International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Forecasting Volume of Patients in the Queue Using Monte Carlo Simulation Model

Mr. Amos Langat*

Jomo kenyatta university of agriculture and technology po. Box. 62000-00200, nairobi, kenya

Email: moskiplangat@gmail.com

Abstract

Healthcare is essential to the general welfare of society. It provides for the prevention, treatment, and management of illness and the preservation of mental and physical well-being through the services offered by medical and allied health professions. Hospitals crowding causes a series of negative effects, e.g. medical errors, poor patient treatment and general patient dissatisfaction. In light of these challenges, a need for review and reform of our healthcare practices has become apparent. One road to improve the typical clinical system is to describe the patient flow in a model of the system and how the system is constrained by available equipment, beds and personnel. Various predictive control models have been developed to try and ease overcrowding in hospitals. Such model is the Model Predictive Control to control the queuing systems developed by Yang Wang and Stephen Boyd. The problem with this model is that it is very slow, and thus not very effective. Others are queuing systems, e.g. Lagrange approach of adaptive control based on Markov Chain model. In this study the research has compared the existing prediction models and come up with Monte Carlo Simulation model to forecasting the volume of patients in the queue. The model uses Poisson distribution on arrival and exponential distribution on service time. The R program was used to run the data where after running, it generate random numbers. After several experiments the model has proved to be very accurate and efficient. This will assist the hospital to utilize the resources and reduces cost of operations.

Keywords: Forecasting; Monte Carlo Simulation; Poisson Distribution.

---

* Corresponding author

1. Introduction

Healthcare is essential to the general welfare of society. It provides for the prevention, treatment, and management of illness and the preservation of mental and physical well-being through the services offered by medical and allied health professions [1,2]. Today, the issue of healthcare is receiving much attention through the media and politics. Healthcare is faced with unprecedented challenges, such as staffing shortages [3,4,5] an aging population [6], rising costs [7,8] and inefficient hospital processes [9]. In light of these challenges, a need for review and reform of our healthcare practices has become apparent. Lean is one way in which overcrowding in Emergency department can be addressed. The basic concept of lean is using less to do more [10]. For healthcare, in particular, one can apply the principles of lean thinking to improve such processes as patient wait time, levels of staffing, and quality of care. Improvements to such processes can greatly impact the health of the community.

[11] States that patient flow can be considered as the movement of patients through a set of locations in a healthcare facility. The patient flow can be considered as a combination of physical flow, information flow and decision flow. Physical flow is the flow of all the existing materials such as patients, test/treatment materials, or caretakers and examples include patient pathway, transport of the blood, or the flow of caretakers. Information flow is the information about the patients and the states in different departments, such as the test results, the occupancy of beds, waiting lists of operation departments, numbers of doctors and nurses who are available, etc. Decision flow depends on the diagnosis of the patient and the state in the hospital. Sometimes, decision flow can be a part of information flow.

The burden on the provision of services can be reduced by model predictive control which will be forecasting the volume of patients in the queue. Hospital administrators can also use model predictive control to forecast the future resident patients' volume in each department/ward.

1.1 Problem Statement

The national budget for the health sector is always criticized for falling short of demand. Hospital cost consumes a significant amount in the national budgets. Patients especially in the public health facilities suffer from the effects of overcrowding. These include lack of vacant beds and caretakers. Congestion is a major headache both to the patients, administrators and health workers who are always overwhelmed by the number of patients [11]. The eight chief wastes in healthcare are unnecessary services or overproduction, mistakes or defects, delays or waiting, unnecessary motion or movement, over-processing, excess inventory, excess transport, unused creativity [12]. Emergency Department [ED] and Inpatients units (IU) crowding causes a series of negative effects such as medical errors, poor patient outcomes and patient dissatisfaction [13]. Patient satisfaction, staff satisfaction, and hospital revenue are all negatively impacted when patients, information, and materials do not move through hospitals in a timely and efficient way.

According to Health, for every dollar spent on healthcare, over 75 cents is spent on the non-patient care activities of communicating, scheduling, coordinating, supervising, and documenting care [13]. This indicates

that there is a huge amount of activity that is not essential to the needs of the patient and illustrates a great potential for improvement to healthcare operations.

Various predictive control models have been developed to try and ease overcrowding in hospitals. Such model is the Model Predictive Control model to control the queuing systems developed by Yang Wang and Stephen Boyd [13]. The problem with this model is that it is very slow, and thus not very effective.

This research seeks to improve on predictive model for forecasting volume of patients in the queue using Monte Carlo simulation model. Simulation provides a model of a real life process, or series of processes, that can be used to see how entities flow through a system. Changes can then be made to the process and the effects can be seen without the commitment of any physical resources or interruption of the system [14].

1.2. Proposed Solution

In this study the research has compared the existing models and come up with Monte Carlo Simulation methods to predict the number of patients in the queue. The simulation model has been chosen in preference to other models because

(i) It allows probabilistic Results - Results show not only what could happen, but how likely each outcome is.
(ii) Sensitivity Analysis - It is easy to see which inputs had the biggest effect on bottom-line results.
(iii) Correlation of Inputs - It is possible to model interdependent relationships between input variables.
(iv) Graphical Results - Because of the data it generates it is easy to create graphs of different outcomes and their chances to occurrence which is important for communication findings to other stakeholders.

1.3. Justification

With Kenya's population growing at a rate of 3 percent annually, the population will continue to place a huge demand for health services. Kenya must continue expanding maternal and child health services while developing the capacity of the health systems to cater for communicable and non-communicable diseases which are on the rise. The Government has committed itself to improving the health Sector infrastructure. Attaining acceptable standards and norms has implications for staffing, equipment, infrastructure, and operating costs [15]. This research has gone a long way in helping achieve vision 2030 by helping in cutting costs in health facilities and patient's time and money.

1.4. Research Questions

The research aims at enhancing models for predicting number of patients in the queue. The main research questions that will guide this research are:

1. Which predictive models are in place and are used by hospitals today?
2. What are difficulties and challenges encountered by hospital staff when using the existing predictive

models?

3. How can the existing predictive models be improved to enable the hospital predict the number of patients in the queue?
4. Can Monte Carlo Simulation model be useful in the hospital to utilize the resources and reduce cost of operations?

1.5. Objective

1.5.1. General Objective

The main goal is to develop queuing forecasting model by studying and comparing existing predictive models in the hospitals.

1.5.2. Specific Objectives

1. To study forecasting models used by hospitals.
2. To identify the short comings associated with the queuing forecasting model.
3. To develop a forecasting model to enable the hospital forecast volume of patients in the queue.
4. To propose forecasting model that will utilize the resources and reduce cost of operations in the hospitals.

1.5.3. Scope of study

1. The objective is to study the forecasting models.
2. The research covers queue models with probabilistic input in a dynamic system.
3. The research uses First-in first-out (FIFO) queuing discipline.

2. Literature review

The chapter discusses the background and problems of overcrowding which could lead to inadequate patient processing.

In the year 2008, use of hospital grew at roughly twice the rate of population growth. With more patients seeking care and fewer inpatient beds available for those who need one, hospitals grew crowded with admitted patients who could not be transitioned to inpatient care [16].

Many hospitals across the country are crowded. Nearly half of hospitals report operating at or above capacity, and 9 out of 10 hospitals report holding or “boarding” admitted patients while they await inpatient beds [17,18].

Across the United States, hospital overcrowding is a significant problem for all types of health care organizations. Overcrowding has become so bad that more than six out of every ten hospitals across the country are operating at or over capacity [20].

Overcrowding is not limited to hospitals either with certain features or in certain areas: overcrowding harms

hospitals in academic, county and private hospitals alike, regardless of whether they are in urban and rural areas [21,22]

Because of the increased capacity at all levels of the health care delivery system, there has been increased pressure for tighter financial management and efficiency. Hospitals are faced with reduced flexibility and ability to accommodate the variations in demand [23].

When the hospitals becomes over-crowded, staff tends to reprioritize patient needs. Typically staff will address the patient's higher level needs because there is neither time, space, nor equipment to address the lower level needs [24]. For example, when the hospital is not at full capacity, a nurse has the time to provide patient education, explain written discharge instructions, and answer any question the patient might have. This is to ensure the patient is well informed of their illness and is aware of what s/he will need to do upon returning home. However, when the hospital is crowded the nurse may only have the time to give the patient written instructions and forego the explanation. The patient's depth of understanding is then compromised and the patient will probably end back up in the hospital for medical treatment [25,26].

In the past, queueing theory has been effectively used in such areas of health care modeling as staff scheduling, policy making (for example, determining how prioritizing certain groups of patients affects wait times), bed requirement analysis and patients waiting in the queue, which is the focus of this thesis.

It is common practice in health services to estimate the required number of beds as the average number of daily admissions times average length of stay in days and divided by average bed occupancy rate (average number of occupied beds during a day) [27,28,29]

$$\text{Bed requirement} = \frac{Average\ number\ of\ daily\ admission}{Average\ bed\ occupancy\ rate} * average\ length\ of\ stay$$

However, as de Bruin et al. mention in [30] "a model, only based on average numbers, is not capable of describing the complexity and dynamics of the in-patient flow."

More recently, queueing models have provided better means of estimating the necessary number of beds based on sound performance measures. Use the M/G/∞ queue as a model for the casualty ward of a hospital. They show that in steady state, the bed occupancy rate follows a Poisson distribution with mean λW, where λ denotes the daily admission rate and W denotes the average duration of stay. Using this model, the author determine the required number of beds in order to guarantee that a given target percentage of arrivals receive a bed immediately.

M/G/∞ system is also used to model the queue of patients needing alternative levels of care in acute care facilities whose treatment is completed and who are waiting to be transferred to an extended care facility (ECF). These patients are kept in the hospital due to unavailability of beds in the ECF and reduce the hospital utilization [31]. The author's model allows managers to forecast the effect of certain policy changes on appropriate access measures. For instance, the cost-benefit trade-off of opening an additional extended care facility within a region

is compared to that of assigning a higher priority to patients going to ECF from acute care facilities than to those coming from other sources. Instead of using an infinite capacity queue, [32] uses an M/G/c queue with a state-dependent arrival rate to address the long hospital-wait list problem. He experiments with various management actions such as increasing the number of beds or decreasing mean service times through appropriate means.

Queueing model was developed for the movement of patients through a hospital department [33]. Performance measures, such as mean bed occupancy and the probability of rejecting an arriving patient due to hospital overcrowding, are computed. These quantities enable hospital managers to determine the number of beds needed in order to keep the fraction of delays under a threshold, and also to optimize the average cost per day by balancing the costs of empty beds against those of delayed patients.

Although service times, unlike inter-arrival times, do not usually have an exponential distribution, such an assumption is often made in order to simplify the analysis greatly. For

instance, [34] use the M/M/c/c queue, referred to as the Erlang Loss model, to investigate the emergency in-patient flow of cardiac patients in a university medical centre in order to determine the optimal bed allocation so as to keep the fraction of refused admissions under a target limit. The authors find the relation between the size of a hospital unit, occupancy rate, and target admission rates. A cancellation rate of 5% is often considered acceptable. However, while the target occupancy rate of 85% has become a golden standard in health care [35] the authors note that using one target occupancy rate for hospital units of different size is not reasonable, for larger hospitals can usually operate at a higher occupancy rate than smaller ones.

After analytically estimating the required number of beds in the First Cardiac Aid (FCA) unit of the medical centre, [34] also use numerical methods to determine the number of beds in the Coronary Care Unit (CCU) and the Normal Care clinical ward (NC). The authors had to rely on numerical techniques at this stage, because the finite capacity of the CCU and the NC leads to blocking in the FCA, making analytical calculations extremely difficult to carry out.

In fact, due to the complexities that arise in analyzing queueing systems with multiple interacting service stations, the study of health care facilities has mainly been done using simulation, with analytical methods applied to the study of one hospital as a whole (represented by a single service station) or of single hospital units, assumed to operate independently of the others. In recent years, however, approximate analytical methods have been developed and used in studying multi-facility interactions.

Another queueing network model applied to a hospital setting is that of [36] who study a specific obstetrics hospital consisting of 8 subunits with 4 different patient arrival streams. The transfer of patients between the different compartments creates blocking in some of the units. Building their simulation model, the authors first use an approximate analysis of the network by ignoring blocking and time dependence of the parameters. This helps to provide quick answers to many of the management questions, in addition to guiding them in validation of their simulation in special circumstances. By using discrete-event simulation to model the full interaction of the different subunits and patient groups, the author then compare alternative methods of reducing blocking

times and increasing the hospital throughput. For example, after identifying the Post-Partum unit as the bottleneck of the system, they show that by adding new beds to this unit or reallocating beds from underutilized units, such as Medicine/Surgery, the maximum number of deliveries per month can be increased; however, in the latter case, reallocating beds beyond a certain threshold causes the bottleneck to shift to a different unit. An interesting result is that increasing the number of beds in the bottleneck unit by 15% yields a 38% improvement in the overall hospital throughput.

As can be seen, the application of queueing models to healthcare is growing more popular as hospital management teams are gaining awareness of the advantages of these operational research techniques in addressing such issues as determining optimal bed counts and making policy decisions with regards to resource allocation. Research in applying queueing networks with blocking is rare in the literature due to the mathematical complexities involved in computing performance measures associated with such systems. As a result, hospitals with interacting subunits are often studied through simulations, for they are able to incorporate much more detail than is affordable by analytical methods.

2.1. Existing model to predict patients in the queue

2.1.1 Stochastic Model

Monte Carlo simulation falls under category of stochastic simulation model where it describes events or systems that are unpredictable due to the influence of a random variable. The research reviews some of the models that fall under stochastic. Stochastic simulation model is a model of having at least same random input components. Most queuing and inventory system are modeled stochastically. Stochastic simulation models produce output that is itself random and must therefore be treated as only an estimate of the true characteristics of the model which leads to main disadvantage of the model. A stochastic process is a collection of similar random variables ordered over time, which are all defined on a common sample space. The set of all possible values that these random variables can take on is called the state space. If the collection is $X_1, X_2 \ldots$, then we have a discrete-time stochastic process. If the collection is $\{X(t), t \geq 0\}$, then we have a continuous-time stochastic process [37]

A stochastic process is a family of random variables X where t is a parameter running over a suitable index set T. (Where convenient, we will write X(t) instead of X,.) In a common situation, the index t corresponds to discrete units of time, and the index set is T = {0, 1, 2, . . .}. In this case, X, might represent the outcomes at successive tosses of a coin, repeated responses of a subject in a learning experiment, or successive observations of some characteristics of a certain population. Stochastic processes for which T = [0, c) are particularly important in applications.

Here t often represents time, but different situations also frequently arise. For example, t may represent distance from an arbitrary origin, and X, may count the number of defects in the interval (0, t] along a thread, or the number of cars in the interval (0, t] along a highway.

Stochastic processes are distinguished by their state space, or the range of possible values for the random

variables X by their index set T, and by the dependence relations among the random variables X, The most widely used classes of stochastic processes are systematically and thoroughly presented for study in the following chapters, along with the mathematical techniques for calculation and analysis that are most useful with these processes. The use of these processes as models is taught by example. Sample applications from many and diverse areas of interest are an integral part of the exposition [38].

2.2.2. Discrete-event simulation and queueing theory

In DES, the operation of a system is represented as a chronological sequence of events. Each event occurs at an instance in time and marks a change of state in the system. The modeled system is dynamic and stochastic. DES includes Clock, Events List, Random Number Generators, Statistics and Ending Condition [39].

For example, in the process that patients wait for a bed in the ward, the system states are queuing length or number of vacant beds. The system events are patients-arrival and patients-departure. The system states, like vacant beds are changed by these events. The random variables that need to be characterized to model this system stochastically are patient arrival time and residence time. To simulate such system, first generate a series of random entities based on the distribution. Let (n,t) be n patients coming into the station at time t. Then all the incoming patients during $(t_1, t_2 = t_1 + dt, t_3 = t_2 + dt, \ldots., t_k)$ can be expressed as $\{(n_1,t_1),(n_2,t_2),\ldots\ldots.,(n_k,t_k t_1)\}$. Here $n_1, n_2 \ldots\ldots n_k$ are random numbers. dt is constant. The simulator generates service rate for each patient, $l_1, l_2 \ldots\ldots l_k$ which are random numbers. All the random numbers obey a certain distribution. The patients leave the ward when the residence time is over. The simulator stores all the data. The patient number and other results can be obtained by analyzing the saved data. Such as to compute the resident patient number at time t_i , the simulator find out the patients that time t_i is between this patients' arrival and departure time.

There are several advantages to build such models [40]. Detailed system behavior can be modeled; It is possible to model the performance, dependability; Less matrices computing.

Also there are some drawbacks compared with other models [30]. Long execution time; Simulation results are difficult to interpret; It is quite likely that some rare events or states are never encountered by the simulation runs.

For example, [40, 41] conduct detailed studies of patient flow in various departments at an Israeli and a US hospital, respectively. Reference [41] do not focus on discharge policies, but they empirically study the transfer process flow from Emergency Department to inpatient wards (which they call internal wards).

Discrete-event simulation and queueing theory are two commonly used approaches for modeling and improving patient flow [42, 43], [44, 45], and [46]. Compared to the rich literature on patient flow models of Emergency Department, inpatient flow management and the interface between Emergency Department and inpatient wards have received less attention. Note that [47] demonstrate that the admission or discharge blocking caused by nurse shortages can have a significant impact on system performance.

2.2.3. Stochastic Simulation and Poisson distribution

Modeled daily bed occupancy variability using stochastic simulation [48]. They found that a flexible bed allocation scheme resulted in fewer overflows with the same level of occupancy compared to a fixed bed allocation. Their model also proves that variable discharge rates are more significant than variable admission rates in contributing to overflows. Numerous investigators have also used regression models to analyses needed bed capacity.

A common statistical assumption in modeling count data is that it follows a Poisson distribution. [49] assumed that the number of patients staying in the OB unit is Poisson distributed.

2.2.4. Weakness of Stochastic Simulation and Poisson distribution

An example of a conclusion from the goodness-of-fit statistical test that is not convincing enough can be found, for instance, in [48]. The authors tried to justify the use of a Poisson process by using a chi-square goodness-of-fit test. The authors obtained the test p -values in the range from 0.136 to 0.802 for different days of the week. Because p -values were greater than 0.05 level of significance, they failed to reject the null-hypothesis of Poisson distribution (accepted the null-hypothesis) [49].

On the other hand, the fundamental property of a Poisson distribution is that its mean value is equal to its variance (squared standard deviation). However, the data indicated that the calculated mean value was not even close to the variance for at least 4 days of the week. Thus, the use of a Poisson distribution was not actually convincingly justified for patient arrivals. Apparently, chi-square test p –values were not large enough to accept the null-hypothesis with high-enough confidence (alternatively, the power of the statistical test was likely too low).

2.2.5. Autoregressive Inductive Moving Average (ARIMA)

Autoregressive Inductive Moving Average (ARIMA) modeling to predict the number of surgical beds required at a UK hospital [50]. Further, McManus describes the use of Queuing Theory to predict monthly responsiveness to changing bed demand [51]. ARMA processes are useful in describing or approximating a wide variety of stationary processes whose auto covariance functions approach zero as the lag approaches infinity.

2.2.6. Weakness of Autoregressive Inductive Moving Average (ARIMA)

Using ARIMA modeling, [52] found that the daily number of occupied beds due to emergency admissions is related to both air temperature and influenza illness rate. It was found that a period of high volatility, indicated by GARCH errors, would result in an increase in waiting time in the A&E department. The model has limitations and especially of the inherent variability of emergency inpatient flow. There have been several methodologies developed for forecasting arrival and census counts in various hospital departments [52] evaluated the use of seasonal autoregressive integrated moving average (ARIMA), time series regression,

exponential smoothing, and artificial neural network models to forecast daily patient volumes in emergency departments at three diverse hospital emergency departments. The time series methods considered in that analysis provided improved absolute prediction error relative to a multiple linear regression approach, considered the benchmark model for forecasting emergency department patient volumes. Additionally, [52] evaluated the use of autoregressive integrated moving average models, adjusted to incorporate various environmental variables, to forecast counts of daily patient attendances in the emergency department of an acute care regional general hospital. The model has limitations and especially of the inherent variability of emergency inpatient flow.

2.2.7. Queuing Theory with Markov Chain (QTMC), and Discrete Event Simulation (DES)

Hidden Markov models (HMMs) have been used in various fields, ranging from Bioinformatics to Storage Workloads [54]. HMMs were first used in the late 1960s in statistical papers by Leonard E. Baum for statistical inference of Markov chains [55] and also for statistical estimation of Markov process probability functions [56]. Speech recognition became a field for training HMMs in the 1970s and 1980s [57, 58], with many such speech models still used today [59].

A hidden Markov model (HMM) is a probabilistic model (a bivariate Markov chain) which encodes information about the evolution of a time series. The HMM consists of a hidden Markov chain {Ct} (where t is an integer) with states not directly observable and a discrete time stochastic process {Ot} $t\geq 0$, which is observable. Combining the two, we get the bivariate Markov chain {(Ct, Ot)} $t\geq 0$.

The first model (QTMC) is only able to consider limited scenarios that can occur. DES has been well recognized in healthcare and is broadly used for the validation of other models. The DES models offer a valuable tool to study the trade-off between the capacity structure, sources of variability and patient flow times [60].

The Hospital arrivals model was found to train successfully on patient arrivals, collected over months of analysis. The means and standard deviations matched well for raw and HMM-generated traces and both traces exhibited little autocorrelation. HMM parameters, fully converged after training, were used to predict the model's own synthetic traces of patient arrivals, therefore behaving as a fluid input model (with its own rates). An enhancement could be to assume instead that the arrival process is Poisson, with corresponding rates, and produce a cumulative distribution function for the patient arrivals workload.

2.2.8 Weakness of Queuing Theory with Markov Chain (QTMC), and Discrete Event Simulation (DES)

The DES plots have more fluctuation between each simulation plots. These unstable properties can be improved by increasing ensemble size, but a large ensemble size of simulations will lead to a higher computation cost. In turn, if multiple patients were admitted simultaneously, queuing analysis would only account for one patient in the model [12].

2.2.9 Stochastic and Simulation Model

A stochastic model is accurately used to represent real world phenomena and processes, particularly in health care and patient monitoring. The model use patient flow through chronic diseases departments. Admissions are modeled as a Poisson process with parameter λ estimated by using the inter-arrival times. The in-patient care time is modelled as a mixed-exponential phase-type distribution. Both geriatricians and hospital administrators agreed that such a model is useful to be applied for optimizing the use of hospital resources in order to improve hospital care [61].

2.2.10 Decision- Tree Model

The DES plots have more fluctuation between each simulation plots. These unstable properties can be improved by increasing ensemble size, but a large ensemble size of simulations will lead to a higher computation cost. In turn, if multiple patients were admitted simultaneously, queuing analysis would only account for one patient in the model [12]. It is not easy to find a proper mathematical model when the process is complex and limitation of the control variable e.g. more efficiently handling the integral variables and the constraints. The model lacked Model Predictive Control theory of practice and should reduce and optimize the matrices computing which are time consuming and take up lots of resources.

2.2.11 Weakness of Decision – tree

The potential limitation is that it did not consider the time-related factors to be potentially correlated with outcomes. It could be difficult to recall the exact time-of-collapse events in emergency situations

2.2.12 Support Vector Machine (SVM) and a Cox Regression based approach

Support vector machine (SVM) based on statistical learning theory has been used generally in machine learning because of its good generalization ability. By using SVM we can classify and identity some probability distributions appeared in queuing system and solve the density function regression problem through using support vector regression (SVR) [62,63]. The framework has the advantage that it can take all variables the hospital has collected for its patient population (possibly with missing values). The model building does not need human involvement (e.g., manual tune-up, heuristics) and is very efficient so one can effectively re-build the model regularly with new patient data (e.g., every night). The use of polynomial kernels or Cox regression did not appear to improve on the performance of the standard linear SVM. The use of Cox regression has the unique advantage that of handling the 30-day readmissions cut-off more robustly, as it does not create arbitrary distinctions between readmissions occurring near the cut-off (e.g. a readmission on the 30^{th} versus the 31st day after a discharge). However, this did not appear to substantially affect performance due to the fact that the great majority of readmissions in our dataset occurred in the first 2-3 weeks of the prior admission.

The foundation of support vector machines (SVMs) has been developed by [64]. SVMs have become in the last few years one of the most popular approaches to learning from examples. Due to many attractive features such as good generalization and anti-noise ability, SVMs have been applied in many fields of science and engineering. The content researched by queuing theory are deducing the performances of queuing system in the

condition of known customer arrival distribution and service time distribution. Since originated by Erlang during his research of telephone communication using probabilistic method in the early 20th century, classical queuing theory has tended to perfection. However some problems appear when applying theory to practice. Queuing theory assumes that customers arrival and service time submit certain prior distributions. How to discriminate these distributions becomes major problem in the practice of queuing system. A.G. Konheim deduced some formula for G1/G/1 queuing system [65], but the results expressions are Laplace transform equations and difficult to solve them. In communication network, cells (customers) arrival has different distributions during different time. The algorithm of SVM based on statistical learning theory solves these problems very well. Train on some samples drawn from customer arrival or service time making use of SVM then establish a system for distribution identification automatically. Experiment shows it works very well [66].

2.2.13 Weakness of Support Vector Machine (SVM) and a Cox Regression based approach

A patient's final LACE score is calculated by summing the points for each attribute. The index was externally validated using administrative data in a random selection of 1,000,000 Ontarians with a reported accuracy of 0.68 in AUC. Since the base readmission rate of the population used for LACE development is around 8%, it is arguably not well suited for the Medicare population in the US, for which the base readmission rate is around 20%. Other models based on general population data include [67].

2.2.14 Forecasting model, Poisson Autoregressive (PAR) model and logistic regression models

In a period of heightened economic burden, efficient and effective allocation of hospital resources is an issue of principal importance. The ability to accurately forecast the number of patient arrivals, as well as predict census counts in hospital departments, have considerable implications for hospital resource allocation, both at the micro and macro level. More importantly, accurate census forecasts can inform scaling up of operations during high census periods, potentially leading to improved patient outcomes [67]. Since staffing levels in hospital units are driven by the census capacity as well as the acuity of unit patients, forecasting methods that incorporate both patient-level severity of illness (which may evolve considerably throughout their stay) and long-term census trends are necessary for informing accurate census predictions. In addition to univariate time series approaches to forecasting emergency department patient volumes, multivariate time series models have also been utilized and have been shown to reliably forecast emergency department patient census.

The ensemble-based method for short-term census forecasts under a framework that simultaneously incorporates (i) hospital unit arrival trends over time and (ii) patient specific baseline and time-varying information. Such approaches represent the future of census forecasting as hospital departments around the country move toward more efficient methods for collecting and processing patient-level information upon admission and through the duration of stay.

2.2.15 Strength of Forecasting model, Poisson Autoregressive (PAR) model and logistic regression models

The model is efficient because it integrates arrival trends over time as well as patient level information. The former is crucial to the development of accurate and reliable models for predicting the probability departure,

while the latter is integral to attainment of a model that can predict the number of census arrivals with a high degree of accuracy.

Our justification for using a conditional logistic regression framework for predicting the number of departures was motivated by two principle issues. As a result of our general forecasting framework, our interest was primarily focused on the expected number of departures for a cohort of patients currently residing in the census. Thus, treating each patient within a cohort as independent, the expected number of departures for a given cohort can be efficiently estimated by summing the individual predictions for departure for each patient. The idea of predicting the probabilities of departure as opposed to length of stay predictions lends itself nicely to a logistic regression framework. An alternative approach involves using length-of-stay distributions within a queuing theory analysis. However, unlike the framework described here, such an approach would not facilitate the attainment of the subject-specific probabilities of departure, which is of interest to clinicians.

2.3 Model in use

In this study the research has compared the existing models and come up with Monte Carlo Simulation methods to forecast the volume of patients in the queue. Monte Carlo Simulation model falls under stochastic model where it describes events or systems that are unpredictable due to the influence of a random variable. From the literature review the existing models are not effective and accurate. They also do not produce optimal results because of the variables they use to compute models.

Monte Carol Simulation is a technique that computes or iterates the project cost or schedule many times using input values selected at random from probability distributions of possible costs or durations, to calculate a distribution of possible total project cost or completion dates state [68]. Monte Carlo simulation samples probability distribution for each system variable to produce hundreds or thousands of possible outcomes.

2.3.1 Application of Monte Carlo Simulation

Monte Carlo Simulation has existed before and has been applied in various fields.

(i) In the field of computer engineering and design, [69] described the use of simulation when optimizing the problem layout of IBM's Blue Gene ® / L supercomputer.
(ii) In geophysical engineering, Monte Carlo analysis has been used to predict slope stability given a variety of factors [70].
(iii) In marine engineering, [71] described a probabilistic methodology they have developed to assess damaged ship survivability based on Monte Carlo simulation.
(iv) Monte Carlo simulation in aerospace engineering to geometrically model an entire spacecraft and its payload, using The Integral Mass Model [72].
(v) In public health, simulation has been used to estimate the direct costs of preventing Type 1 diabetes using nasal insulin if it was to be used as part of a routine healthcare system [73].
(vi) Monte Carlo simulation should be used by research organizations to determine whether or not future possible research is really worth the cost and effort, by modeling possible outcomes of the research

[74].

(vii) Monte Carlo simulation in personal financial planning, especially when estimating how much money one needs for retirement and how much one can spend annually once retirement has begun [75].

2.3.2 Approach of Monte Carlo Simulation

Monte Carlo Simulation has been used in processor performance to predict the Cost Performance Index of in-order architecture and validate it against the Itanium-2 [76]. The research will come up with model design to estimate patients demand in the hospitals which will use arrival time, waiting time and service time. It will also use Poisson rule and exponential distribution to facilitate how Monte Carlo Simulation will work.

In order to facilitate Monte Carlo Simulation, the research will consider the simple multi-server queuing model as *M/M/c/∞*. Suppose arrival time fit Poisson distribution and service time to obey exponential distribution. The research has implemented the Monte Carlo Simulation using R programming.

Poisson distribution

$$= P(x;\lambda)\frac{e^{-\lambda}\lambda^{x}}{x!} \quad \text{for x = 0, 1, 2 …}$$

Exponential distribution

$$= f(x) = \int_0^x \frac{1}{\beta} e^{-t/\beta} dt$$

$$= 1 - e^{-x/\beta}$$

For $x \geq 0, t \geq 0$

2.3.3 Inverse Transform Method

Inverse transform sampling (also known as inversion sampling, the inverse probability integral transform, the inverse transformation method, Smirnov transform, golden rule,) is a basic method for pseudo-random number sampling, i.e. for generating sample numbers at random from any probability distribution given its cumulative distribution function (cdf). Inverse Transform method is a method used to generate random numbers in R programming.

The basic idea is to uniformly sample a number u between 0 and 1, interpreted as a probability, and then return the largest number x from the domain of the distribution $p(X)$ such that $p(-\infty < X < x) \leq u$. For example, imagine that $p(X)$ is the standard normal distribution (i.e. with mean 0, standard deviation 1). Then if we choose $u = 0.5$, we would return 0, because 50% of the probability of a normal distribution

occurs in the region where $X \leq 0$. Similarly, if we choose $u = 0.975$, we would return 1.95996...; if we choose $u = 0.995$, we would return 2.5758...; if we choose $u = 0.999999$, we would return 4.891638...; etc. Essentially, we are randomly choosing a proportion of the area under the curve and returning the number in the domain such that exactly this proportion of the area occurs to the left of that number. Intuitively, we are unlikely to choose a number in the tails because there is very little area in them: We'd have to pick a number very close to 0 or 1.

The inverse transform sampling method works as follows:

1. Generate a random number u from the standard uniform distribution in the interval [0, 1].
2. Compute the value x such that F(x) = u.
3. Take x to be the random number drawn from the distribution described by F.

Suppose we wish to generate random numbers having density h. Let H denote the corresponding cumulative distribution function, and let G = H−1 (inverse in the same sense as square and square root operations are inverses of each other). Set X = G (U), where U has a U (0, 1) distribution. Then

$$F_X(T) = P(X \leq t)$$

$$= P(G(U) \leq t)$$

$$= P(U \leq G^{-1}(t))$$

$$= P(U \leq H(t))$$

= H(t)

3. Research methodology

This chapter presents the description of the methods that were adopted in the study. Here efforts were made to discuss the following:

- Method of Data Collection
- Measure of queue length
- Queueing models Based on the Birth and Death Process

a) Using m/m/s queueing model in measuring system performance in Thika Level 5 Hospital and Kenyatta National Hospital.
b) Formulating Priority – Discipline Queueing Models in Measuring System Performance in Thika Level 5 Hospital and Kenyatta National Hospital.
c) Formulating State – Dependent Service Rate Model in Measuring System Performance in Thika Level 5 Hospital and Kenyatta National Hospital.

- **Method of data collection**

The basic data used for this study is secondary data consisting of recorded information on the arrival times of the patients and the service time for patient. The instruments used for the data collection are recorded sheets from the hospitals.

The data collection span for two days in a week i.e. Monday and Thursday from 8am to 2.30 pm. The days were preferred because on Monday the queue is always long and on Thursday the queue is diminutive. Also most of the times they stop issuing cards at 2.30 pm.

The researcher collected data in the two hospitals because of the number of the patients they handle and the queue they experience.

- **Measures of queue length**

Those measures are:

1) The number of patients waiting in line which is collected after an interval to enable the researcher to construct a model. The researcher decided to collect the data after every 30 minutes. The data are collected from arrival recording book from the consultant receptionist.
2) The time patient is served which is considered constant for all patients and that is five minute per patient.

- **Queuing models based of birth-death process**

Waiting lines (queue) are a direct result of arrival and service variability. They occur because random bunched arrivals and highly variable service patterns cause systems to be temporarily overloaded. In many instances, the variabilities can be described by theoretical Poisson distribution for arrival time and negative exponential distribution for the service time.

In the context of queueing theory, the term birth refers to the arrival of a new patient into queueing system, and death refers to the departure of a served patient.

- **The M/M/S model**

The M/M/c Queue

We now illustrate the ideas introduced in this chapter with the use of an example. Consider the M/M/c queue where the arrival and service rates are λ and μ, respectively. Assuming that steady state exists let p_n be the steady state distribution of the number of units in the system. We proceed to derive the equations involving $p_n$by using the rate-equality principle, which states that the rate at which a process enters a state is equal to the rate at which it leaves that state.

Consider state 0, when there are no units in the system. The process can leave this state only when there is an arrival, which causes the system to transition to state 1. The long-run proportion of time the process is in state 0 is p_0, and since λ is the rate of arrival, the rate at which the process leaves state 0 to go to state 1 is λp_0. Moreover, the process can enter state 0 only from state 1 through a departure or service completion. Since the proportion of time the process is in state 1 is p_1 and the rate of leaving state 1 through service completion is μ, the rate at which the process transitions from state 1 to 0 is μp_1. Using the rate-equality principle, we get

$$\lambda p_0 = \mu p_1 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1$$

Now consider state 0 < n < c. The process can leave state n in two ways, either through an arrival or through a departure. The proportion of time the process is in state n is p_n and the total rate at which the process leaves state n through arrivals or departures is $\lambda p_n + n\mu p_n$ since there are n servers busy (additive property of the Poisson process). The process can enter state n in two ways, either through arrival from state n - 1 or through a departure from state n + 1. Thus, the rate at which the process enters state n is $\lambda p_{n-1} + \mu p_{n+1}$. By the rate-equality principle

$$\lambda p_n + n\mu p_n = \lambda p_{n-1} + (\mathrm{n}+1)\mu p_{n+1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 2$$

Similarly, for the case of n $\geq$ c, we get

$$\lambda p_n + c\mu p_n = \lambda p_{n-1} + \mathrm{c}\mu p_{n+1} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3$$

Repeated application of (2) along with (3) at the last step yields

$$\lambda p_n - (n+1)\mu p_{n+1} = \lambda p_{n-1} - \mathrm{n}\mu p_n.$$

$$= \lambda p_{n-2} - (n-1)\,\mu p_{n-1}.$$

$$\vdots$$

$$= \lambda p_0 = \mu p_1$$

$$= 0.$$

By rearranging terms and iterating we obtain that for 0 < n $\leq$ c

$$p_n = \frac{\lambda/\mu}{n} p_{n-1} = \frac{(\lambda/\mu)^2}{n(n-1)} p_{n-2} = \cdots = \frac{(\lambda/\mu)^n}{n!} p_0 \quad \ldots\ldots\ldots\ldots\ldots\ldots 4$$

In a similar fashion, we get that for n > c

$$p_n = \frac{(\lambda/\mu)^n}{c!c^{n-c}} p_0 \text{.....5}$$

Now for $\lambda/(c\mu) < 1$, the normalization condition $\sum_{n=0}^{\infty} p_n = 1$ gives

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{c!(1-\lambda/c\mu)} + \right] \text{.....6}$$

We now proceed to compute some performance measures. The expected queue length L can be computed as

$$L = \sum_{n=c}^{\infty} (n-c)\, p_n = \frac{\lambda p_c}{\mu(1-p)^2} \text{.....7}$$

Where $p = \lambda/c\mu$ is referred to as the server utilization. Applying formula, we also obtain the expected waiting time in the queue

$$W = \frac{L}{\lambda} = \frac{p_c}{\mu(1-p)^2} \text{.....8}$$

Knowing the probability distribution, we can now directly compute the pgf of the number in the queue

$$P(z) = \sum_{n=0}^{c-1} p_n + \sum_{n=c}^{\infty} p_n Z^{n-c} = 1 - \frac{p_c}{1-\rho} + \frac{p_c}{1-\rho^2} \text{.....9}$$

Which allows us to find the LT of the wait time distribution as

$$w^*(s) = P\left(1 - \frac{s}{\lambda}\right) = 1 - \frac{p_c}{1-\rho} + \frac{p_c}{1-\rho+s/c\mu} \text{.....10}$$

The m/m/s queue is a model with parameters inter-arrival time and the service time that is Poisson and exponentially distributed respectively. The queue discipline here is First – Come, First – Served (FCFS). The space for the waiting line is infinite size.

Consequently, this model is just the special case of the birth-and-death process where the queuing system's mean service rate per busy server is constant.

This information is then used by Health Managers to decide on an appropriate level of service for the facility. The basic objective in most queueing models is to achieve a balance between two costs; cost of offering the service and cost of delay in offering the service.

3.1. Priority – discipline queuing models

In priority – discipline queueing models, the queue discipline is based on a priority system. Thus, the order in which patients of the queue are selected for service is based on their assigned priorities. Many real queueing

systems fit these priority – discipline models much more closely than other available models. Rush jobs are taken ahead of other jobs, and important patients may be given precedence over others. Therefore, the use of priority-discipline models often provides a very welcome refinement over the more usual queueing models. This model incorporates all of the assumptions of the basic multiple – server model and it uses FCFS (first – come, first - served).

3.2. A model with state – dependent service rate and/or arrival rate

All the models thus far have assumed that the mean service rate is always constant, regardless of how many patients are in the system. Unfortunately, this rate often is not a constant in real queueing systems, particular when the servers are people. When there is a backlog of work (i.e. a queue), it is quite likely that such servers will tend to work faster than they do when the backlog is small or none existent. This increase in the service rate may result merely because the servers increase their efforts when they are under the pressure of a long queue. However, it may also result partly because the quality of the service is compromised or because assistance is obtained on certain service phases.

3.3. Description of the patient

The patient arrival at the hospital at random and the queue discipline is first – come, first – served. The hospital under this study operates 24 hours and consultation is open for all patients that are appointment patients, same day appointment patients (walk-ins) and new patients. All patients have to go to the reception desk for submission of their hospital card and if necessary, for initial screening before consultant. Monte Carlo Simulation method was successfully used to describe the complexity and dynamics of patients flow.

3.4. Software used in modeling

The research has used R programming as it produce well-designed publication-quality plots, including mathematical symbols and formulae where needed. Also methodology used in this study was described. The rate-equality principle was used.

$$\lambda p_0 = \mu p_1$$

The queue discipline is based on a priority system which uses FCFS (first – come, first - served).

4. Result and discussion

The basis of Monte Carlo simulation is experimentation on change (or probabilistic) elements by means of random sampling. [77], enumerated the technique breakdown into five steps as follows:

(a) Setting up a probability distribution of important variables.
(b) Building a cumulative distribution for patients in the queue.
(c) Establishing an interval of random numbers for each variable.

(d) Generating random number.
(e) Actually simulating a series of trials.

The distribution of arrivals and accumulated patient in the queue in Thika Level 5 hospital and Kenyatta National Hospital are given below where service time is assumed constant for all patients and thus five minute per patient. The research has used the two hospitals because of the number of patients they handle and the queue they experience.

4.1. Summary of tables 1, 2 and 3

Table 1, 2, and 3 below are secondary data collected from the receptionist arrival patients recording book. The research decided to record at an interval of 30 minutes which is prior to change depending on how to compute the Monte Carlo simulation model. Also there is assumption of cumulative number of patients as doctor increases one by one until the queue diminishes in the last column. The research assumes three doctors is supposed to treat fifteen patients within 30 minutes, four doctors twenty patients, five doctors 25 patients and so on.

4.2. Arrival time and wait time

Table 1: Arrival table for collected on Monday at Thika Level 5 hospital

Inter Arrival Time	**Arrival**	**Cumulative Patient in the queue assuming there are 3, 4, 5 doctors respectively.**		
		3	**4**	**5**
8.00	**24**	**9**	**4**	**0**
8.30	**25**	**19**	**9**	**0**
9.00	**19**	**23**	**8**	**0**
9.30	**21**	**29**	**9**	**0**
10.00	**20**	**34**	**9**	**0**
10.30	**27**	**46**	**16**	**0**
11.00	**25**	**56**	**21**	**0**
11.30	**24**	**65**	**25**	**0**
12.00	**26**	**76**	**31**	**0**
12.30	**20**	**81**	**31**	**0**
1.00	**19**	**85**	**30**	**0**
1.30	**24**	**94**	**34**	**0**
2.00	**25**	**104**	**39**	**0**
2.30	**22**	**111**	**41**	**0**

Table 2: Arrival table for collected on Monday at Kenyatta National Hospital

Inter Arrival Time	Arrival	Cumulative Patient in the queue assuming there are 3, 4, 5, 6 doctors respectively.			
		3 doctors	4 doctors	5 doctors	6 doctors
8.00	32	17	12	7	2
8.30	29	31	21	11	0
9.00	26	42	27	12	0
9.30	27	54	37	14	0
10.00	30	69	44	19	0
10.30	27	81	50	21	0
11.00	29	95	59	25	0
11.30	26	106	67	26	0
12.00	27	118	79	28	0
12.30	30	113	83	33	0
1.00	30	148	90	38	0
1.30	32	165	100	45	2
2.00	31	181	120	51	3
2.30	30	196	127	56	3

Table 3: Arrival table for collected on Thursday at Thika Level 5 hospital

Inter Arrival Time	Arrival	Cumulative Patient in the queue assuming there are 3, 4 doctors respectively.	
		3/15	4/20
8.00	16	1	0
8.30	16	2	0
9.00	15	2	0
9.30	18	5	0
10.00	19	9	0
10.30	19	13	0
11.00	18	16	0
11.30	19	20	0
12.00	16	21	0
12.30	15	21	0
1.00	14	20	0
1.30	14	19	0
2.00	17	21	0
2.30	25	21	0

The model has been computed to use arrival time and servers (doctors) which can keep on changing depending on how many doctors are in operation at that hospital.

Monte Carlo simulation is run for several times with arrival time remaining constant and number of doctor changing. The simulated data generated at random after several trials when the model runs for 1000 times with three slots, four slots, five slots and six slots are as follows. Thika Level 5 hospital table 1: 0, 15, 23, 23, 49, 52, 56, 60, 70, 90, 94, 100, 98, 109; 0, 7, 8, 8, 8, 14, 20, 28, 28, 30, 30, 35, 39, 42; 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 respectively. Kenyatta National Hospital 0, 31, 42, 54, 80, 82, 118, 107, 117, 144, 152, 195, 207, 215; 5, 19, 27, 37, 44, 50, 59, 67, 79, 83, 90, 100, 120, 127; 5, 9, 14, 14, 16, 20, 25, 25, 27, 34, 40, 45, 50, 55; 0, 0, 0, 2, 1, 2, 0, 0, 0, 0, 1, 2, 2, 1 respectively. Thika Level 5 hospital table 3: 0, 2, 5, 6, 4, 13, 16, 18, 25, 18, 16, 19, 21, 20; 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,0

4.3. Performance of different waiting times

Below are results for different graphs for the two hospitals when data collected from hospital is plotted against simulated data from the model.

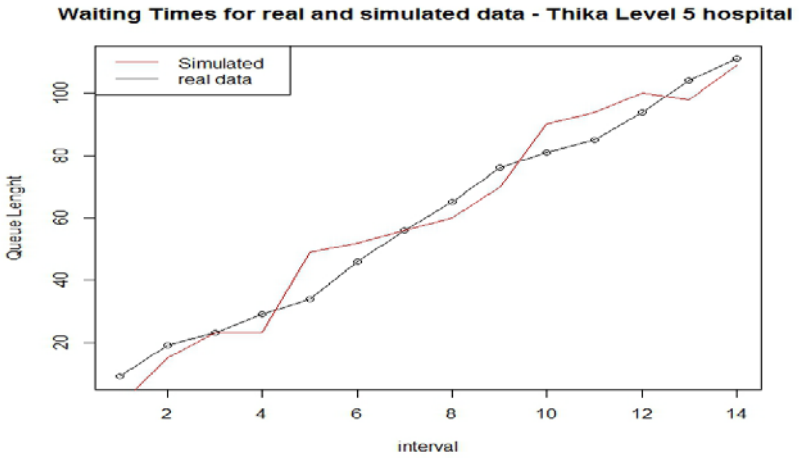

Figure 1: Thika Level 5 hospital on Monday for three slots

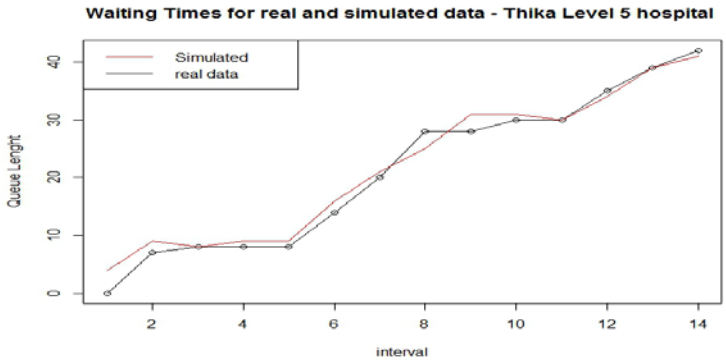

Figure 2: Thika Level 5 hospital on Monday for four slots

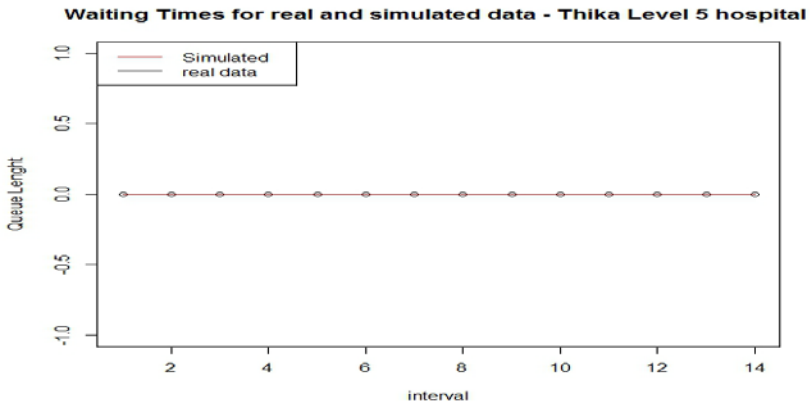

Figure 3: Thika Level 5 hospital on Monday for five slots

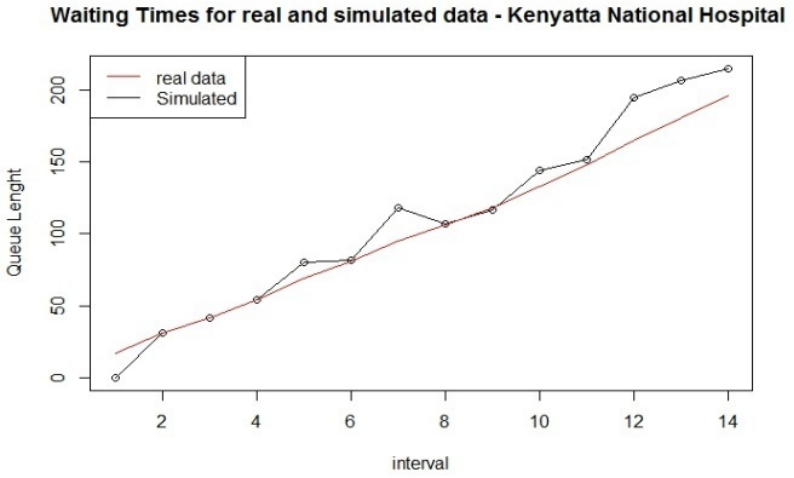

Figure 4: Kenyatta National Hospital on Monday for three slots

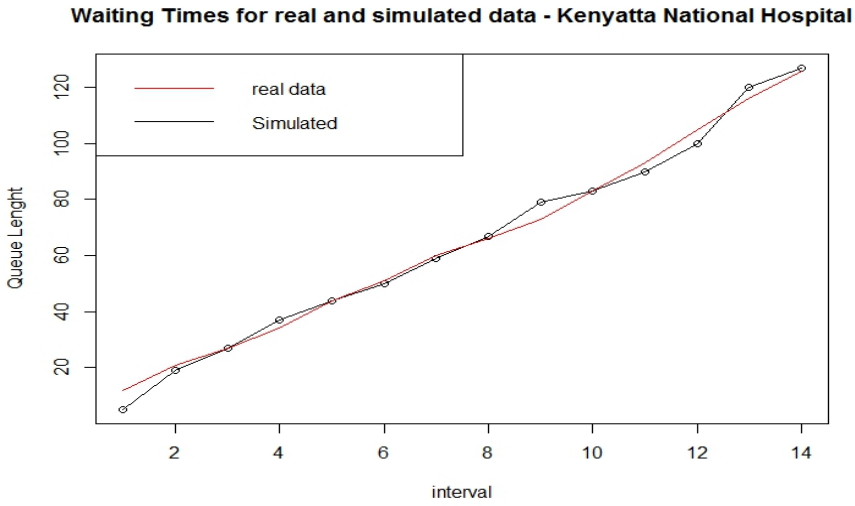

Figure 5: Kenyatta National Hospital on Monday for four slots

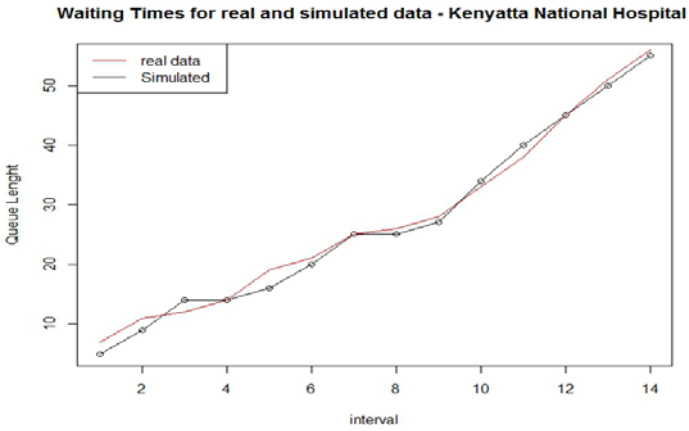

Figure 6: Kenyatta National Hospital on Monday for five slots

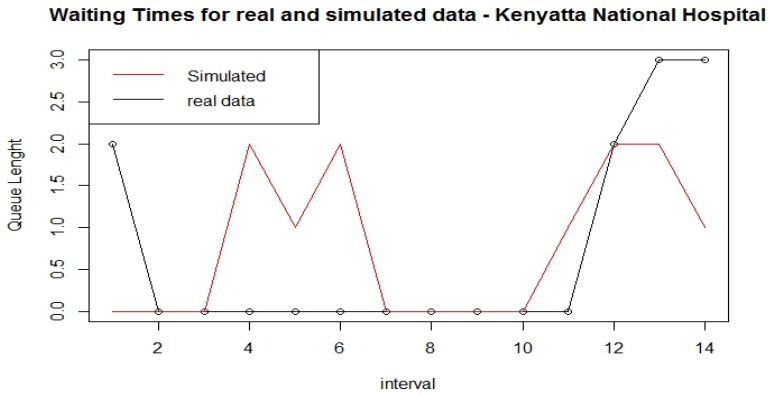

Figure 7: Kenyatta National Hospital on Monday for six slots

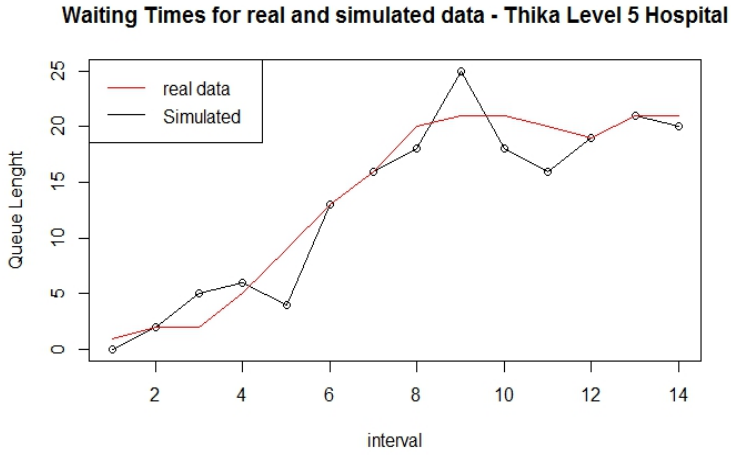

Figure 8: Thika Level Hospital on Thursday for three slots

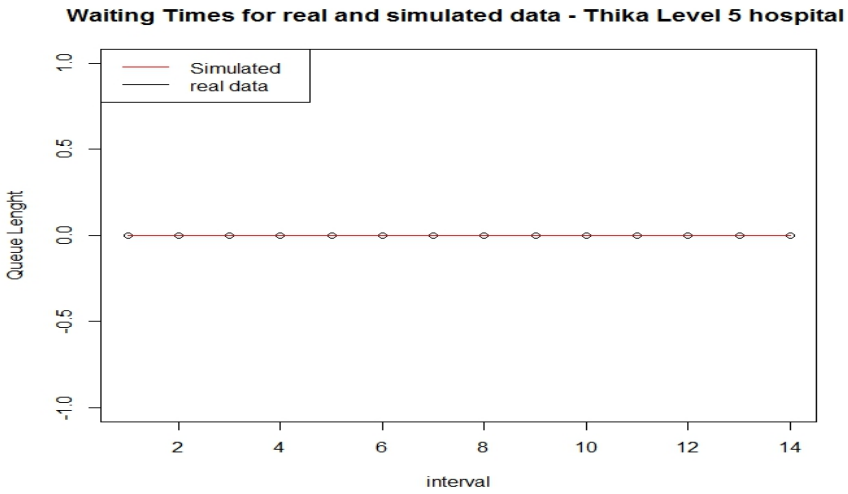

Figure 9: Thika Level Hospital on Thursday for four slots

4.4. Summary of figures 1 - 9

From the figures above the label indicate real data which is data collected from the hospital and simulated data which is data collected from Monte Carlo Simulation model. Also the queue length in y-axis is measured in terms of patients and interval in x-axis in time (hours).

Figure 1, 2 and 3 are results generated when data from Thika Level 5 hospital is plotted against data generated from the Monte Carlo Model. Figure 1 show the output for real data collected from the hospital where the hospital is under operation of three doctors in a day. The result shows the queue is very long of around 111 patients at 2pm. This is as a result from assumption three doctors are supposed to treat fifteen patients per thirty minutes. The service time is not constant for individual patients. It is assumed within 30 minutes there are those patients who can take one minute, three minutes or even 10 minutes to be treated depending on the nature of illness. Figure 2 show result if there is assumption of four doctors in operation. It shows at around 2pm the queue will diminutive to 41 patients in the queue. Figure 3 it was assumed having five doctors in operation there no queue. The research concluded it is optimal to have four doctors in operation on Monday since having five doctors most of the time they will be inactive.

Figure 4, 5, 6 and 7 are results generated when data from Kenyatta National Hospital is plotted against data generated from the Monte Carlo Model. Figure 4 show the output for real data collected from the hospital where the hospital is under operation of three doctors in a day. The result shows the queue is very long of around 196 patients at 2pm. This is as a result from assumption three doctors are supposed to treat fifteen patients per thirty minutes. The service time is not constant for individual patients. It is assumed within 30 minutes there are those patients who can take one minute, three minutes or even 10 minutes to be treated depending on the nature of illness. Figure 5 show result if there is assumption of four doctors in operation. It shows at around 2pm the queue will diminutive to 127 patients in the queue. Figure 6 show result if there is assumption of five doctors in operation. It shows at around 2pm the queue will diminutive to 56 patients in the queue. Figure 7 it was assumed

having six doctors in operation there will be no queue. The research concluded it is optimal to have five doctors in operation on Monday since having six doctors most of the time they will be inactive.

Figure 8 and 9 are results generated when data from Thika Level 5 hospital is plotted against data generated from the Monte Carlo Model. Figure 1 show the result for real data collected from the hospital where the hospital is under operation of three doctors in a day. The result shows the queue length 25 patients at 2pm. This is as a result from assumption three doctors are supposed to treat fifteen patients per thirty minutes. The service time is not constant for individual patients. It is assumed within 30 minutes there are those patients who can take one minute, three minutes or even 10 minutes to be treated depending on the nature of illness. Figure 9 show result if there is assumption of four doctors in operation there is no queue. The research concluded it is optimal to remain with three doctors in operation on Thursday since having four doctors most of the time they will be inactive.

5. Discussion, conclusions and future work

This chapter is aimed at discussing and summarizing the main findings from the study, drawing relevant conclusions and where necessary making some vital recommendations.

5.1. Discussion

On the basis of the results obtained from the plotted graph of three doctors from Thika Level 5 Hospital on Monday, the queue length is high as 111 patients waiting in the queue at 2.30pm. Additional of extra one doctor the queue length reduced to 41 patients at the same time 2.30pm. Adding another doctor to total five the queue length reduced to Zero patient all the time. The research concludes it's optimal to have four doctors on Monday since five is a waste of resources and most of times the doctors will be inoperative.

Still on the basis of the results obtained from the plotted graph of three doctors from Thika Level 5 Hospital on Thursday, the queue length is high as 25 patients waiting in the queue at 2.30pm. Adding another doctor the queue length reduced to Zero patient throughout the time. The research concludes it's optimal to retain three doctors since additional of one more doctors make most of them inactive all through and the purpose of the model is to advice the management on how to utilize resources.

For Kenyatta National Hospital having three doctors the queue length is 196 patients, four doctors the queue length reduced to 127 patients, five doctors the queue length reduced to 56 patients and six doctors most of time there is no queue. Therefore the research concludes it is optimal to have five doctors on Monday since doctors are not machines they also get tired. Six doctors is a waste of resources and most of times the doctors will be inoperative.

According to the study the research has shown Monte Carlo is more accurate than other models since it runs several times which generate more numbers. When this numbers are picked at random and plotted against real or collected data they produce exact results. That is the two lines discrepancy is very small.

This shows and proves Monte Carlo simulation Model is more accurate than other Models which uses probability and generates numbers once.

5.2 Conclusion

As long as increasing the productivity of healthcare organizations remains important, analysts will seek to apply relevant models to improve the performance of healthcare processes. The research shows that many models for estimating waiting time and utilization are available today but not adopted in our country. However, analysts will increasingly need to consider the ways in which distinct queuing systems within an organization interact. Thus, this thesis surveys the contributions and application of queuing models in the field of health care with simulation models.

The aim of the study is two-fold: Firstly, develop a predictive model to enable the hospital predicts the number of patients in the queue. Secondly, because patient waiting is undesirable, limiting waiting time is an important objective and therefore look for ways to reduce congestion at waiting room.

Experimental results show that Monte Carlo Model provides a high accuracy for the prediction of queue. This is because Monte Caro Model runs for several times and generate numbers randomly. It also uses Poisson distribution for arrival time and exponential distribution for service time. The research used random generated numbers and data collected from different hospital to plot the graphs which results shows the discrepancy is very minimal thus concluded Monte Carlo Simulation Model can be used to predict patient queue in the hospitals.

5.3 Future Work

As seen from the study Monte Carlo Simulation Model can be used to forecast volume of patients in the queue. Additional research in this area could yield interesting results. It is recommended that similar experiments be conducted with finite queue. It is also recommended to use Monte Carlo Simulation Model to calculate waiting time for each patient. Finally further research can be done using finite queue and Model Monte Carlo Simulation based on waiting time.

Acknowledgement

I acknowledge the support of my family and closest friends; Patrick Stanley Ngetich, Dancan Kitur, David Koskei and Linus Ngeno whose help and emotional support has seen me through my lowest moments.

And above all God, the creator, giver of all that is good for seeing me through this process.

Dedication

This research paper is dedicated to my loving wife Rose Langat and my son Ramsey Kipkirui

Competing interest

Author have declared that no competing interest exist

References

[1]. Mifflin., H. (2006). The American Heritage® Medical Dictionary. USA: Houghton Mufflin Harcourt

[2]. Houghton. (2007). The American Heritage® Medical Dictionary. USA: Houghton Mufflin Harcourt

[3]. Group, C. C. (2001). Analysis of American Nurses Association Staffing Surve.

[4]. Bazzoli, G. (2003). Does U.S. Hospital Capacity Need to be Expanded? Health Affairs 22(6) , 40-54.

[5]. American Nurses Association, I. (2010). Nurse Staffing. Retrieved from; USA: American Nurses Association.

[6]. Reports, B. E. (2013). The Evolving Role of Emergency Departments in th United states.

[7]. Hellmich, N. (2008). Aging population making more visits to the doctor's.

[8]. Bodenheimer, T. (2010). High and Rising Health Care Costs. Part 1: Seeking an Explanation.

[9]. Vissers, J. A. (2005). Health Operations Management. Patient Flow Logistics in Healthcare 1st Ed. New York, NY: Routledge Publishing.

[10]. New England Healthcare, I. (2010). Waste and Inefficiency in Health Care. England: NEHI.

[11]. Institute for Healthcare, I. (2005). Going Lean in Health Care. Innovation Series white paper, Institute for Healthcare

[12]. Lei Zhao, A. B. (2008). Modeling and Simulation of Patient Flow in Hospitals for Resoure Utilization. 1-10.

[13]. Hadfield, D. (2006). Tools for the Elimination of Waste in Hospitals, Clinics, and Other Healthcare Facilities. In D. Hadfield, The Lean Healthcare Pocket Guide. MCS Media, Inc.

[14]. Yang Wang, A. S. (2010). Fast Model Predictive Control Using Online Optimizaton.

[15]. Butcher, C. (2010). Emergency Department Patient Flow Simulation at HealthAlliance. USA: University of Iowa, 1-10.

[16]. Goverment, of. K. (2012). Health Sector Working Group Report. Republic of Kenya

[17]. American Nurses Association, I. (2013). Nurse Staffing. Retrieved from; USA: American Nurses Association.

[18]. Babbie, E. (2004). The practice of social research 10th Ed. Belmont CA: Wadsworth.

[19]. Mchugh, M. &. (2011). Improving Patient Flow and Reducing Emergency Department Crowding. 1-8.

[20]. McGaig, L. A. (2001). National Hospital Ambulatory Medical Care Survey. Advanced Data from Vital Health Statistics. Center for Disease Control and Prevention. 335.

[21]. Mccaig, L. E. (2003). National Hospital Ambulatory Medical Care Survey 2001: Emergency Department Summary. 335.

[22]. McCaig L, A. N. (2006). National Hospital Ambulatory Medical Care Survey 2004 Emergency Department Summary. Washington, DC: U.S. Department of Health & Human Services, Center for Disease Control & Prevention. National Center for Health Statistics, 372

[23]. Joint, C. R. (2004). Managing Patient Flow: Strategies and Solutions for Addressing Hospital Overcrowding. USA: Texas State University, 1-88.

[24]. Dexter, F. a. (2000). Optimal number of beds and occupancy to minimize staffing costs in an obstetrical unit.

[25]. Derlet R, a. R. (2000). Overcrowding in the Nation's Emergency Departments: Complex Causes and Disturbing Effects. Annals of Emergency Medicine , 35(1), 63-68.

[26]. Derlet, R. a. (2001). Frequent Overcrowding in the U.S. Emergency Departments. Academy of Emergency Medicine. 8 (2), 151-155.

[27]. Asplin, B. R. (2000). Measuring emergency department crowding and hospital capacity. Academic Emergency Medicine (9), 366-367

[28]. Asplin, B. R. (2002). Measuring emergency department crowding and hospital capacity. Academic Emergency Medicine. (9), 366-367.

[29]. Asplin, B. R. (2003). A Conceptual Model of Emergency Department Crowding. Annals of Emergency Medicine , 173-179.

[30]. Banks, J. and Carson J. (2005). Discrete-event system simulation - fourth edition. New Yolk: Pearson.

[31]. Weiss E. N. and McClain J. O. (1986), Administrative days in acute care facilities: a queueing- analytic approach, Operations Research, 35 (1), 35-44.

[32]. Worthington, D. J. (1987), Queueing models for hospital waiting lists, The Journal of the Operational Research Society, 38 (5), 413-422.

[33]. Gorunescu F., McClean S. I., and Millard P. H., (2002) A queueing model for bed-occupancy management and planning of hospitals, The Journal of the Operational Research Society, 53 (1), 19-24.

[34]. A. M. de Bruin, A. C. van Rossum, M. C. Visser, G. M. Koole (2006) Modelling the emergency cardiac in-patient ow: an application of queueing theory, Health Care Management Science, 10(2), 125-137.

[35]. Green, L.V.(2002), How many hospital beds?, Inquiry, 39(4) 400-412.

[36]. Cochran J. K., Bharti, A. (2006) Stochastic bed balancing of an obstetrics hospital, Health Care Management Science, USA: Health Care Management Science 9(1), 31-45.

[37]. Averill M. Law (2007) Simulation Modeling and Analysis, Seiten; McGraw-Hill, 226

[38]. Howard M. Taylor and Samuel Karlin (1998) Introduction to stochastic modeling 3rd edition, Califonia: Stanford University, 5.

[39]. Trivedi, K.S. (2001). Probability and Statistics with Reliability, Queuing and Computer Science Applications, Second edition, New Yolk: Wiley & Sons.

[40]. Hall, R.,Belson, D.,Murali, P., and Dessouky, M., (2006) "Modeling patient flows through the healthcare system," in Patient Flow: Reducing Delay in Healthcare Delivery (Hall, R., ed.) New Yolk: Springer.

[41]. ARMONY, M. and Zacharias C., (2013)"Panel sizing and appointment scheduling in outpatient medical care".

[42]. GREEN, l., (2006)"queueing analysis in healthcare," in patient flow: Reducing Delay in Health Delivery (HALL, R. W., ed), vol. 91 of International Series in Operations Research and Management Science, 281-307, Springer US

[43]. Greene J. (2007). Emergency Department Flow and the Boarded Patient How to Get Admitted Patients Upstairs. Annals of Emergency Medicine,49(1), 68-70.

[44]. Cassandras C.G., a. S. (1999). Introduction to Discrete Event Systems Springer 1 edition. New Yolk: Springer.

[45]. Jacobson, S. H.,Hall, S. N., and Swisher, J. R.,(2006) "Discrete-event simulation of health care systems," in Patient Flow: Reducing Delay in Healthcare Delivery (Hall, R. W., ed.), vol. 91 of International Series in Operations Research and Management Science, pp. 211-252, US Springer.

[46]. Zeltyn, S.,Marmor, Y. N.,Mandelbaum, A.,Carmeli, B.,Greenshpan, O.,Mesika, Y., … and Basis, F., (2011) "Simulation-based models of emergency departments: Operational, tactical and strategic staffing" ACM Trans. Model. Comput. Simul. 21, (24), 1-24

[47]. Yankovic, N. and Green, L. V. (2011) "Identifying good nursing levels: A queuing approach," Operations Research, 59(4), 942-955.

[48]. Harrison, G. A. (2001). Modelling Variability in Hospital Bed Occupancy.

[49]. Kolker, A. (2012). The use of operations Management Methodology for Quantitative Decision.

[50]. Farmer, R. a. (1990). Models for forecasting hospital bed requirements in the acute sector. Epidemiological community Health ..

[51]. McManus, M. L. (2004). Queuing theory accurately models the need for critical care. Anaesthesiology, 100, 1271-1276 .

[52]. Jones, S. a. (2002). Forecasting demand of emergency care. Healthcare Management Science, 5, 297 .

[53]. Harrison P.G., S.W.M. Au-Yeung and W.J. Knottenbelt (2005) A Queueing Network Model of Patient Flow in an Accident and Emergency Department.

[54]. Harrison, P. G., Harrison, S. K., Patel N. M., Zertal, S. (2012) Storage Workload Modelling by Hidden Markov Models: Application to Flash Memory, In: Performance Evaluation, 69 pp. 1740

[55]. Baum, L. E., Petrie, T. (1966) Statistical Inference for Probabilistic Functions of Finite Markov Chains, In The Annals of Mathematical Statistics, UK: University of Washington 37, 1554-63

[56]. Baum, L. E., Eagon, J. A. (1967) An Inequality with Applications to Statistical Estimation for Probabilistic Functions of a Markov Process and to a Model for Ecology, USA: In Bulletin of the American Mathematical Society, 73, 360-3

[57]. Shwartz, E. A. (1991). Adaptive control of constrained Markov chains: criteria and policies', . Annals of Operations Research 28, special issue on `Markov Decision Processes , 101-134,

[58]. Rabiner, L. R. (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, In IEEE, 77, 257-286

[59]. Ashraf, J., Iqbal, N., Khattak, N. S., Zaidi, A. M. (2010) Speaker Independent Urdu Speech Recognition Using HMM

[60]. . Arnoud M, a. G. (2007). Bottleneck analysis of emergency in-patient flow.

[61]. PRODAN, A and PRODAN, R. (2002) Stochastic Simulation and Modelling Romania: Iuliu Hatieganu University, 13, 461-466

[62]. Yoshikazu Goto, T. M. (2013). Decision-tree model for predicting outcomes after out-of-hospital cardiac arrest in the emergency department.

[63]. Gen-sheng, H. (2004) The analysis of queuing system based on support vector machine

[64]. Vaplik, V. N. (1998) foundation of support vector machines, New York: Springer-Verlag

[65]. Vapnik, V. N. (1995) The nature of statistical learning theory, New York: Springer-Verlag

[66]. Gensheng, H. and Deng, F. (2004) Application of support vector machine in queuing system

[67]. Sochalski, J. (2000). Nursing Shortage Redux: Turning the Corner on an Enduring Problem. Health Affairs (11), 57.

[68]. Project Management, I. (2004). A Guide to the Project Management Body of Knowledge: PMBOK Guide , (3rd ed.). Newton Square, Pennsylvania: Project Management Institute

[69]. Bhanot, G. .. (2005). Optimizing Task Layout on the Blue Gene/L Supercomputer . IBM Journal of Research and Development 49 (2/3), 489 .

[70]. El-Ramly, H. .. (2002). Probabilistic Slope Stability Analysis for Practice. Canadian Geotechnical Journal . 39 (3), 665.

[71]. Santos, T. .. (2005). Monte Carlo Simulation of Damaged Ship Survivability. Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment . 219 (1) , 25.

[72]. Lei, F. .. (1999). The INTEGRAL Mass Model – TIMM . Astrophysical Letters and Communications . 39 (1 – 6) , 841 .

[73]. Hahl, J. .. (2003). A Simulation Model for Estimating Direct Costs of Type 1 Diabetes Prevention . Pharmacoeconomics 21 (5) , 295 .

[74]. Phillips, C. (2001). The Economics of ‘ More Research is Needed ’ . . International Journal of Epidemiology . 30 (4) , 771 .

[75]. Boinske, C. .. (2003). How Much Can I Spend? . Journal of Financial Service Professionals 57 (1) , 33.

[76]. Cook, R. S. (2006). Performance Modeling Using Monte Carlo Simulation. USA: New Mexico State University

[77]. Heizer, J. and Barry Render. (2001) Operations Management, sixth edition, New York: Pearson