Application of a patient flow model to a surgery department

Application of a patient flow model to a surgery department

Dario Antonelli
Dept. Production Systems and Economics
Politecnico di Torino
Torino, Italy
dario.antonelli@polito.it

Teresa Taurino
Dept. Production Systems and Economics
Politecnico di Torino
Torino, Italy
Teresa.taurino@polito.it

***Abstract*—Surely the main performance index for a surgical hospital division is the utilization rate of the operation room. In the non emergency divisions, when interventions can be planned in advance, the maximization of the utilization index is accomplished by a careful scheduling of the patient arrivals into the hospital. In this paper a model of patient flow is built, indentified and applied to demonstrate that the optimization of individual stages of the process is impossible without a concurrent control of the entire routing from the incoming to the dismissal. Using data from a real hospital division we simulate the effect of different allocation schemes on the performance of the operation room.**

Health care systems; queueing networks, patient flow, stochastic processes

I. INTRODUCTION

The application of managerial models to the healthcare system is increasing in order to keep under control the exploding costs of the health and to improve the efficiency of the system itself with a concurrent improvement of the service level.

The main performance indexes that are monitored, and possibly enhanced, are: the utilization of surgical rooms, beds and other critical system resources, the waiting time and the queue length, the efficient utilization of staff time. In present paper we focus our attention on the optimization of surgical room utilization in the non emergency hospital divisions. There is an obvious distinction between surgical rooms assigned to emergency operations and the others. In this last case, it is possible to schedule the patient arrivals in order to maximize the utilization of the resource, provided that there is a number of patients sufficient to saturate the surgery capacity.

If the problem is analyzed using the methods of processes analysis we realize that this performance index alone can lead to deceiving results as there are other correlated factors that must be taken into account. Firstly, a rule of production planning says that in a stochastic process there is an utilization rate that can be safely attained. Above this value, small increase in utilization correspond to a large increases in the cycle time and in the work in process (WIP) that in terms of hospital recovery means the time to dismissal and the number of patients in the department's beds. The maximum use of the operation room is obtained only at the expense of the other hospital resources. Secondly, another rule says that, in a production sequence, the maximum throughput of a system is the bottleneck rate, that is the rate of the process with the maximum utilization. In terms of the hospital division this means that if the beds used for the recovery "post intervention" are saturated, the rate at which it is possible to make surgical interventions is modulated by the rate of dismissals of patients from the hospital beds, no matter how productive is the surgery staff.

Therefore the operation room utilization must be regarded as only one performance parameter that cannot be optimized by itself without considering the whole routing that the patient make through the hospital facilities. For this reason a model of the patient flow has been proposed in Section II, using the queueing network. The model is instanced in Section III by applying parameters extracted from actual data collected from the Division of General Surgery in a Italian Hospital during a one year time period. The collected data allow to find reliable values for the time distributions used by the stochastic model. The model is used to simulate a different bed allocation scheme, Section IV, in order to verify its usefulness to improve the operation room utilization. The scheme recommends the generalized adoption of partial hospitalization for the diagnosis and the separation of beds in two batches: limiting the beds dedicated to long terms patients.

II. MODEL DEFINITION

We need as first thing to give a definition of some terms diffusely used in process analysis, in order to apply them to the healthcare system [1].

Routing is the sequence of workstations that a job must pass through in order to be produced. It is the patient flow from arrival to dismissal. Unfortunately patient flow is far from linear, with frequent transfers to and from other hospital divisions. Throughput rate is the average output of the process per unit time. Considering only one Division and not the whole

978-1-4244-4998-9/10/$25.00 ©2010 IEEE

Hospital we will relax the constraint that throughput rate be equal to the arrival rate unless system instability. Capacity of every process is the maximum throughput attainable. The Work in Progress (WIP) is the number of patients hospitalized. Eventually utilization rate is defined as:

$$u = \frac{arrival\ rate}{capacity} \quad (1)$$

The patients enter randomly in the hospital and the healing time follows a stochastic distribution with high variability. Patient flow has been extensively studied in literature. Models have been proposed for inpatient and outpatient flow, for general hospital services, for chronic patients assistance and for emergency departments. For every case there is a model that is more effective in describing the problem. For the present case of patient flow in an hospital Department, there is a wide concordance on the use of Markov chains describing the flow through health states [2-5]. Nevertheless, in present study a queueing network model was preferred. The reason is that the study try to highlight the correlation between two parameters measured on different time scales. The waiting lists is made of patients waiting at home their entrance into the hospital, the elapsed time is in the order of weeks, sometime months. In all the models usually the exact time of the arrival day is of no significance. The utilization rate of the operation room makes reference to the number of hours during the day the room is used divided by the available hours.

The necessity of a model that combines patient flow and operation room scheduling is explained by this consideration: as far as a long term patient occupy a bed, one cannot have a new arriving patient and thus operate her/him. In other words, the allocation rule for the beds can have a strong influence on the running of the operation room without the occurrence of starving cases.

The model of the patient flow takes the form of a queueing network with G/G/m servers. The interarrival times and the process times follow a general distribution. There are m workstations in the server and the queue, intended as the waiting list, is virtually unlimited.

The Poisson stochastic distribution is adopted for patient arrivals because the number of patients considered is very high and memoryless: every arrival is independent from the others. Obviously patients do not base the decision to undergo surgical operations on the basis of other patients' diseases. The distributions of the service times variables are found experimentally from the hospital data, as described in the next Section. The services are: the pre-operation hospitalization full or partial for the users whose disease has already been diagnosed, the operation, the post intervention recovery. The entities using the services are the patients. Services consume some resources, available in limited amounts. Among them our model considers: beds, doctors and nurses. A length limited buffer is used only to force virtually no queue after the intervention (buffer set to one). A further complication is the possibility of flows to and from other divisions. As told before, the pre-intervention is distributed over two services: recovery of patients with a complete diagnosis (they have made the clinical analysis before the recovery) and recovery of patients that need further laboratory exams. The bed resources is divided in two parts: beds dedicated to expected long stay patients (recovery time longer than two times the average) and beds dedicated to standard patients.

It is noteworthy that the resources are shared among the services. As an example, the bed occupied before the intervention is the same bed that will be used during the recovery post intervention.

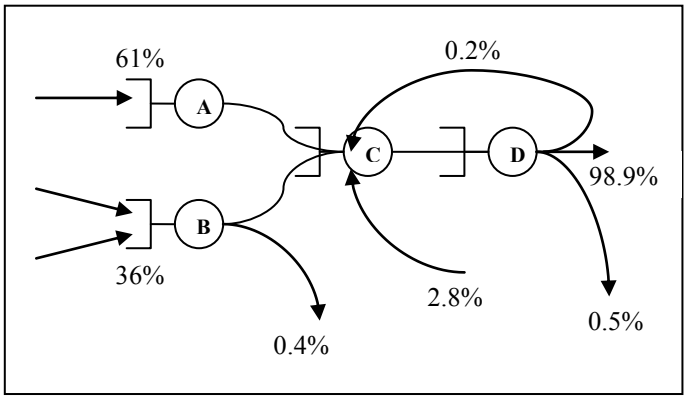

Figure 1. Queueing network model of patient flow.

Figure 1. shows the different possible patient routings through the hospital facilities, representing them as a queue network. The state **A** represents Partial Hospitalized Patients, namely patients that continue to reside at home during the diagnostic process and are admitted in the hospital the day of the surgical intervention; the state **B** represents patients that are hospitalized for the diagnostic process; state **C** represents the surgical room; state **D** represents the recovery.

III. ANALYSIS OF THE EXPERIMENTAL DATA

Patients flows have been obtained from the SDO (Scheda di Dimissione Ospedaliera – Hospital Discharge Board) data base of the hospital accesses of the Ospedale Cardinal Massaia in Asti. The considered year has been the 2007.

To obtain distributions of the lengths of patients stay in each state, data concerning chirurgical patients during the year have been considered, taking into account their days for acceptances, recoveries, surgeries and discharges [8, 9].

Real data concerning days spent by a patient hospitalized before the surgery intervention have been obtained considering the interval between the day of admittance and the day of the surgical intervention; to obtain data about the Recovery length, number of days between the surgical intervention and the discharge day have been considered. For the distribution of surgical intervention duration, the daily number of intervention was available querying the data-base. Assuming that the operating room works 8 hours a day it's possible to obtain the intervention duration distribution.

Incorporating the real data values in the Input analyzer software of Arena a probability distribution to them was fitted.

To pick the best distribution the Kolmogorov-Smirnov (K-S) test was applied; once fixed the distribution, maximum likelihood method for estimating distribution parameters have been used. The K-S test is based on the empirical distribution function (ECDF). Given N ordered data points Y_1, Y_2, ..., Y_N, the ECDF is defined as

(2)

where *n(i)* is the number of points less than Y_i and the Y_i are ordered from smallest to largest value. This is a step function that increases by *1/N* at the value of each ordered data point.

The Kolmogorov-Smirnov test is defined by the null hypothesis :

H_0: The data follow a specified distribution.

The Kolmogorov-Smirnov test statistic is defined as

$$D = \max_{1 \le i \le N} \left(F(Y_i) - \frac{i-1}{N}, \right. \quad (3)$$

where F is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution. The hypothesis regarding the distributional form is rejected if the test statistic, D, is greater than the critical value obtained from a table of the Kolmogorov distribution. [10]

For hospitalized patients, to fit the distribution of number of days spent in the hospital before the surgical intervention, the best probability distribution with a square error of 0.007 is the Beta distribution of parameters:

$$0.5 + 22 * BE \quad (4)$$

Histogram of frequencies with real data and the estimated probability density function is shown in Figure 2, where classes represent days from 1 to 22, rectangle height is proportional to the frequencies and rectangle width is equal to one..

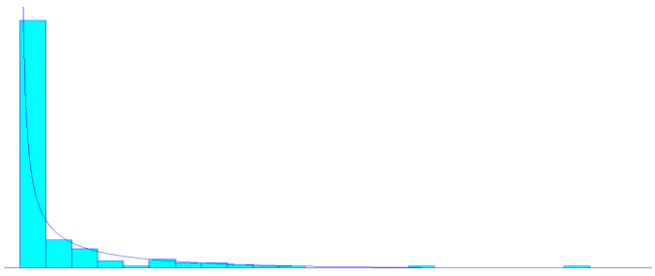

Figure 2. Distribution of days of recovery for hospitalized patients.

The recovery length can be of regular stay or long stay; it's impossible, of each patient, to know in advance the kind of his stay in the hospital. The rule used is to consider as regular stay patients who spend in the hospital a time less equal than

(5)

where μ is the average and σ is the standard deviation of recovery's lengths. The percentage of this kind of recovery is the 90% and the distribution of days of recovery is shown in the next histogram (Figure 3.); rectangle height is proportional to the frequencies while rectangle width is equal to one; classes represent days of recovery from 0 to 7.

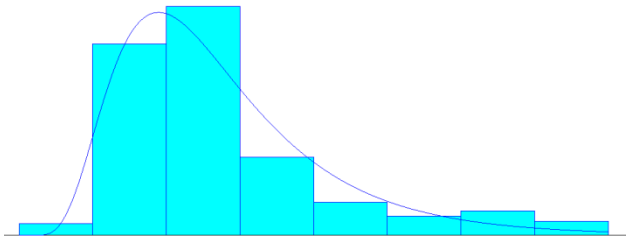

Figure 3. Distribution of days of regular recovery.

the fitted distribution is a log normal with expression

$$-0.5 + LOGN(2.76, 1.48) \quad (6)$$

while the expression for the long stay recovery is

$$7.5 + LOGN(5.33, 10.9) \quad (7)$$

And the histogram of the frequencies with real data and fitting distribution is in Figure 4.; rectangle height is proportional to the frequencies while rectangle width is equal to one; classes represent days from 8 to 35.

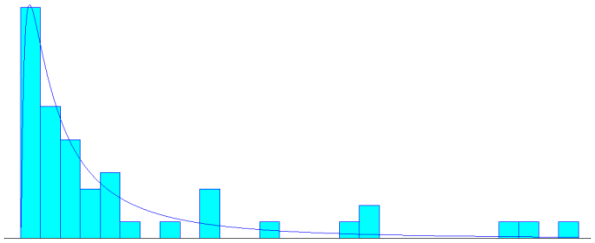

Figure 4. Distribution of days of long recovery.

In TABLE I. there is a summary of the distribution for each variable.

TABLE I. NETWORK PARAMETERS

Service	Distribution	Expression
A - Partial Hospitalized length (days)	costant	0
B – Hospitalized length (days)	Beta	0.5 + 22 * BETA(0.239, 3.3)
C - Surgical intervention	Exponential	EXP(2.44)
D – Recovery length (days)	Lognormal	standard -0.5 + LOGN(2.76, 1.48) (90%)
		long terms 7.5 + LOGN(5.33, 10.9)

Service	Distribution	Expression
		(10%)

The throughput of inpatients in the surgical division is 516 patients/year and the resources are 14 beds (equally divided in rooms for male and rooms for female patients). The number of doctors and nurses is variable along the considered year. The queueing network is solved by Discrete Event Simulation performed on a Rockwell Arena simulation engine. The simulation experiment requires the definition of the key elements of the system and their interrelationships [6-7].

As apparent form the flows of Figure 1, the flow of patients between Departments, in both directions in and out the examined Department, is negligible, exception made for the flow to the surgical room. This is not a common situation. It can be estimated that in the overall hospital the inter-Departments flows be in the order of 20%.

TABLE II shows the results of an experiment made of 30 simulations on a time range of 60 days, after an initial transitory of 30 days. The performance parameters are reported as average value and as range of variability along the 30 runs. Variability is significant, as can be expected from the scattering of the time distributions. Only the working days are considered, as the surgery is not working on holidays and the simulation is surgery centered.

TABLE II. PERFORMANCE PARAMETERS WITH ACTUAL DATA

Parameter	Description	Average Value	Half width
Patients out	Number	146	2.71
Cycle time	Days to dismission	9.08	1.07
Effective time	Days in bed	4.68	0.11
Wait time partial hospitalized (A)	Days before hospitalization	3.81	1.06
Wait time hospitalized (B)	Days before hospitalization	3.88	1.09
WIP	Patients inside the system	23.6	3.77
Number in queueA	Patients in the waiting list	6.43	1.81
Number in queueB	Patients in the waiting list	3.64	1.11
Beds utilization	Rate in / No. beds	0.96	0.02
Surgery utilization	Rate in / capacity	0.84	0.02

The results give an interesting insight in the functioning of the Division. The waiting list is an average 7.8 days and is larger than the effective hospital time that is only 4.70. The number of patients waiting is significantly higher of the number of patients hospitalized: 20 waiting against 12 in beds. In other terms the WIP is high. This is reasonable if we consider the utilization rate of the beds that is near to 1. The beds are always occupied and they represents the true bottleneck of the system. The utilization of the surgery is a good 0.79 (despite the hospital has a target of 0.9) but it is impossible to increase its value, having the bottleneck process, that is the recovery, reached the maximum utilization.

The bottleneck is caused by the recovery process, while the hospitalization plays a non significant role. If we compare the waiting time of both process A and B, we find the same value: 7.8 days. As the effective process time of B is nearly exactly the double of the A process time, this means that the number of people waiting in queue before A should be nearly the double of the number waiting before B. An this is what happen in the experiment as can be seen from TABLE II.

From the results the only solution to increase the utilization of the surgery seems to pass from the elimination of the bottleneck by increasing the number of the beds. For apparent economic reasons, it is advisable to explore other strategies before this. A possible strategy considers the possibility of differentiate the queues among regular and long terms patients. The long terms patients seize the bed resources for a time quite longer than the regular patients. They lead the production flow to blocking. Next section will investigate the possibility of selectively increase the long term patients queue for the sake of reducing the average effective recovery time and consequently increasing the surgery utilization.

IV. MANAGING THE QUEUE

The presence of queues is not a peculiarity of the healthcare system. In every service system, a queue is caused by demand variability and by market dynamics. Randomness in the demand of a service is managed by increasing the productive capacity in order to satisfy the peaks of demand. If the system is stable, the average queue length in terms of waiting time and waiting persons does not increase with time. A condition to assure stability is that the capacity of the service be in excess of the arrival rate.

It is important to differentiate the queue, depending on the time span, in short period queue and in waiting list. The queue time is the time the patient wait inside the service facility before being served. It is hopefully in the order of minutes and can be reduced by improving the scheduling schemes applied. The waiting list is made of patients that are at home and not in the service facilities and that are waiting to be called in order to be hospitalized. The waiting time now is measured in the order of days. The management of the waiting list is a matter of aggregate programming and need different strategies [12-14].

Queue should be reduced as much as possible and can be avoided in the hospital Departments that have a negligible number of emergencies by a proper schedule of the accesses. On the contrary waiting list are not only an issue but can be seen as a tool to manage the offer of services. The demand of health care services is a tradeoff between the expected benefits and the costs of the cure [11]. If the healthcare has no direct costs for the patients, as in the public hospitals, the tradeoff is obviously unfavorable to the service provider and the demand is always in excess. That is, all the people that expect a minimal benefit from the cure will make demand for the service. This is obviously not sustainable from the system. The presence of a waiting list introduces a form of cost that balance the expected benefits. Therefore if the waiting list is not

increasing with time, that is the case of an unstable system in which the capacity of the offered service is inadequate, it is preferable not to modify the offer but to discipline the queue. The discipline is the protocol of access to the service that is corresponding to the individual waiting time in queue. Presently hospital have a priority scale in order to hospitalize at once the acute and then the other patients in the order of the severity of the disease. The number of waiting persons is not diminished but the order with which the patients enter in the system is changed.

It is worth of consideration the possibility to separate the beds dedicated to standard stay patients from the beds dedicated to long terms patients. Long terms patients occupy for a longer time the beds during the recovery. As the beds for the admission to the hospital are obviously the same, they delay new arrivals to the surgery. It is possible that sometime nearly all the beds be occupied by long terms and this would put the system to a blocking state. If the hospital differentiates the beds, a flow of new standard patients will be always guaranteed. In TABLE III the new solution is simulated with 2 beds dedicated to long stay and 12 to the standard patients. It is reproduced the proportion measured between short and long stay patients.

TABLE III. PERFORMANCE PARAMETERS WITH SEPARATED BEDS

Parameter	Description	Average Value	Half width
Patients out	Number	145	3.01
Cycle time	Days to dismission	8.47	0.83
Effective time	Days in bed	4.5	0.07
Wait time partial hospitalized (A)	Days before hospitalization	2.20S	1.03
		26.7L	5.59
Wait time hospitalized (B)	Days before hospitalization	2.22S	1.0
		24.0L	6.91
WIP	Patients inside the system	25.8	3.22
Number in queueA	Patients in the waiting list	3.46S	1.84
		4.76L	1.18
Number in queueB	Patients in the waiting list	2.00S	0.98
		2.75L	0.58
Beds utilization	Rate in / No. beds	0.91S	0.03
		0.99L	0.01
Surgery utilization	Rate in / capacity	0.84	0.02

The results are interesting but not revolutionary. As a matter of fact, the cycle time is reduced, the WIP is slightly increased and the surgery utilization rate is stationary. The price for the better overall performances are paid only by long terms patients that see a significant increase in the waiting times and a slight increase in the queue length. Conversely the standard patients see a significant reduction of their queues both in time and in number of persons waiting. The key factor is still represented by the utilization of the hospital beds that is near to 100%. This factor is a source of instability in the results: small changes of the input parameters, as the number of arrivals, can produce large change to the performance of the system. It is a serious limit to the possibility of intervention on the system. Nevertheless the solution of a different bed allocation proved to go in the expected direction: give a faster service to some patients that deserve it more. It is not the case of long terms patients, but the scheme of bed differentiation is open to other criteria of patient selection.

There is a weak point in the simulation. The allocation of patients in the real hospital can be made only on the basis of the estimated gravity of their disease (ex ante). It is possible that some patients have a faster recovery or a longer than estimated. In the simulation the patients have a recovery length that always fit in the corresponding distribution: long term patients will have a long term stay. The simulation uses to chose the patients length of stay the time distribution obtained from experimental data provided after the dismission (ex post). This fact could bring to differences between the simulated performances and the real ones.

V. CONCLUSIONS

The study provide a model of inpatient flow in a surgery division of an Italian hospital. The main differences with other studies is that the focus is on the non-emergency functioning of the hospital and therefore on the possibility to reduce queues by interventions on the scheduling of the patients arrival and on the allocation of resources. Another difference is the stress on the necessity to simulate at the same time both the short term (hours) behavior of the surgery as the long term (days) response of the recovery division. The model has been identified by using experimental data and has been used to verify the possibility of changing the allocation of beds. Obviously this is only one of the many possible management strategies that can be tested virtually before the practical implementation.

A possible improvement of the allocation scheme would update continuously the proportion of standard and long stay patients and would change the corresponding bed proportion. Present allocation scheme is static, i.e. it is independent from the actual value of patient flows. A better bed allocation scheme would be to dynamically assign beds based on their actual number. If there is an exceeding number of standard patients with respect to the available beds and there are some beds not used in the long term portion of the Department, the beds are re-allocated to the standard patients.

This scheme is surely more effective in the simulation but it is difficult to apply in the real world. The facilities of the rooms for short and long stay patients are different. Rooms are separated for males and females. It is not impossible to change allocation to beds but it is reasonable to think that the dynamic allocation should be slow to give time to the Department to rearrange the internal logistic.

REFERENCES

[1] Young T, Brailsford S, Connell C, Davies R, Harper P, Klein J H., 2004, Using industrial processes to improve patient care, The British Medical Journal (BMJ), 328, 162–164.

[2] Vissers J. (1998). Patient flow-based allocation of inpatient resources: A case study, European Journal of Operational Research, 105: 356-370. Elsevier Science.

[3] Coté M.J. (1999). Patient flow and resource utilization in an outpatient clinic. Socio-Economic Planning Sciences, 33: 231-245. Elsevier Science.

[4] T.J. Coats and S. Michalis. Mathematical modelling of patient flow through an Accident and Emergency department. Emergency Medicine Journal, 18:190–192, 2001.

[5] Xiong, H.H., Zhou, M.C., Manikopoulos, C.N., 1994. Modeling and Performance Analysis of Medical Services Systems Using Petri Nets. Proc. IEEE International Conference on Systems, Man and Cybernetics, 2339-2342L.G. Connelly and A.E. Bair. Discrete event simulation of emergency department activity: A platform for system-level operations research. Academic Emergency Medicine, 11(11):1177–1185, 2004.

[6] R. Davies and H.T.O. Davies. Modelling patient flows and resource provision in health systems. Omega: The International Journal of Management Science, 22:123–131, 1994.

[7] A. C. Virtue. Simulating accident and emergency services with a generic process model. Nosokinetic News, December 2005.

[8] V. Navarro, R. Parker, K.L. White, A stochastic and deterministic model of medical care utilization, Health Services Research 5 (4) (1970) 342–357.

[9] MacCarthy B L, Wasusri T., 2002, A review of non-standard applications of statistical process control (SPC) charts, The International Journal of Quality & Reliability Management, 19, 3, 295 -320.

[10] T.W. Anderson, L.A. Goodman, Statistical inference about Markov chains, Annals of Mathematical Statistics 28 (1) (1957) 89–110.

[11] Culyer J.G. –Cullis J.G. (1976). Some Economics of Hospital Waiting Lists in the NHS, Journal of Social Policy, 5 (3): 239-64.

[12] Villa, A., Bellomo, D. , Cassarino, I., 2005, "Uncertain demand and supply network management: application to a regional health care service", 16th IFAC World Congress, Prague (c-disk edited proceedings).

[13] Qi, E., Xu, G., Huo, Y, Xu, X., 2006. Study of Hospital Management Based on Hospitalization Process Improvement. Proceedings of the IEEE IEEM, 74-78.

[14] Kumar, A., Shim S.J., 2007, Eliminating Emergency Department Wait by BPR Implementation. Proceedings of the 2007 IEEE IEEM, 1679-1683.