Improving OPD efficiency through queuing theory: The case of a regional medical center

Marilou L. Andaya-Basilio
Department of Industrial Engineering
Saint Louis College, San Fernando City, La Union
Contact information: mbas_72@yahoo.com or mbas_99@hotmail.com

Abstract:

Premised on the assumption that patient waiting time is an important quality health care indicator, this research assesses the current queuing models of the outpatient department in a regional medical center in the hope of coming up with new queuing models that can improve the efficiency of the OPD operations. The new queuing models are developed by means of the following directly related components, namely: hospital policies which specifically include the number of doctors and nurses deployed, the service time and the OPD operation time, and the demand for health care based specifically on patient arrival and pattern of arrival.

Keywords: *Queuing theory, efficiency improvement, outpatient Department, waiting time, operating characteristic, ITRMC*

Queuing analysis has contributed materially to improvements in outpatient systems, where a central issue is the tension between patient waiting time and the efficient utilization of medical personnel. On one hand, excessive patient waiting time undermines the hospital's efficiency. Such delay leads to patient dissatisfaction and thus may eventually result in loss of equity of a competitive health care system (De Man, 2004). On the other hand, while it is true that the hospital management can easily and indiscriminately increase capacity to reduce waiting time, the fact is, there is an ever-present constraint with respect to a hospital's budget. Hence, to justify any decision to increase capacity or status quo, OPD system performance measures should be studied and critically analyzed as basis for capacity planning. These performance measures are determined using queuing theory.

This research looks at the case of the outpatient department of the Ilocos Training and Regional Medical Center (ITRMC) where large numbers of patients waiting are a common sight. A preliminary inquiry was conducted to determine the overall performance measures of the outpatient department compared with the expectations of the patients. From this inquiry, it was apparent that patients expected to be seen and served within thirty minutes upon arrival at the outpatient department. The actual experience of the patients deviates from their expectations, however. The Public Assistance and Complaints Unit of ITRMC has never received complaints from patients but the absence of complaints does not necessarily mean that the patients are satisfied with regard to the services experienced in the OPD departments.

There is a need, therefore, to carry out a systematic study on patient waiting time, with a view to identifying the factors that affect this and then recommend ways of minimizing delays through appropriate queuing models for the outpatient operations. To carry these out, this research aims at the following:

1. Find out if there is a significant difference between the actual waiting time and the expected waiting time of the patients from the departments of Medicine, Pediatrics, Obstetrics and Gynecology and Surgery.

2. Describe the existing queuing model of the outpatient departments of Medicine, Pediatrics, Obstetrics and Gynecology, Surgery along the different queuing parameters, namely: average number of customers in the system (waiting and being served), average number of customers in the waiting line, average time a customer spends in the system (waiting and being served), average time a customer spends waiting in line, probability of no (zero) customer in the system, and utilization rate – i.e., the proportion of time the system is in use.

3. Illustrate the appropriate queuing model for the outpatient departments of Medicine, Pediatrics, Obstetrics and Gynecology and Surgery.

This study intends to develop a queuing model by means of the following directly relating components, namely: hospital policies which specifically include the number of doctors and nurses deployed, the service time and the OPD operation time, and the demand for health care based specifically on patient arrival and pattern of arrival.

Design and methods

A combination of methods in the form of questionnaires, direct observation, and interviews with both patients and hospital personnel was utilized in this inquiry. The respondents of the study were selected from the stream of patients arriving for consultation in the four selected departments of the outpatient clinic. A questionnaire was given to a patient immediately upon registration. The registration time was automatically generated by the computer. Such record was extracted from the Registration database. The patient was observed throughout his/her service encounter. A time study was conducted to determine the actual waiting times (Barnes, 1980). The queuing system's operating characteristics were approximated using the actual number of servers, actual arrivals distribution and actual service time distribution. The arrival rate was computed for each department using the data taken from the medical records section. The service rate was computed for each department using the time study data. The last problem is to show a graphical presentation to illustrate the appropriate queuing model of the selected departments. The model was developed based on the expected waiting time of the patients as gathered from the survey. The mean expected waiting time was used as the basis as this is the most acceptable representative of the data. Since the output of the study is a recommended model, such model was tested in order to validate its claims on improvement.

Findings and discussion

Significant difference between the actual waiting time and the expected waiting time

To determine the significant difference between the actual and expected waiting time, the Kolmogorov-Smirnov test was used (Downie, 1983). The theoretical frequency distribution is the expected waiting time of the patients taken before the actual experience of the service and the observed frequency distribution is the actual waiting time of the patients. Table 1 presents the computed and critical Dn for the selected departments.

Table 1. Actual and Expected Waiting Time

Departments	Waiting Time	Observed Frequency	Observed Relative Cumulative (Fo)	Expected Frequency	Expected Relative Cumulative (Fe)	Dn
MEDICINE	< 10mins	0	0	7	0.1707	computed = 0.6829 critical = 0.2124
	10 - 20	2	0.0488	17	0.5854	
	20 - 30	3	0.1220	9	0.8049	
	30 - 1 hr	17	0.5366	6	0.9512	
	> 1 hr	19	1	2	1	
PEDIAT-RICS	< 10mins	0	0	8	0.2667	computed = 0.4667 critical = 0.2483
	10 - 20	3	0.1	9	0.5667	
	20 - 30	10	0.4333	6	0.7667	
	30 - 1 hr	12	0.8333	3	0.8667	
	> 1 hr	5	1	4	1	
OB-GYNE	< 10mins	0	0	3	0.1071	computed = 0.7500 critical = 0.2570
	10 - 20	1	0.0357	11	0.5000	
	20 - 30	0	0.0357	8	0.7857	
	30 - 1 hr	17	0.6429	2	0.8571	
	> 1 hr	10	1	6	1	
SURGERY	< 10mins	0	0	8	0.2353	computed = 0.6176 critical = 0.2332
	10 - 20	4	0.1176	10	0.5294	
	20 - 30	2	0.1765	9	0.7941	
	30 - 1 hr	6	0.3529	3	0.8824	
	> 1 hr	22	1	4	1	

From the information above, it is evident that in all the departments, the Dncomputed is greater than the Dncritical at 0.05 level of significance. There is a significant difference, therefore, between the patients' expectations and their actual waiting time, and that the expectations are definitely lower than the actual experience of the patients. This means that the queuing models should then be improved.

Existing queuing model of the outpatient departments of medicine, pediatrics, obstetrics and gynecology, and surgery

In order to describe the queuing system of the departments, the average customer arrival rates in patients per hour, the average service rates in patients per hour, and the average number of servers per hour will be considered. From observation and interviews with the medical staff, the consultation starts only an hour in the morning and afternoon after the start of registration. The primary reason for this is that in the morning, the doctors do the rounds and in the afternoons they have a short conference. The patients were informed of this matter as it is part of the regular information always posted in the OPD registration area. Using data taken from Medical Records, the average arrival rate per hour per department were noted. Actual service times of the observed patients by the nurse and the doctor respectively were used to approximate the average service rate of the server in terms of patients seen per hour for each department. Table 2 shows the summary of all the values needed to come up with the existing queuing system of the four departments.

Table 2. Number of Servers, Patient Arrival Rate and Service Rates

Department	Ave No. of nurses (s1)	Arrival Rate for nurse (λ1) PATIENTS/HR	Service Rate, nurse (μ1) PATIENTS/HR	Ave No. of doctors (s2)	Arrival Rate for doctor (λ2) PATIENTS/HR	Service Rate, doctor (μ2) PATIENTS/HR
MEDICINE	2	7.3	7.5	1.2	7.3	3.75
PEDIATRICS	1.8	4.1	10	1.3	4.1	4.6
OB-GYNE	1.2	6.5	7.5	1.2	6.5	3.5
SURGERY	2	9.3	6	1.4	9.3	4.3

The number of doctors and nurses available per hour for the entirety of the consultation period was expressed as averages since the nurses and doctors are called once in a while and, hence, were not available for consultation. Since all the departments follow the multiple servers - multiple phase model (see Annex B), there would be a different arrival rate for the nurse and the doctor. The nurse and the doctor work in pairs, meaning a patient who queues up in the Medicine Department can be seen by only one set of nurse-doctor tandem but most of the time, the nurse and the doctor work simultaneously so they are treated separately. (There are times, however, that the nurse assists the doctor.) The arrival rate for the doctor is determined by getting the lowest value between patient arrival rate and the service rate of the nurse. For better comprehension of the existing queuing models, a graphical presentation was developed to illustrate the queuing models and the flow of patients from server to server as shown in Figure 1 below.

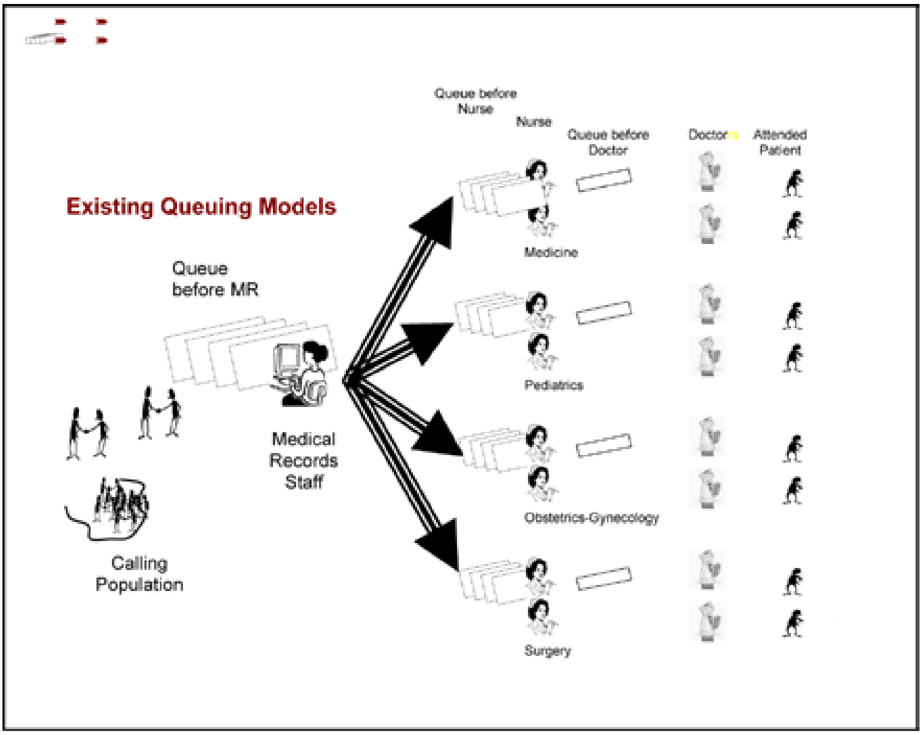


Figure 1: Existing Queuing Models

Using the appropriate queuing model which is the multiple server - single phase (see Annex B), all values from Table 2 were substituted in the working equations (see Annex C), thus coming up with the different operating characteristics or performance measures that describe the existing queuing system. The results are presented in Table 3.

Table 3. Operating Characteristics, Existing Models

Nurse Notation	Description	Medicine	Pediatrics	Ob-Gyne	Surgery	Unit
L	Average number of patients in the system (waiting and being served)	1.28	0.44	1.15	4.78	Patient/s
Lq	Average number of patients in the waiting line	0.3	0.03	0.28	3.23	Patient/s
W	Average time a patient spends with the nurse (waiting and being served)	10.48	6.39	10.58	30.83	Minutes
Wq	Average time a patient spends waiting in line	2.48	0.39	2.58	20.83	Minutes
P0	Probability of no (zero) patient waiting in the system	0.35	0.65	0.37	0.11	
ρ	Utilization rate; the proportion of time the system is in use	65%	35 %	63%	89%	
Doctor Notation	**Description**	**Value**	**Value**	**Value**	**Value**	**Unit (hr)**
L	Average number of patients in the system (waiting and being served)	10.91	1.24	13.48	6.76	Patient/s
Lq	Average number of patients in the waiting line	9.08	0.42	11.62	5.54	Patient/s
W	Average time a patient spends with the doctor (waiting and being served)	89.64	18.21	124.44	67.6	Minutes
Wq	Average time a patient spends waiting in line	10.22	6.21	107.30	55.35	Minutes
P0	Probability of no (zero) patient waiting in the system	0.05	0.36	0.04	0.09	
ρ	Utilization rate; the proportion of time the system is in use	95%	64%	96%	91%	
Total Waiting Time		**12.7**	**6.6**	**109.88**	**76.18**	**Minutes**

The waiting times of 12.7 and 6.6 minutes at the departments of Medicine and Pediatric, respectively, are low compared to European standards of 30 minutes (Standard Patient's Charter of England). This means that both departments were operating within acceptable standards during the time of the study. For Ob-Gyne, patients are waiting on the average of 109.88 minutes. The utilization rate of 63% and 96% means that the nurses and doctors are busy 63% and 96% of the time. The waiting time of 109.88 minutes is very high as compared to European standards, which means that the Obstetrics-Gynecology department was operating below acceptable standards during the time of the study. The same thing can be said of the Surgery whose waiting time is 76.18 minutes.

The appropriate queuing models

Using the waiting times for the four departments studied, the most optimal combinations were arrived at to meet patient waiting time expectations. Technically, it would have been possible to set the most acceptable waiting time and determine the needed number of servers but it would result to a fraction. To counter this, the capacities were assigned using the trial and error method, in order to arrive at the best combination resulting to the least time that will meet global standards of 30 minutes maximum waiting time. A motion and time study of a well trained nurse's activities was conducted, coming up with the standard processing / service time per department. The standard time and the values for all parameters used are shown in Table 4. A standard time of consultation cannot be established for the doctor's activities because the time to attend to a patient will depend upon the complaints, severity of the complaint and the amount of information asked for by the patient. Because of this variation, the previous serving time of doctors for the four departments will be utilized again since these are actual service times.

Table 4. Standard Time of Nurse's Activities

Department	Selected Time (min)	Pace Rating (%)	Normal Time (min)	Personal Allow-ances (%)	Personal Allow-ances (min)	Standard Time (min)	Total Stan-dard Time (min)	Service Time (patients/ hr)
MEDICINE	1.03	100.00	1.03	2.50	1.05	2.08		
	0.56	100.00	0.56	2.50	0.58	1.14		
	0.53	100.00	0.53	2.50	0.55	1.08		
	2.81	100.00	2.81	2.50	2.88	5.68	9.98	6.0
PEDIATRICS	0.22	100.00	0.22	2.50	0.23	0.45		
	0.73		0.73		0.75	1.48		
	3.09	100.00	3.09	2.50	3.17	6.26	8.19	7.3
OB-GYNE	0.96	100.00	0.96	2.50	0.99	1.95		
	0.44	100.00	0.44	2.50	0.45	0.88		
	0.77	100.00	0.77	2.50	0.79	1.56		
	1.43	100.00	1.43	2.50	1.47	2.90	7.29	8.2
SURGERY	0.83	100.00	0.83	2.50	0.85	1.68		
	0.43	100.00	0.43	2.50	0.44	0.87		
	0.89	100.00	0.89	2.50	0.91	1.80		
	2.81	100.00	2.81	2.50	2.88	5.08	9.43	6.4

The summary of values needed in the computation of the new operating characteristics is presented in Table 5. The number of doctor servers was increased to three for all departments and the doctor-intern was instructed to start earlier than the actual post-lunch arrival of two o'clock with the approval of the OPD head. This was possible because of the increase in the number of clerks and interns at the time of the study.

Table 5. Proposed Models and Waiting Times

Departments	Patient Arrivals (Patients/hr)	Serving Time Nurse (Patients/hr)	Serving Time Doctor (Patients/hr)	Expected Waiting Time (min)	Proposed Number of Servers	
					Nurse	Doctor
MEDICINE	7	6.0	3.75	16.19	2	3
PEDIATRICS	7	7.3	4.6	11.20	2	3
OB-GYNE	7	8.2	3.5	28.63	2	3
SURGERY	7	6.4	4.3	11.56	2	3

The new operating characteristics that were computed for each of the departments for both nurse and doctor are shown in Table 6.

Table 6. New Operating Characteristics

Nurse Notation	Description	Medicine	Pediatrics	Ob-Gyne	Surgery	Unit
L	Average number of patients in the system (waiting and being served)	1.77	1.25	1.04	1.56	Patient/
Lq	Average number of patients in the waiting line	0.6	0.29	0.19	0.47	Patient/s
W	Average time a patient spends with the nurse (waiting and being served)	15.16	10.67	8.95	13.38	Minutes
Wq	Average time a patient spends waiting in line	5.16	2.45	1.63	4	Minutes
P0	Probability of no (zero) patient waiting in the system	0.26	0.35	0.4	0.29	

ρ	Utilization rate; the proportion of time the system is in use	74	65	60	71	%
Doctor Notation	**Description**					
L	Average number of patients in the system (waiting and being served)	2.7	2.7	5.09	2.3	Patient/s
Lq	Average number of patients in the waiting line	1.1	1.06	3.15	0.81	Patient/s
W	Average time a patient spends with the doctor (waiting and being served)	27.03	21.79	43.67	21.52	Minutes
Wq	Average time a patient spends waiting in line	11.03	8.74	27	7.56	Minutes
P0	Probability of no (zero) patient waiting in the system	0.09	0.36	0.09	0.2	
ρ	Utilization rate; the proportion of time the system is in use	91	64	91	80	%
Total Waiting Time		**16.19**	**11.19**	**28.63**	**11.56**	**Minutes**

To have a better view of the interaction of the models, a graphical presentation is shown in Figure 2. As can be seen in the figure, the calling population consists of many prospecting patients classified as infinite. When the patient has decided to join the queue, he is now part of the OPD queuing system. The queue was treated finite since the study was not continuous; instead it was treated on a daily arrivals basis. Once the patient has registered with the medical records staff, the patient queues up in the specific department visited. The patient waits until a nurse gets his/her papers and notes his/her vital signs. After being seen by the nurse, the patient waits for a doctor. There should be only one queue for the doctors so that the first come - first served policy will still hold. This signifies a shift from the previous model whereby the patient was seen by only one set of nurse and doctor and where s/he could have possibly been seen either earlier or later than the other queuing members.

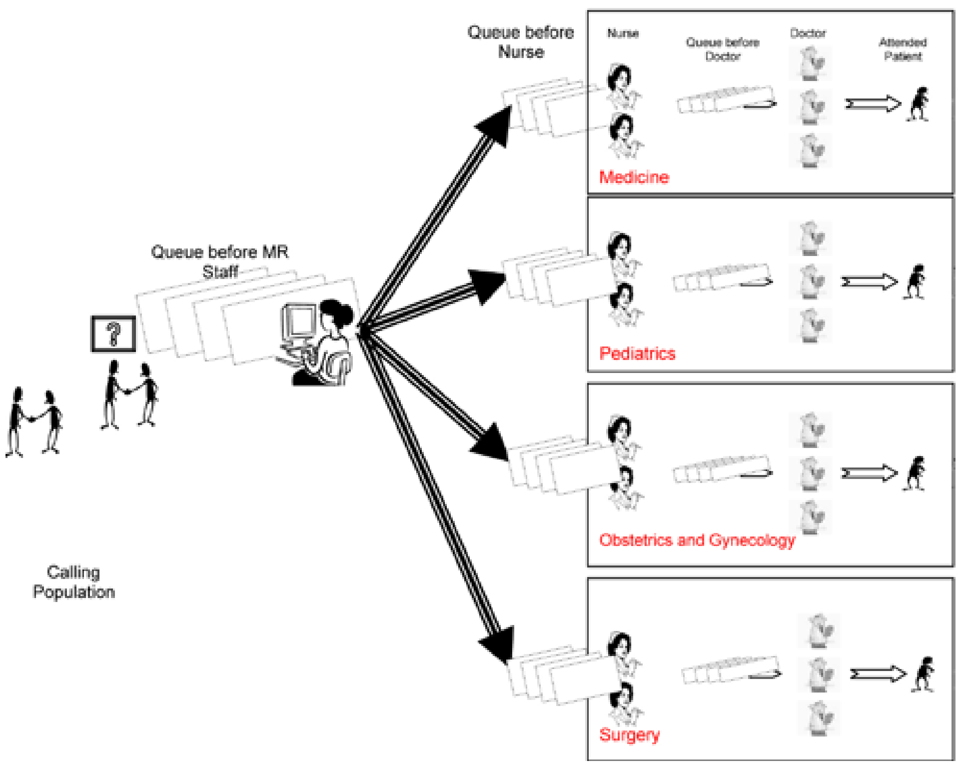


Figure 2. The Proposed Queuing Models

The number of nurses is still two but the number of doctors was increased to three to obtain an optimal combination of trade-off between patient's waiting time and the institution server utilization. Conducting another time study, new data were gathered for the proposed model and these new models have definitely improved the patients' waiting time.

Mathematical model

It is also possible to describe the queuing system by means of a mathematical model called Aspiration–Level Model (Taha, 1997). This method recognizes the difficulty of estimating cost parameters that is why it utilizes available data for analysis. Optimality is viewed in the sense of satisfying certain aspiration levels set by the management. This aspiration level is defined as the upper limits on the values of the conflicting measures that the decision maker wishes to balance. In a multiple server model such as the OPD queuing system of ITRMC, the conflicting measures determined are average waiting time in the system (denoted by W) and percentage of idle time of servers (Po). These two reflect the aspirations of the customer (patients) and the server (ITRMC OPD). The aspiration-level method can be expressed mathematically as follows:

$$W \leq \alpha \qquad \text{and} \qquad Po \geq \beta$$

Trade-off tables are shown in Table 7. To use such table, the waiting time should be set at less or equal to 30 minutes or 0.5 hour waiting time to match global standards. As can be gleaned from the table, the Departments of Surgery and Pediatrics can deploy two doctors and two nurses and can still meet the maximum waiting time of 30 minutes while the Departments of Medicine and Ob-Gyne cannot do the same. If the hospital management sets a percentage idle time of 15%, 3 doctors for Medicine, 2 for Pediatrics, 3 for Obstetrics-Gynecology and 2 for Surgery will be considered optimal. In the same manner, 2 nurses for Medicine, 2 for Pediatrics, 1 for Obstetrics-Gynecology and 2 for Surgery will also be optimal.

Table 7. Aspiration - Level Table, Medicine and Pediatrics

Department	Server	Number				
MEDICINE	*Nurse*	**1**	**2**	**3**	**4**	**5**
	%age Idle Time	0	26	34	38	40
	Waiting Time (hrs)	∞	0.09	0.02	0.01	0.01
	Doctor	**1**	**2**	**3**	**4**	**5**
	%age Idle Time	0	11	23	27	30
	Waiting Time (hrs)	∞	0.47	0.08	0.03	0.02
	Total Waiting Time (hrs)	∞	**0.56**	**0.10**	**0.04**	**0.03**
PEDIATRICS	*Nurse*	**1**	**2**	**3**	**4**	**5**
	%age Idle Time	4	35	42	44	46
	Waiting Time (hrs)	3.06	0.04	0.01	0.01	0
	Doctor	**1**	**2**	**3**	**4**	**5**
	%age Idle Time	0	12	23	28	30
	Waiting Time (hrs)	∞	0.37	0.06	0.03	0.01
	Total Waiting Time (hrs)	∞	**0.41**	**0.07**	**0.04**	**0.01**
OB-GYNE	*Nurse*	**1**	**2**	**3**	**4**	**5**
	%age Idle Time	15	40	46	48	49
	Waiting Time (hrs)	0.61	0.03	0.01	0	0
	Doctor	**1**	**2**	**3**	**4**	**5**
	%age Idle Time	0	1	15	21	24
	Waiting Time (hrs)	∞	4.79	0.14	.05	0.03
	Total Waiting Time (hrs)	∞	**4.82**	**0.15**	**0.05**	**0.03**
SURGERY	*Nurse*	**1**	**2**	**3**	**4**	**5**
	%age Idle Time	0	29	37	40	42
	Waiting Time (hrs)	∞	0.07	0.02	0.01	0.01
	Doctor	**1**	**2**	**3**	**4**	**5**
	%age Idle Time	0	15	25	30	32
	Waiting Time (hrs)	∞	0.29	0.06	0.02	.01
	Total Waiting Time (hrs)	∞	**0.36**	**0.08**	**0.03**	**0.02**

Conclusion and recommendations

There is a significant difference between the actual waiting time and the expected waiting time of the patients for the four departments. The expectations are definitely lower than the actual experience of the patients, which is an indication that the queuing models should be improved.

To further substantiate the claim on the length of waiting time, the existing models were investigated. The operating characteristics computed for each of the departments indicate that the Departments of Obstetrics-Gynecology and Surgery have very high waiting times of more than one hour, while the Departments of Medicine and Pediatrics have lower waiting times not exceeding the European acceptable standards but still higher than the expectation of the patients..

On the basis of the existing models' operating characteristics, new models were conceptualized. The recommended models should have a capacity of 2 nurses and 3 doctors wherein the nurses are available for the entirety of the OPD operation time of 8 o' clock in the morning to 5 o' clock in the afternoon and one doctor for each department is made available at 1 o' clock in the afternoon. In addition, the patients are to follow a single queue for the nurses and another single queue for the doctors.

Given the existing queuing models set-up, the following are recommended: that all the departments increase their capacity by at least one doctor server for their Monday operations; start one hour before the scheduled consultation time to prevent the excessive build-up of patient arrivals; that in the event that a doctor cannot see the registered patients within a 30-minute period, information must be disseminated immediately to the waiting patients about the cause of the delay for them to decide early on whether to wait or just come back the next consultation day and that when a patient particularly requests for a specific doctors, he will be informed of the time the doctor may be available to decide early on whether to wait or avail of other available doctors. And lastly, a further study can be conducted on the optimality of patient flow during the service encounter.

Acknowledgments

This research was funded by Commission on Higher Education – Thesis Grant.

References

Barnes, Ralph M. (1980). *Motion and Time Study: Design and Measurement of Work*. 7th edition. New York: John Wiley and Sons.

De Man, Stefanie. (2004). Employee-Related Antecedents of the Waiting Experience and Service Quality in Outpatient Clinics (Online Version), unpublished Dissertation, Faculty of Economics and Business Administration of Ghent University, accessed September 2006 at http://www.feb.ugent.be/fac/research/proefschriften/deman_s_proefschrift.pdf.

Downie, N.M. and Heath, R. (1983). *Basic Statistical Methods.* New York: Harper and Row Publishers.

Standard Patient's Charter of England (online version), accessed August 15, 2006, from www.pfc.org.uk/node/633

Stevenson, William J. (2002). *Operations Management*. 7th edition. Boston: McGraw Hill Company.

Taha, Hamdy A. (1987). *Operations Research: An Introduction.* New York: Macmillan Publishing Company.

Annex A

Kolomogorov-Smirnov Dncritical Table

Sample Size (n)	Level of Significance for $D = \text{Maximum } \lvert F_e - F_o \rvert$				
	0.20	0.15	0.10	0.05	0.01
1	0.900	0.925	0.950	0.975	0.995
2	0.684	0.726	0.776	0.842	0.929
3	0.565	0.597	0.642	0.708	0.828
4	0.494	0.525	0.564	0.624	0.733
5	0.446	0.474	0.510	0.565	0.669
6	0.410	0.436	0.470	0.521	0.618
7	0.381	0.405	0.438	0.486	0.577
8	0.358	0.381	0.411	0.457	0.543
9	0.339	0.360	0.388	0.432	0.514
10	0.322	0.342	0.368	0.410	0.490
11	0.307	0.326	0.352	0.391	0.468
12	0.295	0.313	0.338	0.375	0.450
13	0.284	0.302	0.325	0.361	0.433
14	0.274	0.292	0.314	0.349	0.418
15	0.266	0.283	0.304	0.338	0.404
16	0.258	0.274	0.295	0.328	0.392
17	0.250	0.266	0.286	0.318	0.381
18	0.244	0.259	0.278	0.309	0.371
19	0.237	0.252	0.272	0.301	0.363
20	0.231	0.246	0.264	0.294	0.356
25	0.21	0.22	0.24	0.27	0.32
30	0.19	0.20	0.22	0.24	0.29
35	0.18	0.19	0.21	0.23	0.27
Over 35	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Note: The values of D given in the table are critical values associated with selected values of n. Any value of D that is greater than or equal to the tabulated value is significant at the indicated level of significance.

Source: Downie, N.M. and Heath, R.(1983), Basic Statistical Methods

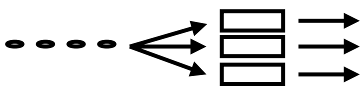
Annex B

Basic Queuing Models (Stevenson, 2002)

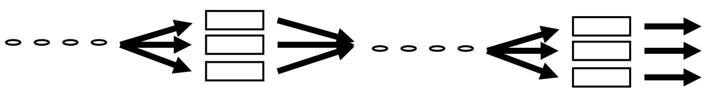
Single Channel Single Phase

Single Channel Multiple Phase

Multiple Channel Single Phase



Multiple Channel Multiple Phase



Annex C

Operating Characteristics and Summary of Formulas Used
in the Computation of Operating Characteristics

Multiple-Server Model

$$P_0 = \frac{1}{\left[\sum_{n=0}^{s-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right] + \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s\left(\frac{s\mu}{s\mu - \lambda}\right)}$$

$$P_n = \begin{cases} \frac{1}{s!\,s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n P_0, & \text{for } n > s \\ \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n P_0, & \text{for } n \leq s \end{cases}$$

$$P_w = \frac{1}{s!}\left(\frac{\lambda}{\mu}\right)^s \frac{s\mu}{s\mu - \lambda} P_0$$

$$L = \frac{\lambda\mu(\lambda/\mu)^s}{(s-1)!(s\mu - \lambda)^2} P_0 + \frac{\lambda}{\mu}$$

$$W = \frac{L}{\lambda}$$

$$L_q = L - \frac{\lambda}{\mu}$$

$$W_q = W - \frac{1}{\mu}$$

Source: Stevenson, William J. (2002). Operations Management 7th edition