;br&gt;• N/A&lt;br&gt;• 22 patients and five therapeutic devices were used.&lt;br&gt;• For every given problem, random test problems were generated and executed.&lt;br&gt;• A number of random schedules were generated.&lt;br&gt;• Each individual was evaluated for the fitness score. The individuals with the best score were selected.&lt;br&gt;• The crossover procedure was applied until new individuals fulfill the complete population. Each time an individual is added, another individual (with the lowest fitness score) is eliminated.&lt;br&gt;• The mutation operator is applied with some probability.&lt;br&gt;• All phases are repeated until an acceptable solution evolved.&lt;br&gt;• A possible solution evolved after 165 generations. The duration of all therapies was 335 minutes.&lt;br&gt;• Idle time for 3 devices is 10, 5, and 30 minutes respectively, without affecting the total time.&lt;br&gt;• The maximum overall waiting time is 15 minutes for one patient&lt;br&gt;• The overall average waiting time is 5 minutes.</td>
</tr>
</tbody>
</table>
## APPENDIX 2: DESCRIPTION OF EVIDENCE FOR GROUP 2

<table>
<thead>
<tr>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeGrano et al. 2009</td>
<td><strong>Intervention</strong>&lt;br&gt;Use of auction and optimization model in nurse scheduling to accommodate individual preferences.&lt;br&gt;Minimize the time allocated to nurse scheduling&lt;br&gt;28 full-time nurses with 36h/week, 6 full-time with 40h/week, six part-time working between 20-28 h/week, 28 nurses working occasionally, 5 traveler nurses working 36h/week.&lt;br&gt;A weight is assigned for each shift (e.g., 1 for a 4-h shift, 2 for 8-h shift, 3 for a 12-h shift, and 6 for day-off). These weights are converted to bid points.&lt;br&gt;Each nurse submits a preferred scheduled. Points are assigned to the schedule. Maximum number of points is 400.&lt;br&gt;Each nurse submits a bid.&lt;br&gt;Two optimization models are used after the bid:&lt;br&gt;• For the award step, the objective function maximizes the point value of bids awarded to the candidate winners.&lt;br&gt;• For schedule completion, the objective function awards bids which were not selected as candidate winners but can create a feasible assignment.&lt;br&gt;It took 2.073s to determine the candidate winner and generate the formulation. For award stage it took 2 minutes and 55 s. To complete the schedule it took 5.74s.&lt;br&gt;Overall it took about 3 minutes to generate the schedule, which is much faster than the manual method.&lt;br&gt;The auction-optimization approach can account for both, the nurse preferences and the hospital constraints, and generate a good schedule.&lt;br&gt;It fulfilled 98.27% of “on” requests and 95.51% of “off” requests.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study population</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Methods</td>
<td></td>
</tr>
<tr>
<td>Main Results</td>
<td></td>
</tr>
<tr>
<td>Source</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Beaulieu et al., 2000</td>
<td><strong>Intervention</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Outcome</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Study population</strong></td>
</tr>
<tr>
<td></td>
<td><strong>Methods</strong></td>
</tr>
<tr>
<td></td>
<td>• Use of a mathematical model to facilitate preparing a schedule for physicians in emergency room.</td>
</tr>
<tr>
<td></td>
<td>• Reduction in time and effort required to develop a six-month schedule.</td>
</tr>
<tr>
<td></td>
<td>• 20 physicians, including 15 working full-time.</td>
</tr>
<tr>
<td></td>
<td>• The model is formulated as a single-objective optimization model, which seeks to minimize a weighted sum of all deviations.</td>
</tr>
<tr>
<td></td>
<td>• Four groups of constraints are defined:</td>
</tr>
<tr>
<td></td>
<td>• Compulsory constraints (e.g., one physician must be assigned to one shift at the time);</td>
</tr>
<tr>
<td></td>
<td>• Ergonomic constraints (e.g., limits on number of weekly hours of certain types of shifts);</td>
</tr>
<tr>
<td></td>
<td>• Distribution constraints (e.g., seniority), and</td>
</tr>
<tr>
<td></td>
<td>• Goal constraints (e.g., number of worked hours per week).</td>
</tr>
<tr>
<td></td>
<td>• Two programs are developed:</td>
</tr>
<tr>
<td></td>
<td>• One to generate the model in a format accessible to the branch-and-bound software, by reading input file;</td>
</tr>
<tr>
<td></td>
<td>• One to read the solution, create an output file and to identify violations of the ergonomic rules.</td>
</tr>
<tr>
<td></td>
<td>• This approach can take into account more rules than any human expert.</td>
</tr>
<tr>
<td></td>
<td>• It violated ergonomic rules 40% less than the human expert (111 compared with 185 violations by human expert).</td>
</tr>
<tr>
<td></td>
<td>• It is faster than a manual method.</td>
</tr>
<tr>
<td>Source</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------------</td>
</tr>
<tr>
<td>Aickelin and Dowsland, 2004</td>
<td>Use of an indirect genetic algorithm method and a heuristic method in nurse scheduling.</td>
</tr>
<tr>
<td></td>
<td>52 real hospital datasets</td>
</tr>
<tr>
<td></td>
<td>The problem is formulated as an integer linear programming.</td>
</tr>
<tr>
<td></td>
<td>The objective function is:</td>
</tr>
<tr>
<td></td>
<td>$\text{MIN } \sum_{i=1}^{n} \sum_{j \in F(i)}^{m} p_{ij} x_{ij}$</td>
</tr>
<tr>
<td></td>
<td>Where: $x_{ij}$ is decision variables, and equals “1” if nurse $i$ works shift pattern $j$, and “0” else; $p_{ij}$ is preference nurse $n$ is number of nurses, $m$ is number of shift patterns.</td>
</tr>
<tr>
<td></td>
<td>The genetic algorithm tries to find the best possible ordering of the nurses, and the decoder builds the actual solution.</td>
</tr>
<tr>
<td></td>
<td>The decoder calculates the fitness function.</td>
</tr>
<tr>
<td></td>
<td>$\text{MIN } \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} x_{ij}$</td>
</tr>
<tr>
<td></td>
<td>$+ w_{\text{demand}} \sum_{k=1}^{14} \sum_{s=1}^{p} \max \left[ R_{ks} \right.$</td>
</tr>
<tr>
<td></td>
<td>$- \sum_{l=1}^{n} \sum_{j=1}^{m} q_{ls} a_{jk} x_{ij}; 0 \left. \right]$</td>
</tr>
<tr>
<td></td>
<td>Where: $R_{ks}$ demand of nurses, $a_{jk}$ and $q_{ls}$ are decision variables, and $w_{\text{demand}}$ is a penalty weight.</td>
</tr>
<tr>
<td></td>
<td>51 out of 52 of datasets are solved to or near to optimality and a feasible solution is found.</td>
</tr>
<tr>
<td></td>
<td>Indirect GA approach showed to be more flexible and robust than tabu search.</td>
</tr>
</tbody>
</table>
APPENDIX 3: VBA MODULE TO POPULATE MISSING DATA

Sub Populate()
    ' Macro2 Macro
    ' Dim ws As Worksheet
    Dim sourcePid

    Set ws = Excel.ActiveSheet

    For i = 2 To ws.Rows.Count
        pid = ws.Range("I" & i).Value
        bdy = ws.Range("J" & i).Value
        If pid <> "" Then
            If bdy = "" Then
                If pid = sourcePid Then
                    ws.Range("J" & i & ":S" & i).Select
                ws.Paste
                End If
            Else
                ws.Range("J" & i & ":S" & i).Select
                Selection.Copy
                sourcePid = pid
            End If
        Else
            Exit Sub
        End If
    Next i
End Sub
APPENDIX 4: RULES DERIVED FROM DECISION TREES FOR PRIMARY CLINIC

/* Node 79 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4
AND VALUE(Events) NE 12) AND (SYSMIS
(SAppointmentTime) OR VALUE(SAppointmentTime) NE 5 AND
VALUE(SAppointmentTime) NE 6 AND VALUE(SAppointmentTime) NE 1) AND
(SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND
(SYSMIS(Precipitation) OR (VALUE(Precipitation) LE 0))
AND VALUE(PovertyLevel) LE 12.8).
COMPUTE nod_001 = 79.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.972222.
END IF.
EXECUTE.

/* Node 108 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4
AND VALUE(Events) NE 12) AND (SYSMIS
(SAppointmentTime) OR VALUE(SAppointmentTime) NE 5 AND
VALUE(SAppointmentTime) NE 6 AND VALUE(SAppointmentTime) NE 1) AND
(SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND
(SYSMIS(Precipitation) OR (VALUE(Precipitation) LE 0))
AND (SYSMIS(PovertyLevel) OR (VALUE(PovertyLevel) GT 12.8)) AND
(SYSMIS(SProviderType) OR (VALUE(SProviderType) LE 411)) AND
(VALUE(Distance) LE 2.2177).
COMPUTE nod_001 = 108.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.970771.
END IF.
EXECUTE.

/* Node 112 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4
AND VALUE(Events) NE 12) AND (SYSMIS
(SAppointmentTime) OR VALUE(SAppointmentTime) NE 5 AND
VALUE(SAppointmentTime) NE 6 AND VALUE(SAppointmentTime) NE 1) AND
(SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND
(SYSMIS(Precipitation) OR (VALUE(Precipitation) LE 0))
AND (SYSMIS(PovertyLevel) OR (VALUE(PovertyLevel) GT 12.8)) AND
(SYSMIS(SProviderType) OR (VALUE(SProviderType) LE 411)) AND
(VALUE(Distance) LE 2.2177).

/* Node 108 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4
AND VALUE(Events) NE 12) AND (SYSMIS
(SAppointmentTime) OR VALUE(SAppointmentTime) NE 5 AND
VALUE(SAppointmentTime) NE 6 AND VALUE(SAppointmentTime) NE 1) AND
(SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND
(SYSMIS(Precipitation) OR (VALUE(Precipitation) LE 0))
AND (SYSMIS(PovertyLevel) OR (VALUE(PovertyLevel) GT 12.8)) AND
(SYSMIS(SProviderType) OR (VALUE(SProviderType) LE 411)) AND
(VALUE(Distance) LE 2.2177).
COMPUTE nod_001 = 112.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.971771.
END IF.
EXECUTE.
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4 AND VALUE(Events) NE 12) AND (SYSMIS(SAppointmentTime) OR VALUE(SAppointmentTime) NE 5 AND VALUE(SAppointmentTime) NE 6 AND VALUE(SAppointmentTime) NE 1) AND (SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND (SYSMIS(Precipitation) OR (VALUE(Precipitation) LE 0)) AND (SYSMIS(PovertyLevel) OR (VALUE(PovertyLevel) GT 12.8)) AND (SYSMIS(SProviderType) OR (VALUE(SProviderType) LE 411)) AND (SYSMIS(Distance) OR (VALUE(Distance) GT 2.2177)) AND (SYSMIS(SAppointmentday) OR VALUE(SAppointmentday) NE 3).

END IF.

EXECUTE.

/* Node 114 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND (SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4 AND VALUE(Events) NE 12) AND (SYSMIS(SAppointmentTime) OR VALUE(SAppointmentTime) NE 5 AND VALUE(SAppointmentTime) NE 6 AND VALUE(SAppointmentTime) NE 1) AND (SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND (SYSMIS(Precipitation) OR (VALUE(Precipitation) LE 0)) AND (SYSMIS(PovertyLevel) OR (VALUE(PovertyLevel) GT 12.8)) AND (SYSMIS(SProviderType) OR (VALUE(SProviderType) LE 411)) AND (SYSMIS(Distance) OR (VALUE(Distance) GT 2.2177)) AND (VALUE(SAppointmentday) EQ 3) AND (SYSMIS(Averageincome) OR (VALUE(Averageincome) LE 55870)).

END IF.

EXECUTE.

/* Node 115 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND (SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4 AND VALUE(Events) NE 12) AND (SYSMIS (SAppointmentTime) OR VALUE(SAppointmentTime) NE 5 AND VALUE(SAppointmentTime) NE 6 AND VALUE(SAppointmentTime) NE 1) AND (SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND (SYSMIS(Precipitation) OR (VALUE(Precipitation) LE 0)) AND (SYSMIS(PovertyLevel) OR (VALUE(PovertyLevel) GT 12.8)) AND (SYSMIS(SProviderType) OR (VALUE(SProviderType) LE 411)) AND (SYSMIS(Distance) OR (VALUE(Distance) GT 2.2177)) AND (VALUE(SAppointmentday) EQ 3) AND (VALUE(Averageincome) GT 55870).

END IF.
EXECUTE.

/* Node 99 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND (SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4 AND VALUE(Events) NE 12) AND (SYSMIS (SAppointmentTime) OR VALUE(SAppointmentTime) NE 5 AND VALUE(SAppointmentTime) NE 6 AND VALUE(SAppointmentTime) NE 1) AND (SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND (SYSMIS(Precipitation) OR (VALUE(Precipitation) LE 0)) AND (SYSMIS(PovertyLevel) OR (VALUE(PovertyLevel) GT 12.8)) AND (VALUE(SProviderType) GT 411).

COMPUTE nod_001 = 99.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.974428.
END IF.
EXECUTE.

/* Node 55 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND (SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4 AND VALUE(Events) NE 12) AND (SYSMIS (SAppointmentTime) OR VALUE(SAppointmentTime) NE 5 AND VALUE(SAppointmentTime) NE 6 AND VALUE(SAppointmentTime) NE 1) AND
(SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND
(VALUE(Precipitation) GT 0).
COMPUTE nod_001 = 55.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.966391.
END IF.
EXECUTE.

/* Node 56 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4
AND VALUE(Events) NE 12) AND (SYSMIS
(SAppointmentTime) OR VALUE(SAppointmentTime) NE 5 AND
VALUE(SAppointmentTime) NE 6 AND VALUE(SAppointmentTime) NE 1) AND
(VALUE(SRacedescription) GT 102) AND (SYSMIS(SMothersEmployment) OR
(VALUE(SMothersEmployment) LE 2022)).
COMPUTE nod_001 = 56.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.987699.
END IF.
EXECUTE.

/* Node 57 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4
AND VALUE(Events) NE 12) AND (SYSMIS
(SAppointmentTime) OR VALUE(SAppointmentTime) NE 5 AND
VALUE(SAppointmentTime) NE 6 AND VALUE(SAppointmentTime) NE 1) AND
(VALUE(SRacedescription) GT 102) AND (VALUE(SMothersEmployment) GT
2022).
COMPUTE nod_001 = 57.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.984694.
END IF.
EXECUTE.
/* Node 100 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4 AND
VALUE(Events) NE 12) AND (VALUE
(SAppointmentTime) EQ 5 OR VALUE(SAppointmentTime) EQ 6) AND
(SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 13)) AND
(SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 301)) AND
(VALUE(Averageincome) LE 30090) AND (SYSMIS
(SAppointmentday) OR VALUE(SAppointmentday) NE 5).
COMPUTE nod_001 = 100.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.982922.
END IF.
EXECUTE.

/* Node 110 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4 AND
VALUE(Events) NE 12) AND (VALUE
(SAppointmentTime) EQ 5 OR VALUE(SAppointmentTime) EQ 6) AND
(SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 13)) AND
(SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 301)) AND
(VALUE(Averageincome) LE 30090) AND (VALUE
(SAppointmentday) EQ 5) AND (VALUE(Temperature) LE 50).
COMPUTE nod_001 = 110.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.870968.
END IF.
EXECUTE.

/* Node 111 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4 AND
VALUE(Events) NE 12) AND (VALUE
(SAppointmentTime) EQ 5 OR VALUE(SAppointmentTime) EQ 6) AND
(SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 13)) AND
(SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 301)) AND
(VALUE(Averageincome) LE 30090) AND (VALUE

(SAppointmentday) EQ 5) AND (SYSMIS(Temperature) OR (VALUE(Temperature) GT 50)).
COMPUTE nod_001 = 111.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.983193.
END IF.
EXECUTE.

/* Node 82 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND 
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND 
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4 
AND VALUE(Events) NE 12) AND (VALUE 
(SAppointmentTime) EQ 5 OR VALUE(SAppointmentTime) EQ 6) AND 
(SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 13)) AND 
(SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 301)) AND 
(SYSMIS(Averageincome) OR (VALUE(Averageincome) GT 30090) 
).
COMPUTE nod_001 = 82.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.986220.
END IF.
EXECUTE.

/* Node 102 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND 
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND 
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4 
AND VALUE(Events) NE 12) AND (VALUE 
(SAppointmentTime) EQ 5 OR VALUE(SAppointmentTime) EQ 6) AND 
(SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 13)) AND 
(VALUE(SPrimarypayor) GT 301) AND (SYSMIS(Temperature) OR 
(VALUE(Temperature) LE 66)) AND (SYSMIS(SProviderType) OR 
(VALUE(SProviderType) LE 411)).
COMPUTE nod_001 = 102.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.976471.
END IF.
EXECUTE.

/* Node 103 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4
AND VALUE(Events) NE 12) AND 
(SAPpointmentTime) EQ 5 OR VALUE(SAppointmentTime) EQ 6) AND
(SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 13)) AND
(VALUE(SPrimarypayor) GT 301) AND 
(SYSMIS(Temperature) OR 
(VALUE(Temperature) LE 66)) AND 
(VALUE(SProviderType) GT 411).
COMPUTE nod_001 = 103.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.978947.
END IF.
EXECUTE.

/* Node 84 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4
AND VALUE(Events) NE 12) AND 
(SAPpointmentTime) EQ 5 OR VALUE(SAppointmentTime) EQ 6) AND
(SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 13)) AND
(VALUE(SPrimarypayor) GT 301) AND 
(VALUE(Temperature) GT 66 AND VALUE(Temperature) LE 71).
COMPUTE nod_001 = 84.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.938272.
END IF.
EXECUTE.

/* Node 85 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4
AND VALUE(Events) NE 12) AND 
(SAPpointmentTime) EQ 5 OR VALUE(SAppointmentTime) EQ 6) AND
(SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 13)) AND
(VALUE(SPrimarypayor) GT 301) AND 
(VALUE(Temperature) GT 71).
COMPUTE nod_001 = 85.
COMPUTE pre_001 = 1.
COMPUTE prb_001
    = 0.968300.
END IF.
EXECUTE.

/* Node 32 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
    (SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
    VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4
    AND VALUE(Events) NE 12) AND (VALUE
    (SAppointmentTime) EQ 5 OR VALUE(SAppointmentTime) EQ 6) AND
    (VALUE(SAppointmentType) GT 13).
COMPUTE nod_001 = 32.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.949206.
END IF.
EXECUTE.

/* Node 11 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
    (SYSMIS(Events) OR VALUE(Events) NE 24 AND VALUE(Events) NE 1 AND
    VALUE(Events) NE 124 AND VALUE(Events) NE 13 AND VALUE(Events) NE 4
    AND VALUE(Events) NE 12) AND (VALUE
    (SAppointmentTime) EQ 1).
COMPUTE nod_001 = 11.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.970433.
END IF.
EXECUTE.

/* Node 12 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
    (VALUE(Events) EQ 24 OR VALUE(Events) EQ 124 OR VALUE(Events) EQ 4 OR
    VALUE(Events) EQ 12) AND (VALUE(Temperature) LE 66).
COMPUTE nod_001 = 12.
COMPUTE pre_001 = 1.
COMPUTE prb_001
    = 0.969014.
END IF.
EXECUTE.
/* Node 13 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(VALUE(Events) EQ 24 OR VALUE(Events) EQ 124 OR VALUE(Events) EQ 4 OR
VALUE(Events) EQ 12) AND (VALUE(Temperature) GT 66 AND
VALUE(Temperature) LE 71).
COMPUTE nod_001 =
  13.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.935185.
END IF.
EXECUTE.

/* Node 33 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(VALUE(Events) EQ 24 OR VALUE(Events) EQ 124 OR VALUE(Events) EQ 4 OR
VALUE(Events) EQ 12) AND (SYSMIS(Temperature) OR (VALUE(Temperature) GT
71 AND VALUE(Temperature) LE 80)) AND
(SYSMIS(Distance) OR (VALUE(Distance) LE 10.0844)).
COMPUTE nod_001 = 33.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.985138.
END IF.
EXECUTE.

/* Node 34 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(VALUE(Events) EQ 24 OR VALUE(Events) EQ 124 OR VALUE(Events) EQ 4 OR
VALUE(Events) EQ 12) AND (SYSMIS(Temperature) OR (VALUE(Temperature) GT
71 AND VALUE(Temperature) LE 80)) AND
(VALUE(Distance) GT 10.0844).
COMPUTE nod_001 = 34.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.929134.
END IF.
EXECUTE.

/* Node 15 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(VALUE(Events) EQ 24 OR VALUE(Events) EQ 124 OR VALUE(Events) EQ 4 OR
VALUE(Events) EQ 12) AND (VALUE(Temperature) GT 80 AND
VALUE(Temperature) LE 82).
COMPUTE nod_001 =
15.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.949495.
END IF.
EXECUTE.

/* Node 35 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(VALUE(Events) EQ 24 OR VALUE(Events) EQ 124 OR VALUE(Events) EQ 4 OR
VALUE(Events) EQ 12) AND (VALUE(Temperature) GT 82) AND
(SYSMIS(SRacedescription) OR (VALUE(SRacedescription)
LE 103)).
COMPUTE nod_001 = 35.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.971264.
END IF.
EXECUTE.

/* Node 36 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(VALUE(Events) EQ 24 OR VALUE(Events) EQ 124 OR VALUE(Events) EQ 4 OR
VALUE(Events) EQ 12) AND (VALUE(Temperature) GT 82) AND
(VALUE(SRacedescription) GT 103).
COMPUTE nod_001 =
36.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.920354.
END IF.
EXECUTE.

/* Node 37 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(VALUE(Events) EQ 1 OR VALUE(Events) EQ 13) AND (VALUE(Temperature) LE
50) AND (VALUE(SAppointmentday) EQ 6).
COMPUTE nod_001 = 37.
COMPUTE pre_001 = 1.
COMPUTE prb_001 =
0.891566.
END IF.
EXECUTE.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(VALUE(Events) EQ 1 OR VALUE(Events) EQ 13) AND (VALUE(Temperature) LE
50) AND (SYSMIS(SAppointmentday) OR VALUE(SAppointmentday) NE 6) AND
(VALUE(SFathersEmployment) LE 2011).
COMPUTE nod_001 = 60.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 1.000000.
END IF.
EXECUTE.

DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(VALUE(Events) EQ 1 OR VALUE(Events) EQ 13) AND (VALUE(Temperature) LE
50) AND (SYSMIS(SAppointmentday) OR VALUE(SAppointmentday) NE 6) AND
(SYSMIS(SFathersEmployment) OR (VALUE
(SFathersEmployment) GT 2011)).
COMPUTE nod_001 = 61.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.939130.
END IF.
EXECUTE.

DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(VALUE(Events) EQ 1 OR VALUE(Events) EQ 13) AND (VALUE(Temperature) GT
50 AND VALUE(Temperature) LE 66).
COMPUTE nod_001 = 18.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.948718.
END IF.
EXECUTE.

DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
(VALUE(Events) EQ 1 OR VALUE(Events) EQ 13) AND (VALUE(Temperature) GT
66 AND VALUE(Temperature) LE 71) AND (VALUE(SSexDescription) EQ 92).
COMPUTE nod_001 = 39.
COMPUTE pre_001
    = 1.
COMPUTE prb_001 = 0.865385.
END IF.
EXECUTE.

/* Node 40 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
        (VALUE(Events) EQ 1 OR VALUE(Events) EQ 13) AND
        (VALUE(Temperature) GT 66 AND VALUE(Temperature) LE 71) AND
        (SYSMIS(SSexDescription) OR VALUE(SSexDescription) NE 92).
COMPUTE
    nod_001 = 40.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 1.000000.
END IF.
EXECUTE.

/* Node 20 */.
DO IF (SYSMIS(SPrimarypayor) OR (VALUE(SPrimarypayor) LE 304)) AND
        (VALUE(Events) EQ 1 OR VALUE(Events) EQ 13) AND
        (SYSMIS(Temperature) OR (VALUE(Temperature) GT 71)).
COMPUTE nod_001 = 20.
COMPUTE pre_001 = 1.
COMPUTE prb_001 = 0.974425.
END IF.
EXECUTE.

/* Node 6 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND
        (VALUE(SProviderName) LE 63).
COMPUTE nod_001 = 6.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.486842.
END IF.
EXECUTE.

/* Node 62 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND
        (SYSMIS(SProviderName) OR
        (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND
(VALUE(Temperature) LE 37) AND (SYSMIS(SProviderType) OR
(VALUE(SProviderType) LE 411)) AND (VALUE(PovertyLevel)
LE 12.8).
COMPUTE nod_001 = 62.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.615385.
END IF.
EXECUTE.

/* Node 86 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR
(VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND
(VALUE(Temperature) LE 37) AND (SYSMIS(SProviderType) OR
(VALUE(SProviderType) LE 411)) AND (SYSMIS(PovertyLevel
) OR (VALUE(PovertyLevel) GT 12.8)) AND (SYSMIS(SProviderName) OR
(VALUE(SProviderName) LE 67)).
COMPUTE nod_001 = 86.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.815018.
END IF.
EXECUTE.

/* Node 87 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR
(VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND
(VALUE(Temperature) LE 37) AND (SYSMIS(SProviderType) OR
(VALUE(SProviderType) LE 411)) AND (SYSMIS(PovertyLevel
) OR (VALUE(PovertyLevel) GT 12.8)) AND (VALUE(SProviderName) GT 67 AND
VALUE(SProviderName) LE 69).
COMPUTE nod_001 = 87.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.864023.
END IF.
EXECUTE.

/* Node 88 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR
(VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND
(VALUE(Temperature) LE 37) AND (SYSMIS(SProviderType) OR
(VALUE(SProviderType) LE 411)) AND (SYSMIS(PovertyLevel
) OR (VALUE(PovertyLevel) GT 12.8)) AND (VALUE(SProviderName) GT 69).

COMPUTE nod_001 = 88.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.727273.
END IF.
EXECUTE.

/* Node 42 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) LE 37) AND (VALUE(SProviderType) GT 411).
COMPUTE nod_001 = 42.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.661765.
END IF.
EXECUTE.

/* Node 43 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 37 AND VALUE(Temperature) LE 50) AND (SYSMIS(Events) OR VALUE(Events) NE 3 AND VALUE(Events) NE 1).
COMPUTE nod_001 = 43.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.858871.
END IF.
EXECUTE.

/* Node 44 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 37 AND VALUE(Temperature) LE 50) AND (VALUE(Events) EQ 3 OR VALUE(Events) EQ 1).
COMPUTE nod_001 = 44.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.563636.
END IF.
EXECUTE.
/* Node 64 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 50 AND VALUE(Temperature) LE 55) AND (SYSMIS(SAppointmentday) OR VALUE(SAppointmentday) NE 2) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) LE 68)).
COMPUTE nod_001 = 64.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.812298.
END IF.
EXECUTE.

/* Node 65 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 50 AND VALUE(Temperature) LE 55) AND (SYSMIS(SAppointmentday) OR VALUE(SAppointmentday) NE 2) AND (VALUE(SProviderName) GT 68).
COMPUTE nod_001 = 65.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.622951.
END IF.
EXECUTE.

/* Node 66 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 50 AND VALUE(Temperature) LE 55) AND (VALUE(SAppointmentday) EQ 2) AND (SYSMIS(Precipitation) OR (VALUE(Precipitation) LE 0)).
COMPUTE nod_001 = 66.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.540541.
END IF.
EXECUTE.

/* Node 67 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 50 AND VALUE(Temperature) LE 55) AND (VALUE(SAppointmentday) EQ 2) AND (SYSMIS(Precipitation) OR (VALUE(Precipitation) LE 0)).
COMPUTE nod_001 = 67.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.540541.
END IF.
EXECUTE.
(Precipitation) GT 0).
COMPUTE nod_001 = 67.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.826923.
END IF.
EXECUTE.

/* Node 104 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SPrimarypayor) GT 63 AND VALUE(SProviderName) LE 71)) AND (SYSMIS(Temperature) OR (VALUE(Temperature) GT 55 AND VALUE(Temperature) LE 77)) AND (SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 12)) AND (SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND (SYSMIS(Precipitation) OR (VALUE(Precipitation) LE 0)) AND (SYSMIS(Events) OR VALUE(Events) NE 2 AND VALUE(Events) NE 1).
COMPUTE nod_001 = 104.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.885609.
END IF.
EXECUTE.

/* Node 105 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SPrimarypayor) GT 63 AND VALUE(SProviderName) LE 71)) AND (SYSMIS(Temperature) OR (VALUE(Temperature) GT 55 AND VALUE(Temperature) LE 77)) AND (SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 12)) AND (SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND (SYSMIS(Precipitation) OR (VALUE(Precipitation) LE 0)) AND (VALUE(Events) EQ 2 OR VALUE(Events) EQ 1).
COMPUTE nod_001 = 105.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.784000.
END IF.
EXECUTE.

/* Node 90 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (SYSMIS(Temperature) OR (VALUE(Temperature) GT 55 AND VALUE(Temperature) LE 77)) AND (SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 12)) AND (SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND (VALUE(Precipitation) GT 0 AND VALUE(Precipitation) LE 0.07000000000000001).

COMPUTE nod_001 = 90.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.898551.
END IF.
EXECUTE.

/* Node 91 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (SYSMIS(Temperature) OR (VALUE(Temperature) GT 55 AND VALUE(Temperature) LE 77)) AND (SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 12)) AND (SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND (VALUE(Precipitation) GT 0.07000000000000001).

COMPUTE nod_001 = 91.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.805556.
END IF.
EXECUTE.

/* Node 69 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (SYSMIS(Temperature) OR (VALUE(Temperature) GT 55 AND VALUE(Temperature) LE 77)) AND (SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 12)) AND (SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 102)) AND (VALUE(Precipitation) GT 0.07000000000000001).

COMPUTE nod_001 = 69.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.782456.
END IF.
EXECUTE.
/* Node 70 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (SYSMIS(Temperature) OR (VALUE(Temperature) GT 55 AND VALUE(Temperature) LE 77)) AND (SYSMIS(SAppointmentType) OR (VALUE(SAppointmentType) LE 12)) AND (VALUE(SRacedescription) GT 103).
COMPUTE nod_001 = 70.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.786096.
END IF.
EXECUTE.

/* Node 48 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (SYSMIS(Temperature) OR (VALUE(Temperature) GT 55 AND VALUE(Temperature) LE 77)) AND (VALUE(SAppointmentType) GT 12).
COMPUTE nod_001 = 48.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.636364.
END IF.
EXECUTE.

/* Node 71 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 77 AND VALUE(Temperature) LE 80) AND (SYSMIS(Events) OR VALUE(Events) NE 2 AND VALUE(Events) NE 24 AND VALUE(Events) NE 4) AND (VALUE(SAppointmentday) EQ 6 OR VALUE(SAppointmentday) EQ 4).
COMPUTE nod_001 = 71.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.806202.
END IF.
EXECUTE.

/* Node 72 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND
(VALUE(Temperature) GT 77 AND VALUE(Temperature) LE 80) AND
(SYSMIS(Events) OR VALUE(Events) NE 2 AND VALUE
(Events) NE 24 AND VALUE(Events) NE 4) AND (SYSMIS(SAppointmentday) OR
VALUE(SAppointmentday) NE 6 AND VALUE(SAppointmentday) NE 4).
COMPUTE nod_001 = 72.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.936047.
END IF.
EXECUTE.

/* Node 73 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR
(VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND
(VALUE(Temperature) GT 77 AND VALUE(Temperature) LE 80) AND
(VALUE(Events) EQ 2 OR VALUE(Events) EQ 24 OR VALUE
(Events) EQ 4) AND (VALUE(SAppointmentday) EQ 6).
COMPUTE nod_001 = 73.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.580645.
END IF.
EXECUTE.

/* Node 92 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR
(VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND
(VALUE(Temperature) GT 77 AND VALUE(Temperature) LE 80) AND
(VALUE(Events) EQ 2 OR VALUE(Events) EQ 24 OR VALUE
(Events) EQ 4) AND (SYSMIS(SAppointmentday) OR VALUE(SAppointmentday)
NE 6) AND (SYSMIS(SSexDescription) OR VALUE(SSexDescription) NE 91).
COMPUTE nod_001 = 92.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.827586.
END IF.
EXECUTE.

/* Node 93 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR
(VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND
(VALUE(Temperature) GT 77 AND VALUE(Temperature) LE 80) AND
(VALUE(Events) EQ 2 OR VALUE(Events) EQ 24 OR VALUE
(Events) EQ 4) AND (SYSMIS(SAppointmentday) OR VALUE(SAppointmentday) NE 6) AND (VALUE(SSexDescription) EQ 91).
COMPUTE nod_001 = 93.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.651515.
END IF.
EXECUTE.

/* Node 106 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 80) AND (VALUE(SAppointmentday) EQ 6 OR VALUE(SAppointmentday) EQ 5) AND (SYSMIS (SSexDescription) OR VALUE(SSexDescription) NE 91) AND (SYSMIS(Distance) OR (VALUE(Distance) LE 4.3955)) AND (VALUE(SFathersEmployment) LE 2011).
COMPUTE nod_001 = 106.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.803922.
END IF.
EXECUTE.

/* Node 107 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 80) AND (VALUE(SAppointmentday) EQ 6 OR VALUE(SAppointmentday) EQ 5) AND (SYSMIS (SSexDescription) OR VALUE(SSexDescription) NE 91) AND (SYSMIS(Distance) OR (VALUE(Distance) LE 4.3955)) AND (SYSMIS(SFathersEmployment) OR (VALUE(SFathersEmployment) GT 2011)).
COMPUTE nod_001 = 107.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.944444.
END IF.
EXECUTE.

/* Node 95 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 80) AND (VALUE(SAppointmentday) EQ 6 OR VALUE(SAppointmentday) EQ 5) AND (SYSMIS
(SSexDescription) OR VALUE(SSexDescription) NE 91) AND (VALUE(Distance) GT 4.3955).
COMPUTE nod_001 = 95.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.746032.
END IF.
EXECUTE.

/* Node 76 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 80) AND (VALUE(SAppointmentday) EQ 6 OR VALUE(SAppointmentday) EQ 5) AND (VALUE(SSexDescription) EQ 91).
COMPUTE nod_001 = 76.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.766839.
END IF.
EXECUTE.

/* Node 96 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 80) AND (SYSMIS(SAppointmentday) OR VALUE(SAppointmentday) NE 6 AND VALUE(SAppointmentday ) NE 4 AND VALUE(SAppointmentday) NE 5) AND (SYSMIS(Events) OR VALUE(Events) NE 2 AND VALUE(Events) NE 1) AND (SYSMIS(SRacedescription) OR (VALUE(SRacedescription) LE 103)).
COMPUTE nod_001 = 96.
COMPUTE pre_001 = 6.
COMPUTE prb_001 =
 0.875949.
END IF.
EXECUTE.

/* Node 97 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR (VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND (VALUE(Temperature) GT 80) AND (SYSMIS(SAppointmentday) OR VALUE(SAppointmentday) NE 6 AND VALUE(SAppointmentday
) NE 4 AND VALUE(SAppointmentday) NE 5) AND (SYSMIS(Events) OR
VALUE(Events) NE 2 AND VALUE(Events) NE 1) AND
(VALUE(SRacedescription) GT 103).
COMPUTE nod_001 = 97.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.865672.
END IF.
EXECUTE.

/* Node 78 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR
(VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND
(VALUE(Temperature) GT 80) AND (SYSMIS(SAppointmentday) OR
VALUE(SAppointmentday) NE 6 AND VALUE(SAppointmentday)
) NE 4 AND VALUE(SAppointmentday) NE 5) AND (VALUE(Events) EQ 2 OR
VALUE(Events) EQ 1).
COMPUTE nod_001 = 78.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.780000.
END IF.
EXECUTE.

/* Node 53 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (SYSMIS(SProviderName) OR
(VALUE(SProviderName) GT 63 AND VALUE(SProviderName) LE 71)) AND
(VALUE(Temperature) GT 80) AND (VALUE(SAppointmentday) EQ 4).
COMPUTE nod_001 = 53.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.728507.
END IF.
EXECUTE.

/* Node 27 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (VALUE(SProviderName) GT 71)
AND (SYSMIS(Events) OR VALUE(Events) NE 3 AND VALUE(Events) NE 13).
COMPUTE nod_001 = 27.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.739583.
END IF.
EXECUTE.
/* Node 28 */.
DO IF (VALUE(SPrimarypayor) GT 304) AND (VALUE(SProviderName) GT 71)
AND (VALUE(Events) EQ 3 OR VALUE(Events) EQ 13).
COMPUTE nod_001 = 28.
COMPUTE pre_001 = 6.
COMPUTE prb_001 = 0.637500.
END IF.
EXECUTE.
APPENDIX 5: SURVEY AND INTRODUCTION LETTER

Title of Research: A Decision Support Tool to Model No-Show Appointments in Healthcare Industry

Researchers: Maria M. Rinder & Dr. Gary Weckman

You are being asked to participate in a voluntary online survey that should take 5-6 minutes to complete and the results are collected anonymously. The purpose of this survey is to gather information regarding your expertise. However, before you decide whether you will participate, you should understand more about how the research is being conducted, how your information will be used, and how your rights are protected.

The Consent Form on the next page of the survey describes the purpose, procedures, benefits, and risks related to this research. Once you have read the Consent Form, and have had your questions answered to your satisfaction, you will be asked to give your consent to participate in the online survey.

------------------[ Contact Information ]------------------

Maria M. Rinder, (513)720-7295, magdarinder@hotmail.com
Dr. Gary Weckman, (740)593-1548, weckmang@ohio.edu
Jo Ellen Sherow, (740)593-0664, compliance@ohio.edu
Title of Research: A Decision Support Tool to Model No-Show Appointments in Healthcare Industry

Researchers: Maria Rinder & Dr. Gary Weckman

You have been invited to participate in this research because of your knowledge of healthcare systems. You were selected to participate because you were either a personal contact of the research investigators, were referred by a fellow colleague, or were selected based on your public information available from professional networking corporations. The study being investigated is conducted by Maria Rinder (primary investigator) for a dissertation project under the supervision of Dr. Gary Weckman (faculty sponsor) of the Department of Industrial and Systems Engineering at Ohio University. Your expertise is needed to determine the usefulness of the information that is provided by a mathematical model that has been developed for the dissertation research by the investigators.

Your participation in this research is completely voluntary and will take you roughly 5-6 minutes to complete the online survey. If you participate, you will be asked basic questions about your experience and then be asked specific questions that relate to decision tools used in healthcare patient scheduling. The information collected from this survey may not benefit you directly, but will provide evidence to researchers who study
the applicability of decision-support methodologies. To participate in this study, you will need to provide your informed consent and you must be at least 18 years of age. If you have any questions about the research or survey please call the primary investigator (Maria Rinder) at (513)720-7295 or emailing her at magdarinder@hotmail.com. You may also contact the faculty sponsor (Gary Weckman) at (740)593-1548 at weckmang@ohio.edu.

There are no known risks if you decide to participate in this research study and you can drop out at any time without any consequence. This survey is anonymous so you will not be required to provide your name or other contact information that can identify you or your affiliation during the online survey. The Ohio University Institutional Review Board has reviewed a request to conduct this investigation and if you have any concerns about your rights, please contact the Director of Research Compliance at Ohio University (Jo Ellen Sherow) at (740)593-0664 or by email at compliance@ohio.edu.

-----------------[ Contact Information ]-----------------
Maria Rinder, (513)-720-7295, magdarinder@hotmail.com
Dr. Gary Weckman, (740)593-1548, weckmang@ohio.edu
Jo Ellen Sherow, (740)593-0664, compliance@ohio.edu
To participate in this study, you are agreeing that:

- you have read this consent form (or it has been read to you)
- you have been given the opportunity to ask questions and have them answered
- you have been informed of potential risks and they have been explained to your satisfaction.
- your participation in this research is completely voluntary
- you may leave the study at any time without consequence.
- you are 18 years of age or older

Yes, I will give my consent to participate in this research

No, I will NOT give my consent to participate in this research
Invitation letter

Dear Invited Participate,

You are being invited to participate in new research regarding patient no-show forecasting in healthcare scheduling. This online survey requires you to apply your knowledge and expertise on this subject.

This study is being conducted by Maria M. Rinder for her dissertation project under the supervision of Dr. Gary Weckman at Ohio University. The objective of the research is to develop a decision support tool that can incorporate individual probability of patient no-show into a scheduling decision support system. The research involves a mathematical model that calculates the individual probability of patient’s no-show based on patient demographic and environmental variables. The preliminary results of the research correctly predict no-show over 90% of the time. Probability of no-show for each patient can then be used as part of a scheduling system. It is anticipated that this can improve healthcare provider resource utilization. Your expertise is needed to help assess the usefulness of this decision support tool through this online survey.

The survey will take roughly 6 minutes to complete. No risks have been identified in participating in this voluntary study since your information will remain anonymous and you can drop out of the online survey at any time without consequence. However, due to
Ohio University policy, your consent is needed and you must be 18 years of age or older to participate. More information regarding the study can be found by accessing the survey link below or by hitting the 'next' button if you are already accessing the online survey. If you have any questions related to the research, or the survey, please do not hesitate to contact the primary researcher (Maria), her faculty adviser (Gary), or Ohio University’s Director of Research Compliance with the contact information below.

Please feel free to refer this information to your fellow colleagues who might be interested in participating in the survey or learning more about the research being conducted at Ohio University. I would like to thank you for reading this invitation, your time and valuable insight is greatly appreciated.

Sincerely,

Maria M. Rinder
1. Who/what decides whether and when to schedule a patient?
   - Human scheduler
   - Algorithm/model
   - Human scheduler assisted by algorithm/model

2. On a scale from 1 to 5, how satisfied are you with your current method of scheduling?

   1  2  3  4  5

   Unsatisfied          Very Satisfied

3. What information do you or others at your place of employment use when scheduling patients?

   [Blank space]

4. How useful do you believe a decision support tool that predicts patient no-show would be in healthcare patient scheduling?

   1  2  3  4  5

   Not useful          Very useful
5. Gauge your interest level in investigating the possibility of adopting such a decision support tool.

1  2  3  4  5
Little Interest                               Very Interested

6. What is your level of experience in healthcare?
   - None
   - 1-5 years
   - 6-9 years
   - 10-14 years
   - More than 15 years

7. You would describe your current position as which of the following?
   - Process Improvement
   - Scheduling
   - Administrative
   - Clinical
   - Other (please specify)

8. What changes would suggest to most improve the decision support tool capability presented here?
9. Please estimate the number of patients scheduled per month at your enterprise?

10. Please estimate your current no-show rate?
    - Don’t know
    - 0-5
    - 6-10
    - 11-15
    - 15-20
    - 21-25
    - More than 25%
APPENDIX 6: IRB APPROVAL EXEMPTION

A determination has been made that the following research study is exempt from IRB review because it involves:

Category 2. research involving the use of educational tests, survey procedures, interview procedures or observation of public behavior

Project Title: A Decision Support Tool to Model No-Show Appointments in Healthcare Industry

Primary Investigator: Maria M Rinder

Co-Investigator(s):

Advisor: Gary Weckman

Department: Industrial and Systems Engineering

Rebecca Cale, AAB, CIP
Office of Research Compliance

Date 3/24/12

The approval remains in effect provided the study is conducted exactly as described in your application for review. Any additions or modifications to the project must be approved (as an amendment) prior to implementation.