2016

Modeling Clinic Utilization by Considering Panel size, Multicomorbidities and Patient Scheduling

Mahsa Kiani

Modeling Clinic Utilization by Considering Panel size, Multi-comorbidities and Patient Scheduling

Mahsa Kiani

Thesis submitted

to the college of Engineering and Mineral Resources

at West Virginia University

in partial fulfillment of the requirements for the degree of

Master of Science in

Industrial Engineering

Kenneth Currie, Ph.D. Chair

Majid Jaridi, Ph.D.

Michael Carr, M.S.

Department of Industrial and Management Systems Engineering

Morgantown, West Virginia

August 2016

Keywords: Patient panel size, Queuing Theory, No-show rate, Re-scheduling rate, Multi-comorbidity, Discrete event simulation

ABSTRACT

Modeling Clinic Utilization by Considering Panel size, Multi-comorbidities and Patient Scheduling

Mahsa Kiani

Many appointment-based clinical systems experience long waiting times. Consequently, these systems experience higher rates of cancellation or no-show. This problem creates dissatisfaction among customers, as well as inefficiencies in healthcare systems, but more importantly, increases medical complications due to postponement of care. As an added complication, sometimes no-showing patients will reschedule appointments and the rate and reschedule discipline can have significant effects on overall patient satisfaction and system efficiency. In this study, a one server, multi-class queuing network model is proposed in which patients have a probability of no-showing as well as a rescheduling rate. No-show and rescheduling rates are computed based on the current backlog of the system. This model categorizes patients into different classes, based on number of comorbidities, with individual service times and arrival rates. In addition to considering the differences of various classes of patients, the model also decreases the under-utilization of resources by considering the no-show and rescheduling rate of customers. The purpose of the model is to determine the number of patients representing the panel size allocated to a specific physician, with recommendations for adding physicians to alleviate increasing backlogs based on increasing rates of comorbidity. In the second section of the study, the appointment system is simulated, and its results are compared with those generated by queuing theory. A preference model is then introduced which gives patients an option of choosing among all available appointments. The simulation results suggest that allowing patients to choose their favorite appointment time does not affect overall system utilization.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Dr. Kenneth Currie, for his continued guidance and support during my Master's program at West Virginia University. I would also like to thank him for going through my work meticulously and helping me to achieve my goals.

I would like to express my gratitude to Dr. Majid Jaridi and Mr. Michael Carr whose inputs and feedbacks have been very useful.

Finally, I would like to thank my family, all my colleagues and friends who have helped me during these two years.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

ABBREVIATION

ACA	Affordable Care Act
VA	Veterans Affairs
DES	Discrete Event Simulation
IID	Independent and Identically Distributed
PCIM	Primary Care Internal Medicine
FCFS	First Come First Served
PS	Processor Sharing
IS	Infinite Server
LCFSPR	Last Come First Serve with Preemptive-Resume

Chapter 1

1. INTRODUCTION

1.1. Background

In today's healthcare system, the increase in requests for doctors combined with a shortage of physicians has led to an increase in the number of patients who ask for appointments at different clinics. The significant outcome of the increase in appointment demand is the growth of patient panel size, the number of unique patients who are allocated to a specific doctor. More specifically, there are two recent healthcare developments that have heavily contributed to the growing number of requests for appointments at U.S. clinics.

The first consideration is the Medicaid expansion under the Affordable Care Act (ACA), which first requires an understanding of the differences between Medicaid and Medicare. Medicare is an insurance program under which medical bills are paid from trust funds, which those covered have paid into, which primarily serves people over the age of 65 regardless of income. This program also covers younger disabled patients and dialysis people (DCD, 2015). Patients pay part of the cost for hospital stays and other costs through deductibles, which do not vary significantly across the country (DCD, 2015). On the other hand, Medicaid serves low-income people without considering their age. Patients typically do not pay for covered medical charges, however sometimes a small co-payment is required (DCD, 2015). The ACA expanded only Medicaid coverage. Based on reports, by 2022 this program will insure 33 million Americans (Klein, 2012). 30 states as well as the District of Columbia opted to expand Medicaid under the ACA.

Despite the benefits that the ACA brought, it caused some problems as well. This program made cuts in some doctors' payments, prompting some physicians and patients to warn that the reductions could make some problems for patients to get service (Pear, 2014). Considering the fact that this program will add millions of Americans to the government entitlement scheme, the shortage of doctors through the United States is set to get much worse. Also, statistics show that with the rollout of the ACA law, people are more willing to see doctors and physicians. In other words they are willing to see doctors for less important issues. As a result,

the number of appointment requests will increase, further compounding the other problems associated with a shortage of doctors that has resulted in an increasing panel size.

The second contributing factor for a rise in physician appointment requests is the Veteran's Health Administration. The Veteran's Health Administration is part of the Department of Veterans Affairs (VA). It is a government-run network of 1,700 hospitals, clinics, counseling centers and nursing homes across the country which annually serves about 9 million of the nation's 22 million veterans (Somashekhar, 2014). The VA wants all veterans to have access to health care, which is available to anyone who served in the military and was discharged under any condition aside from dishonorable (Somashekhar, 2014). If somebody is under VA's healthcare coverage, he or she does not need to take extra steps to reach the ACA coverage standards (U.S.VA, 2015). Kesling (2014) reported that the VA's main goal is that no more than 14 days would be acceptable for the time between a patient's request date and the actual appointment date. This period is presumed to be the waiting time of patients for getting appointment. The actual waiting time of patients is reported to be 115 days, with 84% of patients having to wait more than 14 days. One reason for the long waiting is the aforementioned shortage of doctors. According to the American Federation of Government Employees, "some VA doctors are carrying workloads of more than 2,000 patients — far more than the 1,200 goal set forth in the Veterans Health Administration handbook" (Somashekhar, 2014). Washington Post's website states that "the agency is struggling to hire 400 primary-care physicians, positions that are notoriously hard to fill because of a nationwide shortage of these types of doctors" (Somashekhar, 2014). Based on this website's report, this is not just a VA matter but an issue troubling the U.S. medical system in general. Meanwhile, the demand for VA services has increased (Somashekhar, 2014).

Considering these dual aspects along with the shortage of physicians that exists in many clinics all over the U.S., the growth in panel size of patients to cover increasing number of appointments is not surprising. One significant consequence of increasing panel size is an increase in no-show probability. The no-show probability, or no-show rate, is the rate at which patients do not show up for their appointments. There are different reasons behind no-shows, for example patients may have transportation problems or simply forget about the appointment. The biggest contributing factor, though, is the amount of backlog, which is related to the number of patient requests. The amount of backlog is the maximum number of patients who have already gotten appointment and are in the queue in front of the new patient.

Increased no-show rates has detrimental effects for clinics, and various studies have investigated these consequences. It has been found to have a noticeable effect on annual clinic revenue (Moore et al., 2001), and causes idle time for physicians. No-show patients also affect the arrival rates within the system due to rescheduling of no-show appointments. Consequently, trying to control and reduce the number of no-show patients is an important problem worthy of study. There have been several suggested solutions to decrease the no-show rate. For instance:

1) Reminding patients about their appointments by email or phone,
2) Providing patient transportation to facilities,
3) Providing nursery care for patients with babies,
4) Updating personal and contact information of patients, and tracking patients who historically do not show up for their appointments, and
5) Sending a gift card for patients who show up for their appointments.

Some clinics have a policy of charging no-showing patients a fee to deter them from not showing up to scheduled appointments. Despite all of these strategies, high no-show rates continue to exist in health care systems and reduce their efficiency.

One of the other important issues facing the healthcare system, is the subject of multi-comorbidity patients. Feinstein (1970) first defined the word 'comorbidity' as "any distinct additional clinical entity that has existed or may occur during the clinical course of a patient who has the index disease under study". Multi-comorbidity is a situation in which two or more comorbidity conditions exist. Lui et al., (2013) state that "Because different types of patients may have different visit frequencies as well as various demand for providers' consultation time, multi-comorbidity situations directly influences both the "demand" and "supply" side of a practice" (Lui et al., 2013).

In most studies patients are considered the same, however in reality patients have different attitudes which is related to the multi-comorbidity issue. To clarify, patients with more comorbidities may need to see doctors more times in a year that patients with fewer comorbidities, or their care sessions may be longer. Thus, in the study of healthcare systems, the issue of multi-comorbidity must be seriously considered.

1.2. Research Objective

Considering the fact that the number of patients who request appointments is increasing each year, hospitals and clinics are facing a large volume of demand. Therefore, their objective must be to optimize their appointment system by allocating the maximum number of patients to doctors as possible. It is precisely this problem that is the focus of this project, resulting in obtaining the optimal panel size of patients to assign to an individual physician. To obtain the optimal panel size, two main factors are considered. The first factor is the rate of no-show patients, a function of the amount of backlog. As the state of the system changes, the no-show rate will be updated and calculated based on the current amount of backlog. Among no-show patients, some proportions will reschedule their appointments, which is called the rescheduling rate. During this study, a function based on the amount of backlog is proposed to update the rescheduling rate of the system at each step. This is the second considered factor. The third factor is the problem of different groups of patients with different number of comorbidities, which helps in allocating the right number of appointment spots for each group.

Chapter 2

2. LITERATURE REVIEW

2.1. Background

For many service systems, obtaining the optimal amount of servers is a basic issue, since it affects the quality of services, wait time of the patients, and total revenue of the system. For health care systems, this concern is a critical matter.

Panel size of a clinic, is the number of unique patients who are allocated to a specific doctor. Estimating the panel size can be a useful method for server planning. Altschuler et al. (2012) estimated the patient panel size for primary care doctors by considering different models that allocated some parts of preventive and chronic care to non-physician persons, and observed that they could offload preventive care and chronic care which were possible with their existing workforce. The amount of backlog determines the maximum number of patients in the appointment queue who have gotten their appointment but still have not met with a physician. Hawkins (2011), by perusing 1,162 medical offices in different areas in the U.S., found that waiting time for getting an appointment depended on the specialty of the treatment. For example, the waiting time was 22.1 days for dermatology, 20.3 days for family practice and just 16.8 days for orthopedic surgery. In 2014, a problem at the VA concerning scheduling timely access to medical care was published. Kesling (2014) reported that the VA's main goal is that no more than 14 days would be acceptable as waiting time of patients for getting appointment. However, the actual waiting time of patients is reported 115 days, and 84% of patients had to wait more than 14 days which is a far cry from their goal. One of the outcomes of a large backlog is increase the no-show rate.

2.1.1. No-showing rate

All appointment-based service systems, and particularly health care systems, suffer from high no-show rates. Based on the characteristics of each clinic, the no-show rate is different and can reach up to 60% (Cayirli et al., 2006). Defife et al. (2010) reported a 21% no-show rate for

a psychotherapy clinic while Dreiher et al. (2008) found a 30% no-show rate in an outpatient obstetrics and gynecology clinic. Green and Savin (2008) mentioned that "no-show patients create a paradoxical situation in which a physician is under-utilized while patients have long waits before getting appointments" (Green and Savin, 2008). Moore et al. (2001) estimated that about 31% of the patients who had an appointment did not show up in a family clinic. They investigated the consequences of this above average rate of no-show, and noticed between 3% and 14% of annual revenue was lost because of missed appointments.

Evidence shows that the rate of no-shows increases with the growth of appointment backlogs (Gallucci et al., 2005). Gallucci et al. (2005) considered the effect of different variables on the probability of keeping an appointment. The major predictor is the number of days between asking for an appointment and the available time, while age, sex and comorbidity are considered potential confounding variables. The chi square test for trend was used to assign the relationship between appointment delay and missed appointment. Multivariate logistic regression was used to appraise the magnitude of the relationships between the predictors and missed appointments. The results show that gender, age and number of comorbidities influence the no-show rate for some patients. However, the most important factor is appointment delay or the amount of backlog. For every day of increased appointment delay, the possibility of a no-show increases. Wang and Gupta (2011) investigated the effective factors on no-show probability, and proved that the history of a patient can change no-show rate in addition to backlog. If a person has a record of not coming to his or her appointment, the probability of a no-show for him or her will increase next time. Lacy et al. (2004) interviewed 34 patients of an outpatient care clinic and found three main reasons for not showing: emotions (like fear and anxiety), perceived disrespect, and not understanding the scheduling system.

Various methods have been proposed to decrease the rate of no-shows, including reminder calls, charging no-show patients or providing some transportation to the clinic. Pesata et al. (1999) conducted a survey of patients who called for an appointment at a clinic but did not show up, and 51% of them stated that a transportation problem was the cause. Tuso et al. (1999) worked on reducing backlog to decrease the rate of no-shows, and found that about 25-50% of the patients on the waiting list did not need a return appointment. They suggested that sending a letter to the remaining patients to remind them to call for an appointment would reduce the backlog to an acceptable amount. Schmalzried and Liszak (2012) generated an intervention program to reduce the no-show rate. Based on their plans, a clinic sends e-mails to patients explaining the consequences of no-showing. Based on these policies, if a person does not show

up for three times, he or she must attend a reinstatement class. This approach reduced the no-show rate from 34% to 10%. Some other strategies like tracking every no-show patient in computer systems, following up with missing patients or charging absent patients can also be helpful in decreasing the no-show rate.

The effect of no-show on appointment scheduling has been shown in numerous simulation studies that allow general complexities in appointment systems while investigating the effects of varying the service time mean and variability (Robinson and Chen, 2010; Cayirli et al., 2006; Ho and Lau, 1992; Ho and Lau, 1999). There are also several analytical papers which involve no-show rates in appointment and system planning. The earliest studies consider patients who may arrive late or not at all in queuing models (Mercer, 1960; Mercer, 1973). Liu and Ziya (2013) considered the relationship between the no-show rate and appointment delay, and made demand and capacity control decisions. They used a one-server queuing model after considering two models. In the first, the fixed service capacity and the decision variable is the panel size, while in the second, the panel size and the service capacity are both decision variables. Their purpose was to maximize the net reward function. It is said that, "in addition to the magnitude of patient show-up probabilities, patients' sensitivity to incremental delays is an important determinant of how demand and capacity decisions should be adjusted in response to anticipated changes in patients' no-show behavior" (Liu and Ziya (2013)). Kaandorp and Koole (2007) developed an algorithm to obtain the optimal appointment times, considering exponential service times and the existence of no-show patients. Zeng et al. (2008) applied heterogeneous no-show rates to this model.

2.1.2. Optimizing panel size using queuing theory

As stated in the beginning, a useful method for handling the increase in demand for appointments in healthcare systems is optimizing the panel size of the patients. Based on the concepts of backlog and no-show rate, there are some studies which involve these facts in analytical methods to obtain the panel size. Garcia et al. (2002) have proposed a closed form solution for the M/D/1/K queue to estimate the panel size. Green and Savin (2008) updated Garcia's method to include a no-show rate. They proposed two methods: an M/D/1/K queue with a state-dependent no-show and an M/M/1/K queue with a state-dependent no-show. They assumed that there was a non-eligible no-show rate even for same day appointments. The rate of no-show increases as backlog increases until it reaches a maximum, at which point the rate

of no-show stabilizes at its maximum value. In the M/D/1/K queue model, it is assumed that patients' service time is constant, however in the M/M/1/K queue model, their service times are independent and identically distributed (i.i.d.) exponential random variables. In this model a constant rescheduling rate was considered for the patients who did not show up to their appointments. The reschedule rate was set equal to 1, meaning that all the patients who were absent would reschedule their appointments. In addition, this study assumed that patient mean arrival and service times were identical. The research presented in this thesis relaxes both of these assumptions.

The assumption that is common among many of the queuing analyses is that the arrival rate and service time of patients are constant, despite the fact that patients who enter the model do not have similar medical needs. A significant factor that causes these differences is the existence of multi-comorbidity conditions. Different numbers of comorbidities bring about various inter-arrival times (frequency of visits) and length of service times. Fortin et al. (2005) mentioned the scarcity of research in the area of multi-comorbidity patients in comparison to specific diseases, even though the behavior of these patients can significantly affect the efficiencies of healthcare systems. The queuing method that can be used to show the differences between multi-comorbidities is multi-class closed queuing networks, which provides a convenient framework with which to evaluate the impact of population constraints on the stochastic interactions between different classes at various nodes of the network (Satyam et al., 2013). The research detailed in this thesis utilizes a multi-class, closed queuing network to investigate the impact of multi-comorbidity patients on patient backlog. Baynat and Dallery (1993) offered a method for obtaining estimated solutions of general closed queuing networks with a number of classes of customers. Their proposal was to associate a single-class closed queuing network with load-dependent exponential service stations to each class of patients. Satyam et al. (2013) presented a new approach to analyzing general multi-class closed queuing networks. Satyam et al. (2013) mentioned that "this approach is based on parametric characterization of the traffic processes in the network, which uses two-moment approximations to estimate performance measures at individual nodes" (Satyam et al., 2013). This model consists of R classes and J nodes, with each node modeled as a single server queue. The service time distribution of a class r at a node j is characterized by two parameters: the mean, τ_{rj}, and squared coefficient of variation, $c_{s_{rj}}^2$ .As a result, the service rate,μ_{rj} is equal to τ_{rj}^{-1}(Satyam et al., 2013). The arrival process is described by the mean and SCV parameters

$(\lambda_{a_{Aj}}^{-1}, c_{a_{Aj}}^{2})$ (Satyam et al., 2013), and the arrival time distribution and service time distribution for a multi-class queuing network can be estimated.

2.1.3. Application of simulation in healthcare system

An alternative approach to modeling healthcare systems instead of queuing models is simulation. "Simulation is the imitation of the operation of a real-world process or system over time" (Banks, 1998). Simulation has many advantages in comparison to other methods. The level of detail of information that can be obtained using simulation is one of its benefits. Discrete event simulation (DES) is "a type of computer simulation that imitates the operation of a real-world system over discrete units of time" (Hamrock et al., 2013). Discrete event simulation has many application in improving healthcare systems. "Some models of outpatient clinics aim to improve patient flow, reduce wait times, maximize staff utilization, and accomplish other gains in efficiency" (Hamrock et al., 2013). Wang et al. (2011) found that from 1996-2006 the demand for emergency departments of hospitals in the U.S. increased by 30% while the number of emergency departments decreased by 5%. To combat this trend, computer simulation is used in many emergency departments to decrease the length of patient stays. Hashimoto and Bell (1996) improved outpatient clinic staffing and scheduling based on simulation results. They changed the number of different resources and calculated waiting time for nurses and doctors, and obtained the optimal use of resources. Evans et al. (1996) developed a simulation model for emergency department using Arena software to have a means of investigating the desirability of various possible personnel schedules. Raunak and Osterweil (2005) described a resource model based on resource classes, resource instances, their attributes and relationships among them. Considering constraints of resources, they simulated the model and managed the utilization of them.

2.2. Motivation for this project

During the previous literature review, two methods (queuing and simulation) were commonly used to obtain optimal patient panel size, where "optimal" is taken to mean the amount that gives the smallest appointment backlog. The first part of this research has been derived from the Green and Savin (2008) study, which was the most closely related published research, however improvements were suggested to their original model. While they assumed that all patients were identical, with equal arrival rates and service times. The main motivation of this research is that patients were categorized into subgroups according to the number of comorbidities. By considering the system as a multi-class closed queuing network, desirable service times and arrival rate parameters related to each group will be estimated. In these networks, each class of customers has its own inter-arrival and service time. The estimated parameters will then be used in the M/D/1/K queuing model. In this model, K is the maximum capacity of the system. From this the panel sizes for the whole system and each individual group will be calculated. The no-show and rescheduling rates will also be considered in this model. The research will include a no-show function similar to the one used in the Green and Savin (2008) study, and increases as the backlog increases. While Green and Savin (2008) used a constant number for the rescheduling rate, this research will include a rescheduling function based on the size of the backlog. It will be used in the M/D/1/K model, to estimate the panel size for each group and for the whole system. The second phase of this research will be to simulate the queuing model assumptions as a discrete event simulation appointment system. At first, the simulation results will be compared with the ones obtained using queuing theory. Then the simulation model will be exercised for two different situations; (1) The first situation will be based on the first available appointment, meaning that when a person asks for an appointment, he or she is considered for the first available appointment. (2)The second situation will give more authorization to patients by allowing 25% of patients to choose between all available appointments.

Chapter 3

3. QUEUING THEORY

The first section of this research is application of queuing theory in investigation of a clinical appointment system. After patients ask for an appointment, they have to wait until they get a chance to see a physician. Meanwhile, they may change their plan. They might cancel their appointment, or simply not show up on their appointment date. Cancelation gives the system a chance to replace the appointment with a new patient, but no-showing causes various disadvantages. There are several reasons for not showing up, for example: holidays, special days of the week (eg, Mondays or Fridays), transportation problems or forgetting about the appointment. According to past studies, one of the most important factors that increases the no-show rate is patient backlog. Gallucci et al. (2005) reported that the rate of cancellations and no-shows are dependent on the backlog at the time a patient receives an appointment. They presume a function for the no-showing rate based on the amount of backlog. This function has three specific characteristics:

1. There is a no-show rate, albeit low, even for same day appointments.
2. The rate of no-shows monotonically increases with the increase in backlog until it reaches a maximum.
3. The rate of no-shows stabilizes when it reaches this maximum value (Green and Savin, 2008).

Using data from a public mental health clinic at the John Hopkins Bayview Medical Center in Baltimore, Gallucci et al. (2005) fitted the best line of no-show rate versus backlog. Green and Savin (2008) followed this approach and, with applying a line of best fit to their data, obtained the function of no-shows, based on backlog. Considering the three mentioned features, the suggested function is:

$$\gamma(k) = \gamma_{max} - (\gamma_{max} - \gamma_0)e^{-\frac{k}{C}}, \qquad (3.1)$$

Where *k* is the value of the appointment backlog when a patient asks for an appointment, γ_0 is the minimum observed no-show rate, γ_{max} is the maximum observed no-show rate and C is a no-show backlog sensitivity parameter (Green and Savin, 2008).

We fitted the mentioned function to the Columbia MRI facility data which is derived from Green and Savin (2008), and results are shown in Figure 1.

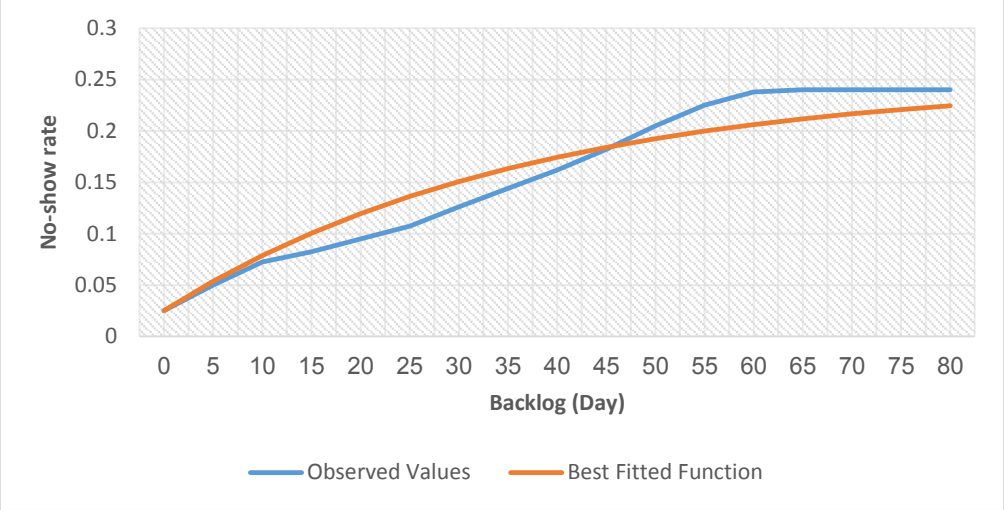

Figure 1. No-show rate based on the amount of backlog

The obtained function based on observed values is:

$$\gamma(k) = 0.24 - (0.24 - 0.025)e^{-k/37} \qquad (3.2)$$

On the basis of the fitted model, the no-show backlog sensitivity parameter is equal to 37. The maximum and minimum no-show rates are 24% and 2.5%, respectively.

3.1. M/D/1/K Queuing Model

The M/D/1/K model is a finite capacity queuing system in which a patient's arrival rate follows a Poisson distribution with mean λ , and constant service rate T. The maximum capacity is K. In the following model, the patient panel size is represented by N, and is measured by computing the number of unique patients seen by an individual doctor within a specific time

frame. Green and Savin (2008) considered arrival as a Poisson process with rate λN where λ is the arrival rate per patient and N is panel size of patients. They mentioned that "although the customer pool is a finite source, N is assumed to be large enough that the arrival rate is constant and is not dependent on the number of patients in service and in the appointment backlog" (Green and Savin, 2008). It is considered that queue length K is finite, and that when a new patient comes into the system, if the number of patients in the waiting line is K, he or she will be lost. The system is FIFO, which means each patient who comes first will get the appointment first, and service rate is deterministic with length T.

Initially, we represent some notations that are used in the model. These notations are derived from Green and Savin (2008).

$D(k,t,t+\Delta t)$ is the probability that a patient finishes his or her service between time instances t and $t+\Delta t$ leaving behind k patients in the appointment backlog, where $0 \le k \le K-1$. As a result, $D(t,t+\Delta t) = \sum_{k=0}^{K-1} D(k,t,t+\Delta t)$, is the probability that service for a patient will be finished in the time interval $[t,t+\Delta t]$ when k patients are in the backlog (Green and Savin, 2008). The corresponding departure rates are defined:

- $d(k,t) = \lim_{\Delta t \to 0} \frac{D(k,t,t+\Delta t)}{\Delta t}, \qquad 0 \le k \le K-1$ (3.3)
- $d(t) = \lim_{\Delta t \to 0} \frac{D(t,t+\Delta t)}{\Delta t}$ (3.4)

(Green and Savin, 2008).

The other important notations are backlog probabilities, $p(k,t)$, for $k = 0,\ldots,K$, which is the probability that the appointment backlog includes k patients at time t, and t is in a set of time intervals: $\Theta_n = \{t : (n-1)T \le t < nT\}, n \in N$ (Green and Savin, 2008).

The basic assumption considered in the model is that there are no patients in the appointment system at time $t=0$. Based on this assumption, $p(0,0)=1,\ p(k,0)=0, k=1,\ldots,K$ $d(k,0)=0, k=0,\ldots,K-1$. $p(k,t)$ and rates follow the below equations:

$$\frac{dp(0,t)}{dt} = -\lambda N p(0,t) \qquad (3.5)$$

$$\frac{dp(k,t)}{dt} = -\lambda N p(k,t) + \lambda N p(k-1,t), \qquad k=1,\ldots,K-1 \qquad (3.6)$$

$$\frac{dp(K,t)}{dt} = \lambda N p(K-1,t) \qquad (3.7)$$

(Green and Savin, 2008; Garcia et al., 2002).

Now for time interval $\Theta_n = \{t : (n-1)T \le t \le nT\}$, it is assumed the backlog probability $p(k,t)$, and departure rates $d(k,t)$, are known.

Let $\alpha_k(t) = \frac{(\lambda N t)^k e^{-\lambda N t}}{k!}$ (3.8) denote the probability that k patients arrive during time interval t.

Let $\rho = \lambda N T$, where T is service time, so we define:

$$\alpha(k) = e^{-\rho}\frac{\rho^k}{k!}, \qquad k \ge 0 \qquad (3.9)$$ (Green and Savin, 2008).

$\alpha(k)$ is the probability that k arrivals happen during a customer's service time.

As a result, the transition matrix P of the Markov Chain is as follows:

$$P = \begin{bmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \dots & \alpha_{K-2} & 1-\sum_0^{K-2}\alpha_k \\ \alpha_0 & \alpha_1 & \alpha_2 & \dots & \alpha_{K-2} & 1-\sum_0^{K-2}\alpha_k \\ 0 & \alpha_0 & \alpha_1 & \dots & \alpha_{K-3} & 1-\sum_0^{K-3}\alpha_k \\ \vdots & 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \alpha_1 & 1-\alpha_0-\alpha_1 \\ 0 & 0 & \dots & 0 & \alpha_0 & 1-\alpha_0 \end{bmatrix} \qquad (3.10)$$

For any time $t \in \Theta_n$, the departure rates are obtained:

$$d(0,t) = p(0,t-T)\lambda N(1-r\gamma(0))\alpha(0) + (1-r\gamma(0))\alpha(0)d(1,t-T), \qquad (3.11)$$

$$d(k,t) = p(0,t-T)\lambda N((1-r\gamma(k))\alpha(k) + r\gamma(k-1)\alpha(k-1)) + (1-r\gamma(k))\alpha(0)d(k+1,t-T) +$$

$$\sum_{i=1}^{k}((1-r\gamma(k))\alpha(k+1-i) + r\gamma(k-1)\alpha(k-i))d(i,t-T), \qquad k = 1,\dots,K-2, \qquad (3.12)$$

$$d(K-1,t) = p(0,t-T)\lambda N\left(\left(1-\sum_{i=0}^{K-2}\alpha(i)\right)(1-\mathrm{r}\,\gamma(K-1)) + r\gamma(K-2)\alpha(K-2)\right) +$$

$$\sum_{i=1}^{K-1} d(i,t-T)\times\left(\left(1-\sum_{j=0}^{K-1-i}\alpha(j)\right)(1-r\gamma(K-1)) + r\gamma(K-2)\alpha(K-1-i)\right), \qquad (3.13)$$

The proofs of all above equations are given by Green and Savin (2008) and Garcia et al. (2002).

To obtain the optimal panel size, the stationary backlog distribution $\pi(k)$ is needed. The stationary distribution satisfies:

$$\pi(k) = \frac{d^*(k)}{\lambda N}, k = 0,\ldots,K-1 \qquad (3.14)$$

where

$$\pi(k) = \lim_{t\to\infty} p(k,t), \qquad k = 0,\ldots,K \qquad (3.15)$$

$$d^*(k) = \lim_{t\to\infty} d(k,t), \qquad k = 0,\ldots,K-1 \quad \text{(Green and Savin, 2008).} \qquad (3.16)$$

Therefore:

$$d^*(0) = d^*(0)(1-r\gamma(0))\alpha(0) + d^*(1)(1-r\gamma(0))\alpha(0)\,, \qquad (3.17)$$

$$d^*(k) = d^*(0)((1-r\gamma(k))\alpha(k) + r\gamma(k-1)\alpha(k-1)) + (1-r\gamma(k))\alpha(0)d^*(k+1)$$
$$+\sum_{i=1}^{k}((1-r\gamma(k))\alpha(k+1-i) + r\gamma(k-1)\alpha(k-i))d^*(i) \qquad (3.18)$$
$$k = 1\ldots.,K-2$$

and

$$d^*(K-1) = d^*(0)((1-\sum_{i=0}^{K-2}\alpha(i))(1-r\gamma(K-1)) + r\gamma(K-2)\alpha(K-2)) + \sum_{i=1}^{K-1}((1-\sum_{j=0}^{K-1-i}\alpha(j))(1-r\gamma(K-1))$$
$$+r\gamma(K-2)\alpha(K-1-i))d^*(i)$$

(Green and Savin, 2008). (3.19)

A new notation, $f(k)$, is described as the proportion of the departure rate at k, over the departure rate at the 0 point, which is shown as:

$$f(k) = \frac{d^*(k)}{d^*(0)} \qquad k = 0,1,\ldots,K-1. \tag{3.20}$$

So for different amounts of k, f is obtained:

- $f(0) = 1,$ (3.21)
- $f(1) = \frac{e^{\rho}}{1 - r\gamma(1)} - 1,$ (3.22)
- $f(k+1) = \frac{e^{\rho}}{(1 - r\gamma(k+1))}.(f(k) - (1 - r\gamma(k+1))\alpha(k) - r\gamma(k)\alpha(k-1)) -$

$$\frac{e^{\rho}}{(1 - r\gamma(k+1))}(\sum_{i=1}^{k}((1 - r\gamma(k+1))\alpha(k+1-i) + r\gamma(k)\alpha(k-i))f(i)), \qquad k = 1,\ldots,K-1$$

(Green and Savin, 2008). (3.23)

Using all previous equations, the stationary backlog distribution is obtained.

$$\pi(0) = \frac{1 - r\gamma(K)}{1 - r\gamma(K) + \rho(\sum_{i=0}^{K-1} f(i)) - r\sum_{i=1}^{K-1}(\gamma(K) - \gamma(i))f(i)}, \tag{3.24}$$

$$\pi(k) = \frac{(1 - r\gamma(K))f(k)}{1 - r\gamma(K) + \rho(\sum_{i=0}^{K-1} f(i)) - r\sum_{i=1}^{K-1}(\gamma(K) - \gamma(i))f(i)}, \qquad k = 1,\ldots,K-1 \tag{3.25}$$

and

$$\pi(K) = 1 - \frac{(1 - r\gamma(K))(\sum_{i=0}^{K-1} f(i))}{1 - r\gamma(K) + \rho(\sum_{i=0}^{K-1} f(i)) - r\sum_{i=1}^{K-1}(\gamma(K) - \gamma(i))f(i)}.$$

(Green and Savin, 2008; Garcia et al., 2002). (3.26)

Equation (3.26) shows the stationary backlog distribution. In this equation, the no-show and rescheduling rate of patients are considered. For no-show rate, the aforementioned function based on the amount of backlog is used; however, the rescheduling rate is assumed to be a constant number. Green and Savin (2008) considered the rate to be 1 for simplicity.

3.2. Multi-Class Network

The current M/D/1/K model considers all patients the same, but in real healthcare systems there are different factors that make patients distinct from one another. One of these factors is the number of comorbidities. Based on the definition of the National Institute on Drug Abuse, when two disorders or illnesses occur in the same person, simultaneously or sequentially, they are described as comorbid (www.drugabuse.gov, 2010). Comorbidity also implies interactions between the illnesses that affect the course and prognosis of both (www.drugabuse.gov, 2010). Patients with multiple comorbidities need to see doctors more often. In other words, the number of comorbidities affects the frequency that each patient spends visiting his or her doctor and correspondingly has a higher mean arrival rate. Therefore, in this study patients are categorized based on the number of comorbidities, and it is assumed that patients in each group with the same number of comorbidities have the same arrival rate and service time. To achieve this goal, a multi-class queuing network model is used to represent the appointment system characteristics. First different multi-class networks and related parameters are introduced.

A multi-class queuing network, is one that services multiple groups of customers which may have various services and arrival rates, different routes through networks and per unit of waiting time cost (Bertsimas et al., 1994). Multi-class queuing networks consist of two special categories: closed multi-class queuing networks, open multi-class queuing networks.

In an open model jobs enter the network at random from outside at a fixed rate, are received at one or more nodes and eventually leave the network (Whitt, 1984). Thus, with an open model, the total external arrival rate or throughput is an independent variable, and the number of jobs in the system is a dependent variable (Whitt, 1984). On the other hand, in a closed model there is a fixed number of jobs in the network. Therefore, with a closed model the number of jobs in the system is an independent variable and the throughput is a dependent variable (Whitt, 1984). Closed queuing networks are classified into two groups: product-form networks, non-product-form networks. A queuing network is said to have a product-form solution when:

$$\mathrm{p}(n_1, n_2, \ldots, n_K) = \prod_{k=1}^{K} p_k(n_k) \qquad (3.27)$$

where $p_k(n_k)$ is a function only of the k^{th} node (Sahner et al., 2012).

This is true when the following characteristics hold:

1. The routing of customers from one service center to the next must be history independent, i.e., memory less (or Markovian).
2. The queuing disciplines may be FCFS (First Come First Served), PS (Processor Sharing), IS (Infinite Server) or LCFSPR (Last Come First Serve with Preemptive-Resume).
3. For an FCFS center, the service time distribution must be exponential; for other servers, the service time distribution does not have to be exponential but must be differentiable.
4. A product-form network may have multiple chains (multiple classes) of jobs and may be open with respect to some chains of jobs and closed with respect to others. External arrivals for all open chains must be Poisson distributed (Sahner et al., 1996).

However, product form networks need assumptions to obtain a product-form solution, and in the real world, they are not practical.

As mentioned before, a closed multi-class queuing network is used to show the characteristics of the appointment system. In the following, the applied network will be described using the methodology of Satyam et al. (2013).

The model consists of R classes, and patients in each class have the same number of comorbidities. The number of nodes is considered to be 1 since the panel size in being obtained for one physician. The service time mean of each class is given by τ_r. This means that the service rate for class r, μ_r, is equal to τ_r^{-1} (Satyam et al., 2013). The list of all notations used in the model is as follows:

R : Number of classes in the system.

μ_r : Service rate of class r, equal to τ_r^{-1}.

τ_A : Mean of service time of an aggregate class.

λ_r^{-1} : Mean of inter-arrival time of class r.

λ_A^{-1} : Mean of inter-arrival time of an aggregate class.

A is used to indicate the parameter of an aggregate class. The analysis in Whitt (1984) and Bitran and Tirupati, (1988) has led to the following equations:

$$\lambda_A = \sum_{r=1}^{R} \lambda_r \tag{3.28}$$

and

$$\tau_A = \sum_{r \in E} \phi_r \tau_r \quad \text{where } \phi_r = \frac{\lambda_r}{\lambda_A} \text{ for all } r \in E \,. \tag{3.29}$$

The above equations show the aggregate service rate and inter-arrival time for all classes. Patients with the same number of comorbidities are categorized into the same groups, and it is assumed that they have the same arrival rate and service time. Using these service times and arrival rates of different groups, it is possible to calculate the aggregate service rate and inter-arrival time for all classes and use it in the model, altering the M/D/1/K model of Green and Savin (2008). The panel size of each group of patients was estimated by applying this network. The strategy which is used to obtain the panel size is identical to that used by exactly the same as one that Green and Savin (2008), with a stationary backlog distribution.

3.3. Rescheduling Function

As mentioned before, equations 3.25 and 3.26 show the backlog distribution considering the no-show function and rescheduling rate. The assumption of the model is that each patient that no shows, will reschedule his or her appointment with a probability $r\gamma(k)$, where γ_k is the probability of no-show considering k backlog, and r is the rescheduling rate. In previous studies, the rescheduling rate was assumed to be constant. Green and Savin (2008) assumed a rescheduling rate of 1, indicating that all no-show patients would automatically reschedule their appointments. However, in reality, some patients may give up their appointments, or change the clinic and the physician. One of the factors that affects the behavior of patients who want to reschedule their appointment is the day of the new appointment. If the new date is not close to what they want, they might change the clinic and not reschedule. Green and Savin (2008) showed that, by increasing the amount of backlog and panel size, the rescheduling rate will decrease. Inspired by the no-show function, a rescheduling function was introduced while considering three basic assumptions:

1. There is a maximum rate for rescheduling that exists for same-day appointments.
2. The rate of rescheduling monotonically decreases with the increase in backlog until it reaches the minimum amount.
3. The rate of rescheduling stabilizes when it reaches this minimum value.

Using these assumptions, the re-scheduling function is estimated as:

$$r(k) = r_{min} + (r_{max} - r_{min})\, e^{-\frac{k}{s}} \qquad (3.30)$$

where r_{min} is the minimum rescheduling rate, r_{max} is the maximum rate, k is the amount of backlog at the time the patient wants to reschedule his or her appointment, and S is the re-scheduling function parameter. Green and Saving (2008) based on the Columbia MRI facility data, considered the maximum and minimum rescheduling rates to be 1 and 0, respectively. These, same values were applied to this function. An exponential function for the re-scheduling rate is assumed, however this rate can also be described as a linear function. The purpose of considering a function for re-scheduling is to update this rate based on the current amount of backlog every time that a patients requests to reschedule an appointment.

With these minimum and maximum rates, the rescheduling function can be obtained. It can be found plotted in Figure 2.

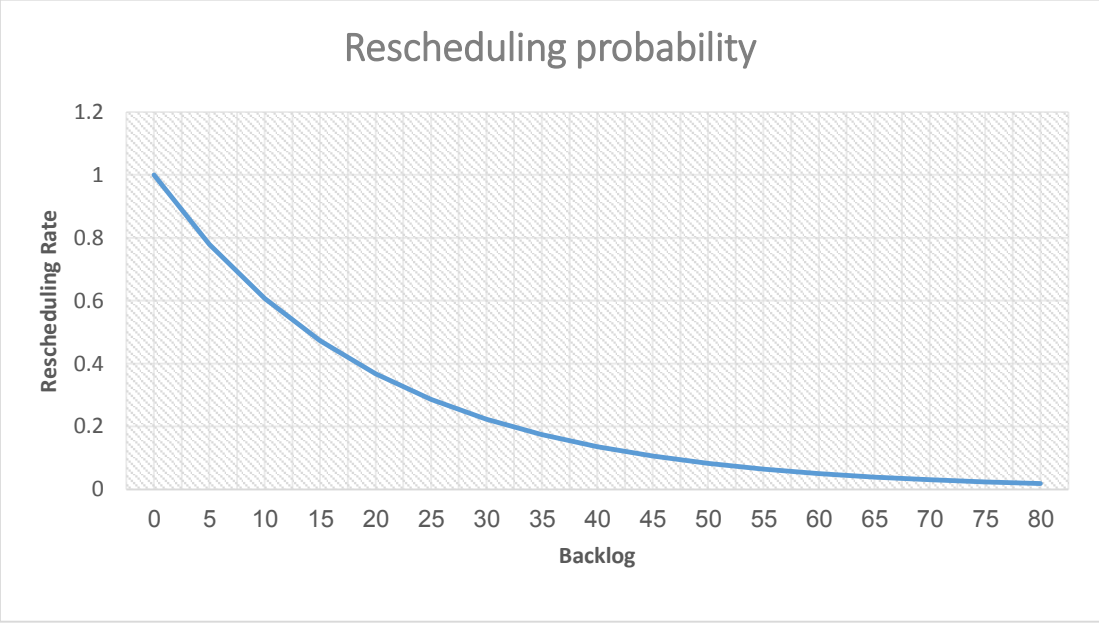

Figure 2. Rescheduling rate based on the amount of backlog

On the basis of above figure, the obtained function is:

$$r(k) = 0 + (1 - 0)\, e^{-k/30} = e^{-k/30} \tag{3.31}$$

The calculated rescheduling function parameter is 30. Thus, this exponential function is used to update the rescheduling rate in each situation based on the current amount of backlog in the system.

3.4. The Composed Model

As mentioned previously, the purpose of this project is to extend the Green and Savin (2008) model to represent the features of our appointment system. The two main features that will be discussed are: 1) considering the multi-comorbidity patients and, 2) updating the rescheduling rate at each entrance. The proposed model is a multi-class queuing network considering no-show rate and rescheduling rate based on the amount of backlog. Green and Savin (2008) used the stationary backlog distribution of an M/D/1/K model to obtain the optimal panel size. We alter the M/D/1/K model to a network to consider patients with different number of comorbidities. The arrival rate and service times are calculated using the aforementioned aggregate formulas. The no-show function is exactly the one that Green and Savin (2008) introduced. However for rescheduling rate the exponential function is used. The stationary backlog distribution is used to estimate the panel size. Based on these extensions we update the stationary backlog distribution according to the following equations:

$$\pi(k) = \frac{(1 - r(K)\gamma(K))\, f(k)}{1 - r(K)\gamma(K) + \rho\left(\sum_{i=0}^{K-1} f(i)\right) - \sum_{i=1}^{K-1} (\gamma(K) - \gamma(i))(r(K) - r(i))\, f(i)}, \tag{3.32}$$

$k = 1, \ldots, K-1$

$$\pi(K) = 1 - \frac{(1 - r(K)\gamma(K))\left(\sum_{i=0}^{K-1} f(i)\right)}{1 - r(K)\gamma(K) + \rho\left(\sum_{i=0}^{K-1} f(i)\right) - \sum_{i=1}^{K-1} (\gamma(K) - \gamma(i))(r(K) - r(i))\, f(i)}. \tag{3.33}$$

For different amounts of k, the above equations gives the probability that the amount of backlog during a large run is k.

To obtain the expected amount of backlog for each panel size point, the following formula is used:

$$\bar{k} = \sum_{i=0}^{K} k.\pi(k) \qquad (3.35)$$

Chapter 4

4. DATA AND RESULTS

The composed model was applied to the data from the Mayo Clinic Primary Care Internal Medicine (PCIM) study. Table 1 shows the rate of arrival based on the number of comorbidities. Based on the experience of Mayo Clinic physicians, Liu et.al (2015) considered a constant service time on the basis of the number of comorbidities. According to this assumption, the service time for patients with 0, 1 or 2 comorbidities is 20 minutes, and for patients with 3 or more comorbidities is 40 minutes. The maximum number of comorbidities considered was 7. The details of the PCIM data are represented in Table 1.

Table 1. Arrival rate and service time of different groups of patients

Number of comorbidities	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7 or more***
Arrival rate per patient per day	0.006	0.011	0.015	0.02	0.026	0.03	0.038	0.041
Average number of patients per day	0.83	2.75	3.69	4.56	4.04	1.81	0.70	0.10
Proportion of each group	0.045	0.148	0.200	0.247	0.218	0.098	0.038	0.006
Mean of Inter arrival time(minutes)	575.12	174.79	129.99	105.18	118.76	264.90	681.06	4680.33
Average of service time (minutes)	20	20	20	40	40	40	40	40

Based on the above table, the average number of patients in each day is obtained by multiplying the arrival rate per patient per day and the total number of unique patients of each group who request for appointment, which is derived from PCIM data. As the number of comorbidities increases the arrival rate per patient per day increases. Therefore patients with 7 or more comorbidities need to visit physician more than others. However since the number of patients in this group is so low, they have less average number of patients in each day.

The mean of service time and inter-arrival times in minutes of the aggregate class are:

$\tau_A = 32.1$

$\lambda_A = 18.03$

As mentioned before, the arrival rate is a Poisson process with mean λN, where λ is the arrival rate per patient and N is the panel size of patients. Using the aggregate arrival rate formula, the rate of arrival per patient per day for the aggregate class, λ_A, can be obtained. In this case $\lambda_A = 0.017$

Therefore $0.017N$ is used as the arrival rate in the model. Testing different values of N, this rate would be diverse in different situations.

The stationary backlog distribution function for composed model was applied to PCIM data, producing an arrival rate of patients of $0.017N$. The stationary backlog distribution was estimated by using varying panel sizes. The results are shown in Figure 3.

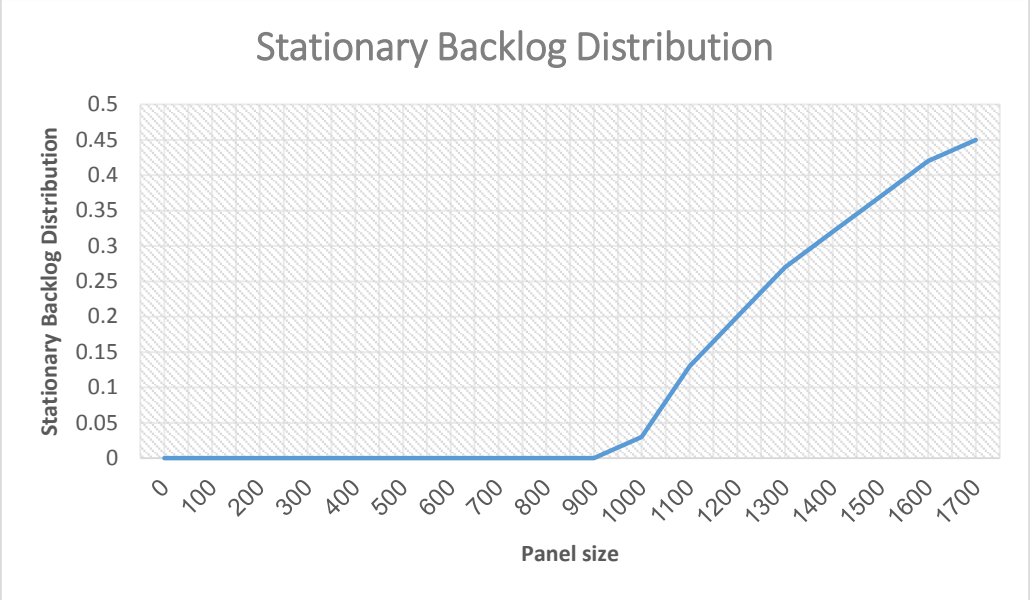

Figure 3. Stationary backlog distribution for different panel sizes

Based on the above figures, by increasing the patient panel size, the stationary backlog probability and expected amount of backlog increase gradually. Around N=1000 and N=1100 the growth rate significantly increased. If we accept less than a 15% possibility to reach the maximum amount of backlog, N=1100 is the optimal panel size.

Figure 4 shows the expected appointment backlogs. Therefore, if we choose N=1100, the expected appointment backlog is 80 spots or almost 5 days (Considering this fact that in each day, each physician meets 18 patients).

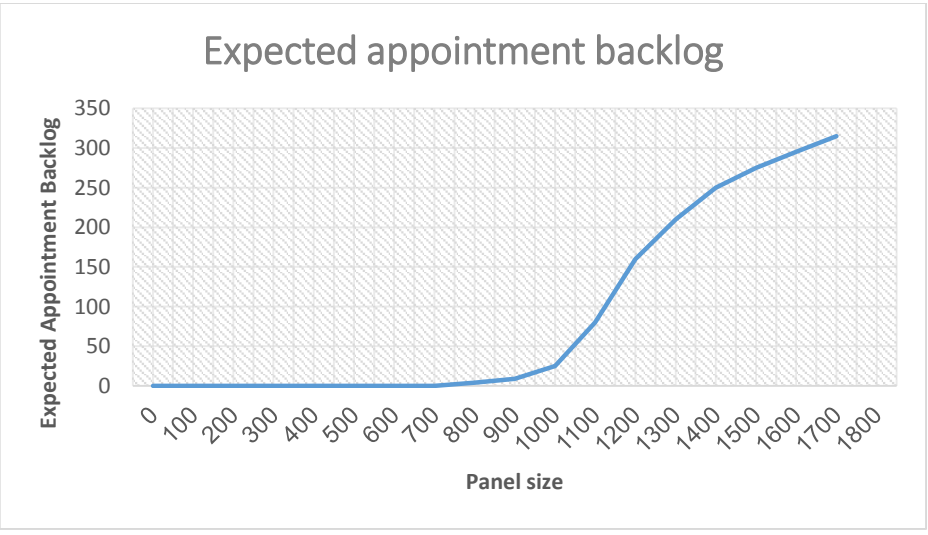

Figure 4. Expected appointment backlog for different panel sizes

Now that the panel size of the system has been estimated, the panel size for each class of patients is needed. Applying the weighted average by using ϕ_r of each group, the panel size of each group was obtained. They are shown in Table 2.

Table 2. Panel size for each group

Number of comorbidities	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***
Panel size of each class	139	250	246	228	155	60	19	3

These numbers are completely dependent on the arrival rate and aggregate service time which is assumed before. For patients with 0, 1, and 2 comorbidities service time is 20 minutes and for ones with more number of comorbidities, it is assume to be 40 minutes. If we change the second assumption from 40 minutes to 30 minutes, the results would be different. By considering this change, the aggregate service time would be $\tau_A = 26$. The following figure is obtained considering new situation.

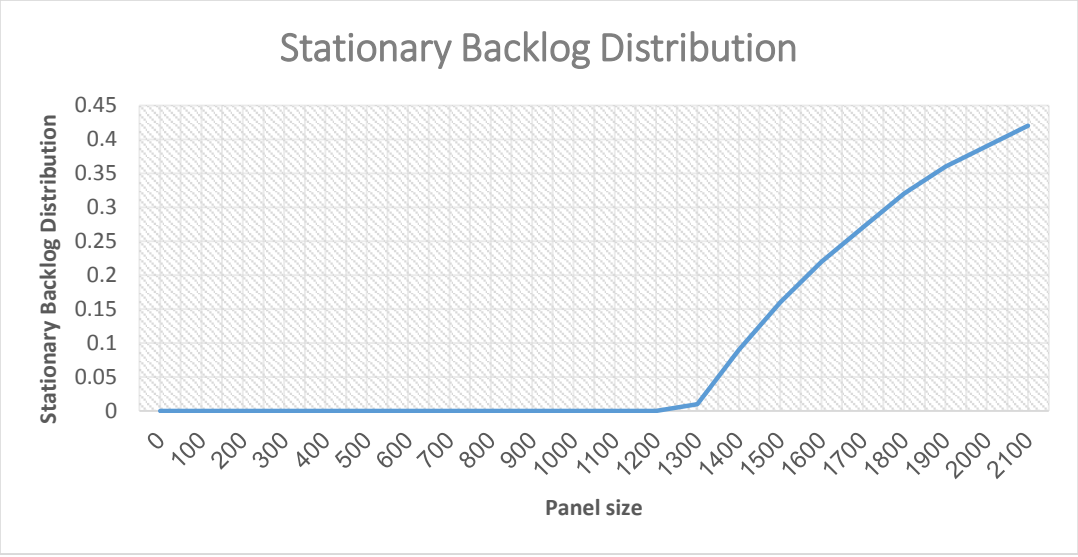

Figure 5. Stationary backlog distribution for new situation

Based on the above figure, the optimum panel size is increased. Considering our previous assumption for choosing optimal panel size, which is having a stationary backlog distribution close to 15%, N=1400 would be acceptable. Therefore, panel size of each group would increase.

Chapter 5

5. SIMULATION

5.1. Methodology

Queuing theory has several applications in healthcare system improvement, but it brings limitations that prevent the proposed models from showing all the features of a real appointment system. In contrast to queuing theory, simulation is more flexible with regards to the features of real systems. One of the primary advantages of simulation models is that they are able to provide users with practical feedback for designing real-world systems (Craig, 1996). Another benefit of simulation is that it permits system designers to study a problem at several different levels of abstraction (Craig, 1996).

For these reasons, simulation was used as the second strategy to investigate the appointment system. In the aforementioned model, events occur during units of time, T. Therefore, discrete event simulation is used to demonstrate the state of the system. The model was coded using the MATLAB software.

The below figure shows an overview of the appointment system.

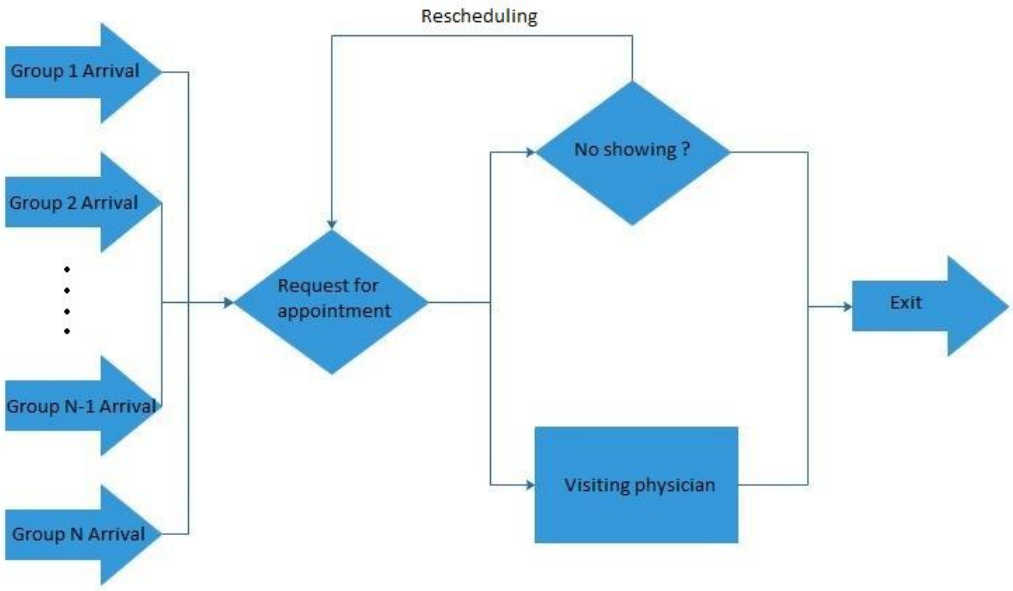

Figure 6. Simulation algorithm

Different groups of patients arrive and request appointments. Patients are categorized based on the number of comorbidities. Each group of patients has a different arrival rate and service time. When a patient requests an appointment, the first available appointment is given to him or her. Some patients may not show up, and may ask to reschedule the appointment. The no-show and rescheduling rates used in the simulation model are the ones obtained based on the amount of backlog, discussed above. Patients do not have any authority to choose their appointment date, and the first available appointment is assigned to them. The state of the system during different units of time is shown with a 3×K matrix, where K is the maximum acceptable backlog. The first row of the matrix shows the presence of patients in the appointment system. Each element of the first row indicates one appointment spot, and is filled by either a 0 or 1. When an appointment position is allocated to a patient, the corresponding element switches to a 1, otherwise it remains 0. Table 3 shows the initial matrix. There are no patients in the system, so all rows are filled with zeroes. When the first patient enters, the first column of the first row is changed to 1. Since there are no patients before the first one, the second row remains 0. Each element of the second row shows the number of patients in the backlog when a patient enters the model. Each spot of the third row is 0 or 1. 1 if a patient is going to be a no-show one and 0 otherwise. Because there is not a no-show patient, the third column is still zero in Table 4.

Table 3. Initial matrix

1	2	…	K-1	K
0	0	…	0	0
0	0	…	0	0
0	0	…	0	0

Table 4. The status matrix after the first arrival

1	2	…	K-1	K
1	0	…	0	0
0	0	…	0	0
0	0	…	0	0

The second step is shown in Table 5. The second patient enters and as a result, the first row of second column is changed to 1. Since there is already one person in the system, the second row is changed to 1, indicating that the number of patients in the backlog is 1. Because of absence of any no-show patients, the third row remains 0.

Table 5. The status matrix after the second arrival

1	2	…	K-1	K
1	1	…	0	0
0	1	…	0	0
0	0	…	0	0

The first time the generated random number is less than or equal to the no-show rate at that point, the first no-show patient arrives. It is assumed that this situation will happen for n^{th} patient, so, the no-show row gets the first 1. In this step the backlog is n-1. This state is shown in Table 6.

Table 6. The status matrix after the nth arrival

1	2	…	n	…	K-1	K
1	1	…	1	…	1	1
0	1	…	n-1	…	0	0
0	0	…	1	…	0	0

This procedure continues until all K available positions are filled up. At this time, no more appointments can be accepted until the next day. At that point, M patients will receive service and M available positions will appear in appointment system. The patients who could not get appointments are called no-service patients. The matrix for the final step is as follows, considering that there is a no-show patient in the K^{th} position.

Table 7. Final matrix

1	2	…	n	…	K-1	K
1	1	…	1	…	1	1
0	1	…	n-1	…	K-1	K
0	0	…	1	…	0	1

At beginning of each day, the matrix is updated. The first M columns are removed, where M is the number of appointment for each day. The remaining patients are brought forward. This results in M empty spots plus the previous number of empty spots being available for new requests.

5.2. Preference model

The previous model is based on the assumption that patients always prefer the earliest available appointment. However, Murray and Tantau (2000), proved that 25% of patients who are offered same day appointment reject it in favor of later appointment. Although, no specific data which shows patient preferences could be found, this assumption and Green and Savin's (2008) supposition were accepted. Based on those findings, 75% of patients prefer the first available appointment while the remaining 25% have preferences with a uniform distribution over all available appointment times. The second model considers patients' preferences, and is thus called the "preference model". The main procedure is exactly the same as previous one; with an increasing backlog, the no- show rate increases while the rescheduling rate decreases.

5.3. Results

5.3.1. First available model

The first scenario which is considered is the model with first available appointment option. PCIM data, which was used in the previous queuing theory section, were also applied to this model. The obtained panel size for each class and its arrival rate and service time were used. Instead of using the aggregate arrival rate and service time, which are applied in multi-class networks, each class's arrival rate and service time were used. This is shown in following table:

Table 8. Arrival rate and service time for different groups

# of comorbidities	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***
Average # of patients per day	0.83	2.75	3.69	4.56	4.04	1.81	0.70	0.10
Mean of Inter arrival time(minutes)	575.12	174.79	129.99	105.18	118.76	264.90	681.06	4680.33
Average of service time(minutes)	20	20	20	40	40	40	40	40

After running the model for 260 days (considering that a year has 260 working days in the US) with 100 replications, the number of patients who could visit the physician, no-show patients, rescheduling patients and the patients who could not obtain an appointment have been counted. These results are shown in the following tables.

Table 9 represents the number of patients in each class who get a chance to visit the doctor during a year.

Table 9. Number patients in each group who visited the physician in a year

# of comorbidities	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***
# of patients who visited physician in one year	209	683	921	1135	1007	449	176	25

The Total number of patients who had a chance to visit the physician is 4,604. Among these patients, 184 of them are patients who rescheduled their appointments after not showing up for the first time. The following table shows the number of no-show and rescheduling patients for the year:

Table 10. Number of no-showing and rescheduling patients

No-showing Patients	***Rescheduling Patients***
314	100

When the number of patients in the system reaches its capacity, K, no more patients can be accepted. The purpose of the proposed model is to decrease the number of no-service patients as much as possible. In this model, the number of no-service patients is 124.

Next, the average backlog during these 260 days with 100 replications was calculated. It was found to be 6 days. In the queuing theory section the expected amount of backlog was calculated for different panel sizes. For the selected panel size of N=1100, it was50 appointment spots, or (almost 5 days). These two numbers are close. Thus the simulation qualifies the proposed queuing model.

Table 11. Expected number of backlog using simulation and queuing theory

Expected amount of backlog obtained by queuing theory for N=1100	***Average amount of backlog obtained by simulation***
5 days	6 days

In the following the utilization of the physician can be found using the following formula.

$$Utilization = \frac{Total\ number\ of\ patients\ who\ visit\ the\ physician}{Total\ demand\ for\ appointments} \quad (5.1)$$

$$Utilization = \frac{4604}{4604 + 124} = 0.973 \quad (5.2)$$

Therefore based on the simulation results the utilization of the physician is almost 97% which is completely acceptable.

5.3.2. Preference model

The second simulated model in this study is the preference model, previously explained. After running the preference model, for 260 days and 100 replication of each year, the following results were found:

Table 12 shows the number of patients in each group who could visit the physician in one year.

Table 12. Number of patients in each group who visited the physician in preference model in one year

# of comorbidities	***0***	***1***	***2***	***3***	***4***	***5***	***6***	***7***
# of patients who visited physician in one year	205	680	910	1128	1000	449	174	25

The total number of patients who visited the doctor was 4,571. The number of no-show and rescheduling patients is shown in Table 13.

Table 13. Number of no-show and rescheduling patients in preference model

Number of No-showing Patients	***Number of Re-scheduling Patients***
308	118

The average amount of backlog during these 100 replications was found to be 6 days. This number is in agreement with the amount obtained by the first model.

The utilization of the physician was also calculated:

$$Utilization = \frac{4571}{4571+152} = 0.967 \qquad (5.3)$$

The utilization is not significantly different between the two models. This means that giving authority to approximately 25% of patients to choose among all available appointment times does not affect significantly the utilization of the system and therefore this model is preferred.

5.4. CONCLUSION

The first section of this study, focused on obtaining an optimal number for a clinical panel size for each physician with the goal of having less waiting time for patients. Patients are categorized into different groups based on their number of comorbidities. Then, according to the expected backlog for each panel size, the optimal panel for each group is estimated.

In the second section, using the calculated panel sizes from the queuing model approach, the appointment model is simulated. Two different scenarios are assumed. The first one is based on the queuing theory's assumption, which says patients always get the first available appointment. The second scnario, or preference model, gives 25% of patients an opportunity to choose between all available appointments. The expected appointment backlog for both models are obtained which close to the number that is obtained using queuing theory. At the end, the utilization of the physician is derived for both scenarios. The results represent that giving authority to patients to choose their appointment will not change the utilization. The application of categorizing patients to different groups based on the number of comorbidities is that the expected panel size of each group would be obtained, therefore in allocating appointments to patients, these number would be considered.

In this research the service time of patients is assumed to be constant, however in future studies, an exponential distribution can be assigned. Also, the used no-show rate is based on the amount of backlog, and is identical for different groups of patients. However, a different function based on the number of comorbidity could be considered.

REFERENCES

(DCD), D. C. (2015, October 2). *http://www.hhs.gov/*. Retrieved from http://www.hhs.gov/answers/medicare-and-medicaid/what-is-the-difference-between-medicare-medicaid/index.html

(2010, September). Retrieved from www.drugabuse.gov.

Altscbuler, J., Margolius, D., Bodenbeimer, T., & Grumbacb, K. (2012). Estimating a Reasonable Patient Panel Size for Primary Care Physicians With Team-Based Task Delegation. *The Annals of Family Medicine*, 396-400.

Banks, J. (1998). *Handbook of Simulation: Principles, Methodology, Advances, Applications, and Practice.*

Baynat, B., & Dallery, Y. (1993). A unified view of product-form approximation techniques for general closed queueing networks. *Performance Evaluation*, 18(3): 205-224.

Bertsimas, Dimitris, Paschalidis, I. C., & Tsitsiklis, J. N. (1994). Optimization of multiclass queueing networks: Polyhedral and nonlinear characterizations of achievable performance. *The Annals of Applied Probability*, 43-75.

Bitran, G. R., & Tirupati, D. (1988). Multiproduct queueing networks with deterministic routing: Decomposition approach and the notion of interference. *Management Science*, 34(1): 75-100.

Cayirli, T., Veral, E. A., & Rosen, H. (2006). Designing Appointment Scheduling Systems for Ambulatory Care Services. *Health Care Management Science*, 9: 47-58.

Craig, C. D. (1996). *Extensible Hierarchical Object-Oriented Logic Simulation with an Adaptable Graphical User Interface.* Department of Computer Science Memorial University of Newfoundland .

Defife, J., Conklin, C. Z., Smith, J. M., & Poole, J. (2010). Psychotherapy appointment no-shows: Rates and reasons. Psychotherapy. *Psychotherapy Theory Research Practice Training*, 413-417.

Dreiher, J., Froimovici, M., Bibi, Y., Vardy, D. A., Cicurel, A., & Cohen, A. D. (2008). Nonattendance in obstetrics and gynecology patients. *Gynecologic and Obstetric Investigation*, 66(1) 40–43.

Evans, G. G., Gor, T. B., & Unger, E. (1996). A simulation model for evaluating personnel schedules in a hospital emergency department. *In Proceedings of the 28th conference on Winter simulation* (pp. 1205-1209). IEEE Computer Society.

Feinstein, A. R. (1970). The pre-therapeutic classification of co-morbidity in chronic disease. *Chronic diseases*, 23(7): 455-468.

Fortin, M., Bravo, G., Hudon, C., Vanasse, A., & Lapointe, L. (2005). Prevalence of multimorbidity among adults seen in family practice. *The Annals of Family Medicine*, 3(3): 223-228.

Gallucci, G., Swartz, W., & Hackerman, F. (2005). Impact of the Wait for an Initial Appointment on the Rate of Kept Appointments at a Mental Health Center. *Institute for Operations Research and the Management Sciences (INFORMS)*, 56: 344–346.

Garcia, J.-M., Brun, O., & Gauchard, D. (2002). Transient Analytical Solution of M/D/1/N Queues. *Applied probability*, 39(4): 853-864.

Green, L. V., & Savin, S. (2008). Reducing Delays for Medical Appointments: A Queueing Approach. *OPERATIONS RESEARCH*, 56 (6): 1526-1538.

Hamrock, E., Parks, J., Scheulen, J., & Bradbury, F. J. (2013). Discrete event simulation for healthcare organizations: a tool for decision making. *Healthcare Management*, 58(2): 110.

Hashimoto, F., & Bell, S. (1996). Improving Outpatient Clinic Staffing and Scheduling with computer simulation. *General internal medicine*, 11(3): 182-184.

Hawkins, M. (2011). *2009 Survey of Physician Appointment Wait Times* . Available at http://www.merritthawkins.com/pdf/mha2009waittimesurvey.pdf.

Ho, C.-J., & Lau, H.-S. (1992). Minimizing Total Cost in Scheduling Outpatient Appointments. *Management Science*, 38(12): 1750–1764.

Ho, C.-J., & Lau, H.-S. (1999). Evaluating the Impact of Operating Conditions on the Performance of Appointment. *European Journal of Operational Research*, 112(3): 542–553.

Kaandorp, G. C., & Koole, G. (2007). Optimal outpatient appointment scheduling. *Health Care Management Science*, 10(3): 217-229.

Kesling, B. (2014). 'Serious conditions' at Phoenix veterans affairs office, watchdog says. *Wall street* .

Klein, E. (2012, June 24). *www.washingtonpost.com*. Retrieved from https://www.washingtonpost.com/news/wonk/wp/2012/06/24/11-facts-about-the-affordable-care-act/

Lacy, N., Paulman, A., Reuter, M. D., & Lovejoy, B. (2004). Why We Don't Come: Patient Perceptions on No-Shows. *National Center for Biotechnology Information*, 2(6): 541-545.

Liu, N., & Ziya, S. (2013). Panel Size and Overbooking Decisions for Appointment-based Services under Patient No-shows. *Production and Operations Management*, 23(12): 2209-2223.

Mercer, A. (1960). A queueing problem in which the arrival times of the customers are scheduled. *The Royal Statistical Society*, 108-113.

Mercer, A. (1973). Queues with scheduled arrivals: A correction, simplification and extension. *The Royal Statistical Society*, 104-116.

Moore, C. G., Wilson-Witherspoon, P., & Probst, J. C. (2001). Time and money: effects of no-shows at a family practice residency clinic. *Family Medicine*, 33(7): 522-527.

Pear, R. (2014, December 27). *www.nytimes.com*. Retrieved from http://www.nytimes.com/2014/12/28/us/obamacare-medicaid-fee-increases-expiring.html?_r=0

Pesata, V., Pallija, G., & Webb, A. A. (1999). A descriptive study of missed appointments: families' perceptions of barriers to care. *National Center for Biotechnology Information*, 13(4): 178-182.

Raunak, M. S., & Osterweil, L. J. (2005). Effective Resource Allocation for Process Simulation: A Position Paper. *ProSim'05*, 14: 175.

Robinson, L. W., & Chen, R. R. (2010). A Comparison of Traditional and Open-Access Policies for Appointment Scheduling. *MANUFACTURING & SERVICE OPERATIONS MANAGEMENT*, 12(2): 330–346.

Sahner, R. A., Trivedi, K., & Puliafito, A. (2012). *Performance and reliability analysis of computer systems: an example-based approach using the SHARPE software package*. Springer Science & Business Media.

Satyam, K., Krishnamurthy, A., & Kamath, M. (2013). Solving generalmulti-class closed queuing networks using parametric decomposition. *Computers & Operations Research*, 40(7): 1777-1789.

Schmalzried, H. D., & Liszak, J. (2012). A Model Program to Reduce Patient Failure to Keep Scheduled Medical Appointments. *Community Health*, 37: 715-718.

Somashekhar, S. (2014, May 30). *www.washingtonpost.com*. Retrieved from https://www.washingtonpost.com/national/health-science/some-of-the-internal-problems-that-led-to-va-health-system-scandal/2014/05/30/399095b4-e81e-11e3-8f90-73e071f3d637_story.html

Tuso, P. J., Murtishaw, K., & Tadros, W. (1999). The Easy Access Program: A Way to Reduce Patient No-Show Rate, Decrease Add-Ons to Primary Care Schedules, and Improve Patient Satisfaction. *The Permanente*, 3(3): 68-71.

U.S.VA. (2015, October 15). *www.va.gov*. Retrieved from http://www.va.gov/health/aca/

Wang, W.-Y., & Gupta, D. (2011). Adaptive Appointment Systems with Patient Preferences. *MANUFACTURING & SERVICE OPERATIONS MANAGEMENT*, 13(3): 373–389.

Whitt, W. (1984). Open and closed models for networks of queues. *AT&T Bell Laboratories Technical Journal*, 63(9), 1911-1979.

Zeng, B., Zhao, H., & Lawley, M. (2008). Primary-Care Clinic Overbooking and Its Impact on Patient No-shows.