IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 10, Issue 4 Ver. VI (Jul-Aug. 2014), PP 47-53
www.iosrjournals.org

Queueing Analysis of Patient Flow in Hospital

Olorunsola S. A, Adeleke R. A and Ogunlade T. O
Department of Mathematical Sciences, Ekiti State University of Ado-Ekiti, Ekiti State, Nigeria.

Abstract: *Waiting on a queue is not usually interesting, but reduction in this waiting time usually requires planning and extra investments. The increasing population and health-need due to adverse environmental conditions have led to escalating waiting times and congestion in hospitals especially in the Emergency and Accident Departments (EAD). It is universally acknowledged that a hospital should treat its patients, especially those in need of critical care, in timely manner. Incidentally, this is not achieved in practice particularly in government owned health institutions because of high demand and limited resources in these hospitals. To enhance the level of admittance to care, optimal beds required in hospital is needed and this can be achieved by adequate knowledge of patient flow. In this paper, we show that queue theory can accurately model the flow of in-patient in hospital; we determine the optimal bed count and its performance measure.*
Keywords: *M/M/C queue, Poisson arrival, Exponential distribution, In-patient.*

I. Introduction

One of the major elements in improving efficiency in the delivery of health care services is in-patient flow. From a clinical perspective, in-patient flow represents the progression of a patient's health status. As such, an understanding of patient flow can offer education and insight to health care providers, administrators, and patients about the health care needs associated with medical concerns like disease progression or recovery status. Equally important, an understanding of patient flow is also needed to support a health care facility's operational activities. From an operational perspective, patient flow can be thought of as the movement of patients through a set of locations in a health care facility. Then, effective resource allocation and capacity planning are contingent upon patient flow because patient flow, in the aggregate, is equivalent to the demand for health care services (M. J Côté, 2000). The rising population and health-need due to adverse environmental conditions have led to escalating waiting times and congestion in hospital Emergency Departments (ED) Derlet R. W et al (2001). It is universally acknowledged that a hospital should treat its patients, especially those in need of critical care, in a timely manner. Incidentally, this is not achieved in practice particularly in government owned health institutions because of high demand and limited resources in these hospitals.

Queueing theory is used widely in engineering and industry for analysis and modeling of processes that involve waiting lines. In appropriate systems, it enables managers to calculate the optimal supply of fixed resources necessary to meet a variable demand. In the past, attempts have been made to apply queuing analysis to a variety of hospital activities, including cardiac care units, obstetric services, operating rooms and emergency departments, as a means of directing the allocation of increasingly scarce resources. More recently, health policy investigators have also sought to apply these techniques more widely across entire healthcare systems. Unfortunately, most proposed queuing models lack real-world validation and perhaps for this reason, have yet to be embraced by physicians and hospital administrators. Therefore, to explore the utility and implications of queuing theory as it relates to the supply and demand for critical care services, we sought to validate a simple queuing model in a busy hospital.

Queueing theory is a mathematical approach in Operations Research applied to the analysis of waiting lines. A.K. Erlang first analyzed queues in 1913 in the context of telephone facilities. The body of knowledge that developed thereafter via further research and analysis came to be known as Queueing Theory, and is extensively applied in industrial settings and retail sectors. Waiting on a queue is not usually interesting, but reduction in this waiting time usually requires planning and extra investments. Queuing theory involves the mathematical study of waiting lines. Queuing systems is a system consisting of flow of customers requiring service where there is some restriction in the service that can be provided. Three main elements are commonly identified in any service centre namely; a population of customers, the service facility and the waiting line.

We usually investigate queues in order to answer questions like, the mean waiting time in the queue, the mean response time in the system, utilization of service facilities, distribution of number of customers in the queue, etc. Decisions regarding the amount of capacity to provide a service must be made frequently by any service provider for optimality. The study of queueing theory requires some background study in probability theory and stochastic analysis.

II. Literature Review

Queueing theory has effectively been applied to various field of endeavour like traffic management, supermarket and health care etc. Weiss and McCliam (1986) used the M/G/∞ system to model the queue of patients needing alternative levels of care in an acute care facilities whose treatment is completed and are waiting to be transferred to an extended care facility. Adeleke R. A et al (2009) considered application of queuing theory to the waiting time of out-patients in a hospital. The average number of patients and the time each patient waits in the hospital were determined. Likewise in his paper Worthington (1987) used queuing theory to model hospital waiting list. He used an M/G/C queue with state dependent arrival rate to address the long- wait list problem. He experimented with various management actions such as increasing the number of beds or decreasing the mean service time through appropriate means. DeBruin *et al* (2006) investigated the emergency in-patient flow of cardiac patient in an hospital in order to determine the optional bed allocation so as to keep the fraction of those refused admission under a target hint. The authors find a relation between the size of a hospital unit, occupancy rate and target admission rates. After analytically estimating the required number of beds in first cardiac Acid (FCA) unit, they also used numerical method to estimate the number of beds in the Critical Care Unit (CCU) and Normal care clinical ward (NC). Jonathan E. H *et al* (2009) characterize an optimal admission threshold policy using control on the scheduled and expedited gate way for a new Markov Decision process model. In their work, they presented a practical policy base on insight from the analytical model that yield reduced emergency blockages, cancelations and off-units through simulation based on historical hospital data.

Application of queueing theory to model health care is growing more popular as hospital management teams are becoming aware of the advantages of these techniques. In this research we will use both analytical techniques and simulation to study a simple queuing network composed of only two service stations placed in tandem. In this paper, we studied all admissions into the Emergency and Accident Department (EAD) of a tertiary hospital. We will show that admissions into this system has a Poisson distribution, hence it has exponential inter-arrival rate. We also examine the average length of stay, the occupancy rate and we determine the optimal bed count in the Intensive Care Wards (ICW) and the Medical and Surgical Wards (MSW). Since the ICW and MSW have multiple beds we will consider the M/M/c queue.

III. The M/M/c Model

For this queueing system, it is assumed that the arrivals follow a Poisson probability distribution at an average of λ customers (patients) per unit of time. It is also assumed that they are served on a first come, first-served basis by any of the doctors. The service times are distributed exponentially, with an average of μ customers (patients) per unit of time and c number of servers. 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. By the rate-equality principle

$$\lambda p_n + n\mu p_n = \lambda p_{n-1} + (n+1)\mu p_{n+1} \qquad 3.1$$

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

$$\lambda p_n + c\mu p_n = \lambda p_{n-1} + c\mu p_{n+1} \qquad 3.2$$

$$\begin{aligned} \lambda p_n - (n+1)\mu p_{n+1} &= \lambda p_{n-1} - n\mu p_n \\ &= \lambda p_{n-2} - (n-1)\mu p_{n-1} \\ &\vdots \\ &= \lambda p_0 - \mu p_1 \\ &= 0 \end{aligned}$$

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

$$p_n = \frac{\lambda\lambda \ldots . \lambda}{(\mu)(2\mu) \ldots (n\mu)} p_0$$

$$p_n = \frac{\lambda/\mu}{n} p_{n-1} = \frac{(\lambda/\mu)^2}{n(n-1)} p_{n-2} = \cdots = \frac{(\lambda/\mu)^n}{n!} p_0 \qquad 3.3$$

In a similar fashion, we get that for $n > c\ (i.e\ n = c, c+1, c+2 \ldots)$

$$\begin{aligned} p_n &= \frac{(\lambda)(\lambda) \ldots . (to\ n\ factors)}{[(\mu)(2\mu) \ldots (c\mu)][(c\mu)(c\mu) \ldots (to\ (n-c) factors)]} p_0 \\ &= \frac{\lambda^n}{c!\, \mu^c\, c^{n-c} \mu^{n-c}} p_0 \\ &= \frac{(\lambda/\mu)^n}{c!\, c^{n-c}} p_0 = \frac{\lambda}{c\mu} p_{n-1} = \rho^{n-c} p_c \end{aligned} \qquad 3.4$$

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

$$p_0 = \left[\sum_{n=0}^{c-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^c}{c!\,(1-\lambda/c\mu)}\right]^{-1} \qquad 3.5$$

Note that the probability an arrival unit has to wait on arrival is given by the probability

$$P(N \geq c) = \sum_{n=c}^{\infty} p_n = \frac{(\frac{\lambda}{\mu})^c}{c!\,(1-\rho)} p_0 = \frac{p_c}{1-\rho} \qquad 3.6$$

This is known as Erlang's C formula or Erlang delay probability. We now proceed to compute some performance measures.

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

3.1 Expected Nunber Of Busy And Idle Servers

The expected number of busy servers E(B) is given by

$$E(B) = \sum_{n=0}^{c-1} np_n + \sum_{n=c}^{\infty} cp_n \qquad 3.7$$

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

$$= \frac{\lambda}{\mu}\left[\sum_{m=0}^{c-2} \frac{(\frac{\lambda}{\mu})^{m}}{m!} + \frac{\{(1-\rho)+\rho\}(\frac{\lambda}{\mu})^{c-1}}{(c-1)!\,(1-\rho)}\right] p_0$$

$$= \frac{\lambda}{\mu}\left[\sum_{m=0}^{c-1} \frac{(\frac{\lambda}{\mu})^{m}}{m!} + \frac{(\frac{\lambda}{\mu})^{c}}{c!\,(1-\rho)}\right] p_0$$

$$= \frac{\lambda}{\mu} p_0^{-1} p_0 = \frac{\lambda}{\mu} = c\rho \qquad 3.8$$

Hence the expected number of idle servers E(I) is given by

$$E(I) = E(c-B) = E(c) - E(B)$$

$$= c - c\rho = c(1-\rho) \qquad 3.9$$

The expected queue length L can be computed as

$$L = \sum_{n=c}^{\infty} (n-c) p_n$$

$$= \sum_{n=c}^{\infty} (n-c) \frac{(\frac{\lambda}{\mu})^n}{c!\,c^{n-c}} p_0$$

$$= \frac{\rho p_c}{(1-\rho)^2} = \frac{\rho}{(1-\rho)} P(N \geq c) \qquad 3.10$$

A special case where c=1, $L = \frac{\rho}{(1-\rho)}$

Where $\rho = \lambda/c\mu < 1$ is referred to as the server utilization. Applying Little's formula, we also obtain the expected waiting time in the queue

$$W = \frac{L}{\lambda} = \frac{\rho_c}{\mu(1-\rho)^2} \qquad 3.11$$

We can find the steady state waiting-time distribution for the M/M/c queue. Let $w_q(x)$ and $w(x)$ be the PDFs of the waiting time W_q and W_s in the queue and in the system, respectively. We obtain $w^*(s)$ by conditioning. If a patient finds on arrival $n < c$, he does not have to wait, and his waiting time in the system equals to his service time, that is,

$$w^*(s|n) = \frac{\mu}{s+\mu} \qquad \text{for } n < c \qquad 3.12$$

If he finds $n \geq c$ patients in the hospital, he has to wait in the hospital until the completion of service of (n-c+1). All the c beds being occupied, then the service rate is cμ. Taking into consideration his own service time, he has to wait in the system for the completion of (n-c+1) services at the rate cμ and his own service at the rate μ. That is,

$$w^*(s|n) = \left(\frac{c\mu}{s+c\mu}\right)^{n-c+1}\left(\frac{\mu}{s+\mu}\right) \qquad for\ n \geq c \qquad 3.13$$

Using PASTA property, $a_n = p_n$. Thus

$$w^*(s) = \sum_{n=0}^{c-1} w^*(s \mid n)p_n + \sum_{n=c}^{\infty} w^*(s|n)p_n$$

$$= \frac{\mu}{s+\mu}\left[\sum_{n=0}^{c-1} p_n + \sum_{n=c}^{\infty} p_n\left(\frac{c\mu}{s+c\mu}\right)^{n-c+1}\right] \qquad 3.14$$

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

$$= \sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}p_0\frac{\mu}{s+\mu} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}p_0\frac{c\mu^2}{(c-1)\mu-\lambda}\left[\frac{1}{s+\mu} + \frac{1}{s+c\mu-\lambda}\right]$$

Inverting the transform, we get

$$w(t) = \sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}p_0\mu e^{-\mu t} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}p_0\frac{c\mu^2}{(c-1)\mu-\lambda} \times [e^{-\mu t} - e^{-(1-\rho)c\mu t}] \qquad 3.15$$

A special case for c=1

$$w(t) = \mu(1-\rho)e^{-(1-\rho)\mu t}$$

Since $w^*(s) = w_q^*(s)[\mu/(s+\mu)]$, we get from 3.14

$$w_q^*(s) = \sum_{n=0}^{c-1} p_n + \sum_{n=0}^{\infty} p_n\left(\frac{c\mu}{s+c\mu}\right)^{n-c+1}$$

$$= \sum_{n=0}^{c-1}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!} + \frac{\left(\frac{\lambda}{\mu}\right)^c}{c!}\frac{c\mu}{s+c\mu-\lambda} \qquad 3.16$$

Inverting the transform gives

$$w_q(t) = \left(\sum_{n=0}^{c-1} p_n\right)\delta(t) + \sum_{n=c}^{\infty} p_n\frac{c\mu(c\mu t)^{n-c}e^{-c\mu t}}{(n-c)!}$$

Simplifying, we get

$$w_q(t) = \left(1 - \frac{p_c}{1-\rho}\right)\delta(t) + c\mu p_c e^{-c\mu(1-\rho)t}, \qquad t > 0 \qquad 3.17$$

Where $\delta(t)$ is the Dirac delta (or impulse function). Note that the coefficient of $\delta(t)$ is the probability of zero wait, or the probability that there is a free server upon arrival.

IV. Length Of Stay Distribution

The number of days in hospital for a patient is described by the term length of stay (LOS). LOS is defined as the time of discharge minus time of admission. Following, the average length of stay is abbreviated as ALOS. The average length of stay in hospitals is a statistical calculation often used for health planning purposes. $\text{Average Length of stay (in days)} = \frac{\text{Total discharge days}}{\text{Total discharges}}$ Or $\text{Average Length of stay (in days)} = \frac{\text{Total inpatient days of care}}{\text{Total admissions}}$

Below are the definitions for each of the four data items included in the above calculations:

TOTAL DISCHARGE DAYS - The sum of the number of days spent in the hospital for each inpatient who was discharged during the time period examined regardless of when the patient was admitted.

TOTAL DISCHARGES - The number of inpatients released from the hospital during the time period examined. This figure includes deaths. Births are excluded unless the infant was transferred to the hospital's neonatal intensive care unit prior to discharge.

TOTAL INPATIENT DAYS OF CARE - Sum of each daily inpatient census for the time period examined.

TOTAL ADMISSIONS - The total number of individuals formally accepted into inpatient units of the hospital during the time period examined. Births are excluded from this figure unless the infant was admitted to the hospital's neonatal intensive care unit.

V. Bed Occupancy

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) Huang X (1995)

$$\text{bed requirement} = \frac{\text{average no. of daily admissions}}{\text{average bed occupancy rate}} \times \text{average length of stay} \qquad 5.1$$

Hospital bed capacity decisions have been made based on Target Occupancy Rate (TOR) – the average percentage of occupied beds and the most commonly used occupancy target has been 85% Linda V. Green (2002). Another metric often cited in the literature is the Target Access Rate (TAR), which measures the percentage of the time that a census count will show that the hospital contains at least one empty bed, de Bruin et al (2007), Kumar and John (2010).

VI. Numerical Solution

We consider a tertiary hospital in the south-western Nigeria and studied the admissions through its Accident and Emergency department to the ICW and MSW. The queueing model selected assumes that daily admissions rate follow a Poisson distribution and this behavior was confirmed here by goodness of fit test as illustrated in the fig1. Green L.V(2002) and Milne and Whitty (1995) have shown that the arrival rate into Intensive Care Unit follows a Poisson distribution.

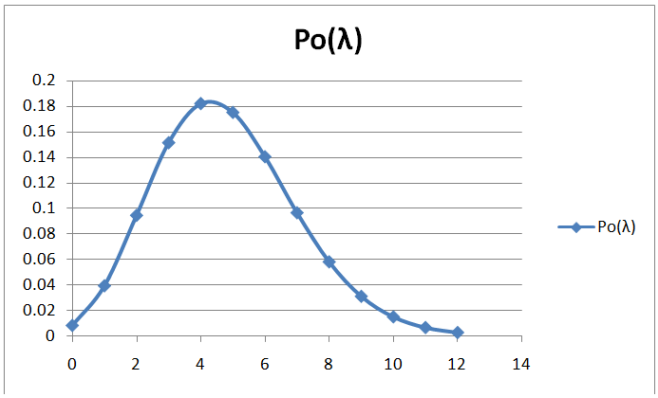
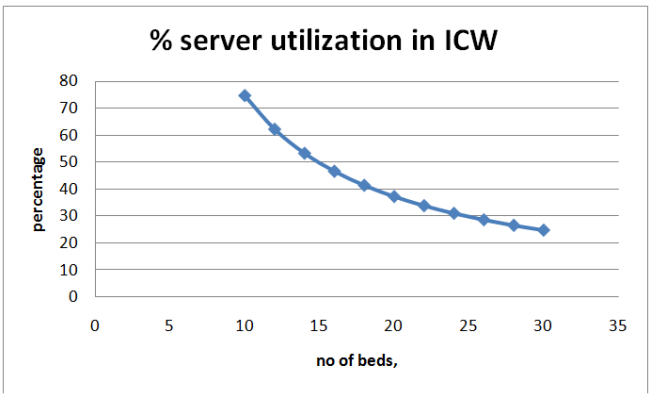


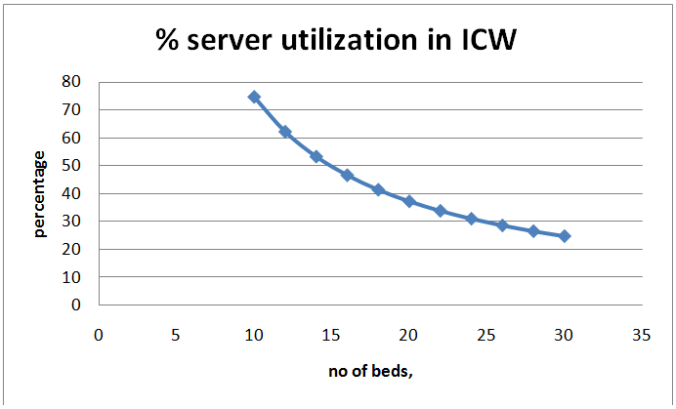
Fig2 Distribution of arrivals into the System.

The occupancy rate (ρ) is related to the real demand (λ) and LOS (μ) and can be defined as follows,

$$\rho = \frac{Average\ number\ of\ beds\ occupied}{number\ of\ beds\ available} = \frac{(1-P_c)\lambda\mu}{c} \qquad 7.1$$

The term $(1-P_c)\lambda$ can be entitled as the effective demand as the refused admissions are subtracted from the real demand. Furthermore, the product $\lambda\mu$ which is the expected number of patients in the system is also known as the workload of the system. Many hospitals use the same target occupancy rate for all hospital units, no matter the size of the unit. The target occupancy rate is typically set at 85% and has developed into a golden standard [Green (2002)]. The conclusion is clear and important. Larger hospital units can operate at higher occupancy rates than smaller ones while attaining the same percentage of refused admissions. Therefore, one target occupancy rate for all hospital units is not realistic [De Bruin (2007)].





Total admission into ICW=634, ALOS in ICW=4.44days, percentage of patients reneged, k=3.4%. We also have the following set of data for the MSW: Total admission into MSW=5073, ALOS=7.24 days.
From the parameter values specified, we estimate the arrival rate to each station as

$$\lambda_{ICW} = \frac{N_{ICW}}{365days} = 1.74days^{-1}$$

$$\lambda_{MSW} = \frac{N_{MSW}}{365days} = 13.9days^{-1}$$

But the queue leading to the MSW is composed of new arrivals and blocked patients from the ICW. Also we have only a fraction $1-k = 96.6\%$ of patients arrived into ICW without reneging during service. So that the effective arrival into the ICW is

$$\lambda_{ICW}^{e} = \lambda_{ICW}(1-k) = 1.681days^{-1}$$

And $$\lambda_{MSW}^{e} = \lambda_{ICW}^{e} + \lambda_{ICW} = 15.581days^{-1}$$

Table1: Performance Measure for ICW

c_1	% server utilization	Probability of delay	Mean waiting time	Mean waiting time in queue (hrs)
10	74.64	0.299850	4.964899	12.600000
12	62.20	0.093022	4.531046	2.1900
14	53.31	0.023395	4.455892	0.5600
16	46.65	0.004763	4.442477	0.0590
18	41.46	0.000791	4.440333	0.0080
20	37.32	0.000108	4.440038	0.0010
22	33.93	0.000012	4.440004	0.0001
24	31.10	0.0000001	4.440000	0.0000
26	28.71	0.000000	4.440000	0.0000
28	26.66	0.000000	4.440000	0.0000
30	24.88	0.000000	4.440000	0.0000

Table2: Performance Measure for ICW

c_2	% server utilization	Probability of delay	Mean waiting time	Mean waiting time in queue (hours)
114	98.95	0.869420	12.513803	126.570000
116	97.25	0.679725	8.780979	36.980000
118	95.60	0.523275	7.969463	17.507112
120	94.01	0.396277	7.638836	9.572064
122	92.46	0.294935	7.472264	5.574336
124	90.97	0.215531	7.379405	3.345720
126	89.53	0.154511	7.324789	2.034936
128	88.13	0.108572	7.291737	1.241688
130	86.77	0.074723	7.271465	0.755160
132	85.46	0.050335	7.258987	0.455688
134	84.18	0.033167	7.251330	0.267192
136	82.95	0.021366	7.246670	0.160080

Tables1 and 2 show result for various values of c_1 and c_2 . From the tables we can see that $c_1 = 24$ guarantees that there is no waiting at the EAW, since urgent patient needing urgent care are brought in through it. In the MSW, $c_1 = 132$, will guarantee an approximate of 85.46% server utilization and a minimum waiting time in queue.

VII. Summary

An admission guarantee should be one of the main goals of any hospital for patients entering through its Emergency and Accident Department. In this work, we analyzed a queueing network model with reneging to study how waiting time in the EAD of an hospital is influenced by the number of beds in the ICW and MSW. The system was decomposed into two independent multi-server queues so as to obtain estimates for the required number of beds in the wards. We found that the required number of beds to ensure that emergent patients are promptly attended and there is easy flow is approximately 24 in the ICW and 132 in the MSW for the test hospital under consideration.

References

[1]. Adeleke. R. A, Ogunwale O. D, Halid O. Y (2009) Application of Queueing Theory to Waiting Time of Out-patients in Hospitals. Pacific Journal of Science and Technology Vol. 10(2) 270-274

[2]. Arun Kumar and John Mo (2010) Models for Bed Occupancy Management of a Hospital in Singapore. Proceedings of the 2010 International Conference on Industrial Engineering and Operations Management Dhaka, Bangladesh.

[3]. DeBruin A.M, A.C van Rossun MC Visser, GM Koole(2006) Modeling the emergency cardiac in-patient flow: an application of queuing theory. Health Care Management Science Vol. 10, No 2, pp 125-137

[4]. Green L. V (2002) How many Hospital Beds? Decision, Risk and Operation. Working Paper Series

[5]. Green LV, Soares J, Giglio JF, Green RA (2006) Using queueing theory to increase the effectiveness of emergency department provider staffing. Academic Emergency Medicine 13(1):61-68

[6]. Jonathan. E H, shervin Ahmad Beygi, Mark P.Van Oyen (2009), Design and Analysis of Hospital Admission Control for operational Effectiveness. Technical Report 09-05. University of Michigan. Michigan

[7]. Murray J. Cote (2000) Understanding Patient Flow. Production/Operations Management Decision line pp 8-10

[8]. Vanberke P. T Boucherier R. J Hans E. W Hurin J. L, Litvak N (2010). A Survey of Health Care Models that Encompasses Multiple Department. Intl Journal of Health Management Information 1(1):37-69

[9]. Weiss E.N, J.O Mcclain (1986) Administrative days in acute care facilities: a queing analytic approach operations Research Vol. 35 No 1 pp 35-44

[10]. Worthington D.J (1987) queueing models for hospital waiting list. The journal of the operational research Society Vol. 38, No 5, pp 413-422