Article Title Page

Hospital capacity management based on the Queueing Theory.

Author Details

Author 1 Name: Otavio Bittencourt
Department: Ciências Exatas e Sociais Aplicadas (DECESA)
University/Institution: Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA),
Town/City: Porto Alegre
Country: Brazil

Author 2 Name: Vedat Verter
Department: Desautels Faculty of Management
University/Institution: McGill University
Town/City: Montreal
Country: Canada

Author 3 Name: Morty Yalovsky
Department: Desautels Faculty of Management
University/Institution: McGill University
Town/City: Montreal
Country: Canada

Corresponding author: Otavio Bittencourt
Corresponding Author's Email: otavion@ufcspa.edu.br

☐ *Please check this box if you do not wish your email address to be published*

Acknowledgments: This research was conducted when the first author was a postdoctoral fellow at McGill's Desautels Faculty of Management and this was supported by the Coordination for the Improvement of Higher Education Personnel (CAPES/Brazil) (postdoctoral fellowship Grant No. 1653-09-1). The language review was funded by the Hospital de Clínicas de Porto Alegre Research and Event Incentive Fund (FIPE-HCPA) and by the Federal University for Health Science of Porto Alegre (UFCSPA). The authors would like to acknowledge the editor and anonymous reviewers of IJSOM for their valuable comments and suggestions.

Biographical Details:

Otavio Bittencourt is Assistant professor at Federal University for Health Science of Porto Alegre. Postdoctoral fellow at McGill's Desautels Faculty of Management, doctorate of Production Engineering, master's degree in Management, and Bachelor of Business Administration at Federal University of Rio Grande do Sul. He has experience in management and academic, focusing on healthcare management, strategic planning, cost management, and service operations management.

Vedat Verter is James McGill Professor, Editor-in-Chief of Socio-Economic Planning Sciences. He is the Founding Director of the Canada-wide NSERC CREATE Program on Healthcare Operations & Information Management. He serves as Director of the McGill MD-MBA Program. Since 2012, he has served as the Associate Director (Health) of the Marcel Desautels Institute for Integrated Management.

Morty Yalovsky is associate Professor of Management Science at the Faculty of Management, McGill University. Associate Dean Academic at McGill University, Area Coordinator of Operations Management at Desautels Faculty of Management. He had several positions at McGill University, such as Vice-Principal (Administration and Finance) and Dean of Continuing Education.

Structured Abstract:

Purpose
This paper focuses on the contributions of Queueing Theory to hospital capacity management to improve organizational performance and deal with increased demand in the healthcare sector.

Design/methodology/approach
Models were applied to six months of inpatient records from a University hospital to determine operation measures such as utilization rate, waiting probability, estimated bed capacity, capacity simulations and demand behavior assessment.

Findings
Irrespective of the findings of the queueing model, the results showed that there is room for improvement in capacity management. Balancing admissions and the type of patient over the week represent a possible solution to optimize bed and nurse utilization. Patient mixing results in a highly sensitive delay rate due to length of stay (LOS) variability, with variations in both the utilization rate and the number of beds.

Practical implications
The outcomes suggest that operational managers should improve patient admission management, as well as reducing variability in length of stay and in admissions during the week.

Originality/value
The Queueing Theory revealed a quantitative portrait of the day-by-day reality in a fast and flexible manner which is very convenient to the task of management.

Keywords: operations management, queueing theory, capacity, hospital, health services sector.

Article Classification: Research paper.

For internal production use only

Running Heads:

Hospital capacity management based on the Queueing Theory.

Abstract

Purpose
This paper focuses on the contributions of Queueing Theory to hospital capacity management to improve organizational performance and deal with increased demand in the healthcare sector.

Design/methodology/approach
Models were applied to six months of inpatient records from a University hospital to determine operation measures such as utilization rate, waiting probability, estimated bed capacity, capacity simulations and demand behavior assessment.

Findings
Irrespective of the findings of the queueing model, the results showed that there is room for improvement in capacity management. Balancing admissions and the type of patient over the week represent a possible solution to optimize bed and nurse utilization. Patient mixing results in a highly sensitive delay rate due to length of stay (LOS) variability, with variations in both the utilization rate and the number of beds.

Practical implications
The outcomes suggest that operational managers should improve patient admission management, as well as reducing variability in length of stay and in admissions during the week.

Originality/value
The Queueing Theory revealed a quantitative portrait of the day-by-day reality in a fast and flexible manner which is very convenient to the task of management.

Keywords: operations management, queueing theory, capacity, hospital, health services sector.

1. Introduction

Society is experiencing a transient epoch in which old problems coexist with new ones. The new shape of the demographic pyramid, with growing aging populations and a decrease in birth rate, increased prevalence of chronic disease, the persistence of infection diseases, the need for more technology and resources and the increment of healthcare cost are concerns for decision makers in many countries (Martens and Huynen, 2003).

In spite of the rapid evolution of the healthcare management in the last thirty years, this situation has resulted in some gaps in research, such as new techniques that support hospital planning, including forecasting the length of stays and other aspects of hospital performance and new ways of staffing with regard to the number and type of nurses in hospitals (Edwards and Harrison, 1999). Such tools are necessary to help create healthcare systems that are safe, effective, patient-centered, timely, efficient and equitable through the use of techniques that measure and optimize system performance to meet performance goals (Reid et al., 2005). The identification of organizational

models based on factory and network focused concepts that provide a real integrated care system are essential to respond to population demands and at the same time deliver this change in a sustainable way (Karakusevic, 2010).

In Brazil, there was an 8.34% decline in the number of hospitals from 1999 to 2005, while the population increased by 12.34%. There was an increase of 16.3% in hospital admissions between 2002 and 2004, and a decline in hospital beds of 18.6% from 1992 to 2005 (IBGE, 2006). Public hospitals absorb 30–50% of government budgetary allocations in the health sector in both developed and developing countries (Barasa et al., 2015). In 2008, the aging population (over 65 years old) was 6.6%, with a tendency of growth. Between 2000 and 2005 hospital expenditures rose 87.7% (IBGE, 2008), while the rate of inflation was 39.8%. The majority of the Brazilian population (75.5%) receives healthcare via the public system, with most coming from low-income families, while high income families are likely to use the private sector (IBGE, 2006).

In addition to the many strategies that could be undertaken at a national level, managerial improvements at organizational level can cope with these issues. At the same time, hospitals are markedly different production systems than those from other industries and require adequate solutions (Morton and Cornwell, 2009). The training and education of hospital management in the public sector is a key element of an effective, timely and efficient healthcare system. Timely access to care is an important component of high-quality healthcare (Lakshmi and Iyer, 2013).

However, many hospitals are operated at a basic deterministic level, using averages and proportions as measures of bed allocations and forecasting bed requirements. Using average length of stay (LOS) alone to calculate future bed needs results in an underestimating of requirements and lacks the necessity detail (Harper and Shahani, 2002). Further development of operational-research capacity and the allocation of specific resources are needed for a more efficient healthcare setting (Zachariah et al., 2009). Hospitals are faced with a tradeoff between having available beds to meet patient demand and keeping bed occupancy (utilization) rates high (Lakshmi and Iyer, 2013).

A relevant approach to addressing capacity management in relation to the allocation of beds is the Queueing Theory. Applying this approach can mathematically formulate the current process and determine the point on the utilization curve that maximizes responsiveness and productivity (Terwiesch et al., 2011). Literature has shown that Queueing-based models are simple and useful tools for analyzing the bed allocation problem in a healthcare facility and provide accurate and rapid estimates of system performance (Lakshmi and Iyer, 2013). By identifying the probabilistic behavior of the arrival and service rates, one can make decisions about the satisfactory number of beds and level of service, identifying the target delay for each type of patient, service, or unit. In the short term, hospital decision-makers can decide how long a patient will have to wait, which disease to treat in each unit, and what might be an acceptable number of patients waiting in line. In the long term, they may be able to evaluate operational capacity and modify the number of beds or human resources.

This paper focuses on contributions of the Queueing Theory to hospital capacity management to improve organizational performance and deal with increased demand in the healthcare sector. We describe the application of Queueing Theory in the setting of a surgery and general ward unit of a public University hospital and demonstrate how to deal with the decisions described above by applying analytical methods. Some models have affordable solutions through the analytic method, as the derived mathematical expressions for operational measures are published in books or papers on operational management. Software packages are also available to compute the formulation. We also apply four queueing models, compute a range of outcomes, demonstrate how different

results can be obtained based on each assumption, study the number of required operational beds and explore the use of hospital resources (beds and nurses) according to the demographic data and length of stay of patients, as information to support decision-making processes in hospital operational management.

The remainder of this paper is organized as follows: a literature review, research design, demonstrations of Queueing Theory model for the general and surgery wards, and the last section, which presents conclusions and remarks.

2. Literature Review

In many developing countries, the need for productivity and performance improvement in the healthcare sector is an urgent issue by the development of effective approaches to reducing healthcare cost and increasing efficiency, without compromising on its quality (Bhat et al., 2016).

Hospital Operations Management addresses decision making at costs reduction, process improvement, productivity expansion and technology enhancement, by continuously look for the most efficient and optimal use of resources, to deal with the inconsistencies or dispersions of inputs and outputs in the healthcare process, but also to improve flow. Staffing and resource consumption should be tied directly with patient volumes and workload. One of the central point in the operations management is to understand the variability that exists in the patient demand, in-between patients and among practitioners, to evaluate the consequences to the capacity, workflow and average of patients that flow through the process (Langabeer II, 2008).

Next, we will present capacity and variability management, which contribute to these four dimensions of hospital operations strategy, and after basics concepts of Queueing Theory as a method to monitor and manage hospital capacity.

2.1. Capacity Management

In a stable process, the average inflow or outflow rate is called throughput. Capacity is the maximum sustainable throughput (Anupindi at al., 2012). Capacity is also defined as the maximum number of customers that can be served per unit of time (Terwiesch et al., 2011). In processes that produce a highly differentiated number of products or services, like healthcare service organizations, capacity during a specific period is then dependent on the mixture of products demanded during the period (Lantz and Rosén, 2016). In such cases, healthcare capacity is typically measured in terms of resources or inputs, beds, operating theatre time or slots, in order to deal with the variety of the patient/service mix. On the other hand, the patient characteristics and professional group related performance in the unit, measure by the average of length of stay, will determine the effective capacity. Therefore, capacity measurement provides the basis for the planning and control activities of the operation, providing information on available levels of activity over a set time period (Bamford and Chatziaslan, 2009).

Since expanding capacity is not feasible due to space limitations within hospitals, workforce shortage and government regulations, neither is necessarily desirable, the capacity management must match the demand with workload to optimize the workflow. It does not mean to operate at high level of utilization, where utilization is defined as the ratio of the number of customers served to the capacity, but to find a balance between efficiency and responsiveness (Terwiesch et al., 2011). Due to shortages and increased cost, managers must control the main healthcare resources, as it

demonstrates Ben-Gal and others (2010) by developing a model for physician staffing requirements. The complexity of healthcare operations is certainly a decisive factor for a problematic capacity measurement, so that a better understanding of underutilized resources not only lead to better capacity management reducing profit loss or unnecessary investments (Bamford and Chatziaslan, 2009), but also the need to use robust methods to measure the capacity.

The time elapsed from the point from which a physician orders an exam or procedure for a patient to the point at which his or her medical service starts is called timeliness, or patient waiting time. Long patient waiting times can potentially increase disease severity, resulting in more intensive treatments and higher costs (Liu and D'Aunno, 2012).

Studies elsewhere have identified sources of waiting time for patients, such as, the limited number of beds, physicians, and nurses, delays between admission referral by doctor, bed allocation, and patient transferal to allocated beds, and inefficient process in laboratory and radiology tests (Yuancheng et al., 2015). Investments in capacity often fail to increase overall output because they are not systematically directed at the real bottlenecks (Rechel et al., 2010). Addressing them will not only reduce the patient waiting time but also increase the revenue. Another important topic related to capacity management is the variability which is discussed in the next section.

2.2. Variability management

Variability is one of the main concerns to Lean thinking that leads the company efforts to create a dynamic process of change, integrated and driven by a systematic set of principles, practices, tools, and techniques that are focused on reducing waste, synchronizing work flows, and managing production flows (Bhat et al., 2016).

In healthcare, there is a high degree of variability, specifically for acute patients where the patient inflow concerning time, health issues and response to treatment is highly variable (Olsson and Aronsson, 2015). Variability is the enemy of operations, yet the risks associated with variability decrease as we aggregate many independent sources of variability (Terwiesch et al., 2011).

Besides the biological and intrinsic variation, there is the extrinsic variation created by the behaviors in healthcare systems such as discontinuous scheduling, variable capacity to discharge, and by splitting demand into groups (Olsson and Aronsson, 2015). It is suggested that the lean concepts often induce a change towards a process orientation by establishing proper planning and control of the patient flow.

McManus and others (2003) found the elective patient flow where more variable than the random demand of emergencies which indicates poor scheduling management. Therefore, there are two types of variability: natural, when it is related to disease and the arrival pattern of patients, and artificial, that are introduced by idiosyncrasies in the systems which would be controlled in the design and management of healthcare system.

Beds and wards become holding areas, which means patients stay not for treatment issue but to accommodate inefficient patient flow, where each hospital department seek to optimize their own functioning without considering how this affects the performance of others (Rechel et al., 2010).

Systems operating near capacity and with few scheduled admissions may benefit greatly by the control of variability (McManus, 2003). The inflow variability greatly accentuates problems of rejection and unused capacity, by the colliding of competing patient flow outstripping the supply of available beds. The variability management is a

promising area that managers may improve patient care without intruding on the specifics of clinical decision.

The key is smoothing patient flow by reconfiguring and decentralizing services, by for example, advances in medical imaging, spreading surgery evenly among the days of the week (Rechel at al., 2010).

According to Noon and others (2003), it may be more beneficial for the overall system to reduce the variability than to speed the tasks up. Variability may be caused also by lack of materials or information, for example, if an operating room knows ahead of time that a patient scheduled for an appendectomy has complicating conditions, an increased time for the procedure can be scheduled (Noon at al., 2003). From the point of view of Lean theory, variability is a kind of waste, and must be eliminated or reduced (Roemeling at al., 2017).

2.3. Queueing Theory

Queueing Theory is an advanced mathematical modeling technique that can estimate waiting times. In general, a queueing system has two main components; customers and servers. The former is seeking a service that can be provided immediately or otherwise by the server depending on the kind of service and the number of customers. If a customer must wait in line, it is referred to as a queue. Since customers arrive randomly and there is variability in the system, the delays they encounter are highly variable and depend upon the number of servers who are working and how fast they can work. A queueing model can be used to translate the arrival patterns and processing times to estimate important system performance measures, such as average customer waiting times and the likelihood of a random customer encountering zero delays, for any number of servers (Liu and D'Aunno, 2012).

If the system has sufficient capacity to deal with demand, waiting occurs primarily because of randomness or variability in the pattern of arrival of units and because of variability in the times required to service those units. Managing both these features can result in improvements in the system (Budnick et al., 1988).

The queueing process can be characterized basically by the arrival patterns of customers, the service patterns of servers, queue discipline, system capacity, number of service channels and number of service stages (Gross et al., 2008). There are many possibilities for each one of these, as well as other properties, but many authors argue that the Poisson arrival rate and exponential service time for many servers (M/M/s in notation) is the most common model for the healthcare services (Bruin et al., 2007; Green, 2003; Green and Nguyen, 2001; Green et al., 2006).

The M/M/s model requires some assumptions before being applied: a single queue with an unlimited waiting room that feeds into **s** identical servers (Green et al., 2006; Green et al., 2007; Budnick et al., 1988); servers operate independently of each other (Sztrik, 2016); arrivals occur according to a time-homogeneous Poisson process (λ) at a constant rate (Green et al., 2006); based on the previous assumption, arrival rate does not change over the day (Green et al., 2006); service time has exponential distribution (μ) (Green et al., 2006); patients are not assigned a bed in an alternative unit or are turned away if delays get long (Green and Nguyen, 2001); interarrival times, as well as service times, are assumed to be statistically independent (Gross and Harris, 1985, p.60); service discipline: the given operational characteristics apply to first-come first-served (FCFS), last-come first-served (LCFS), and service in random order (SIRO) (Budnick et al., 1988); arrival process: a single population with infinite number of units,

single arrivals with no control exercised by the queueing system, and a stationary arrival process exists (Budnick et al., 1988); queue discipline: no rejections (Budnick et al., 1988); service facility: servers are uncooperative and service times across channels are independent and identical (Budnick et al., 1988); the number of customers in the system is a birth-death process, where arrivals mean births and services mean deaths (Sztrik, 2016); as traffic intensity (ρ) equals λ/μ, it is a necessary condition that $\rho < s$, or $(\lambda/\mu s) < 1$, for a steady state to be achieved. In other words, the mean overall service time (μs) for the system must be greater than the mean arrival rate (Budnick et al., 1988).

Another variable is M/M/s/K, in which there is a limit K placed on the number allowed in the system at any one time. This approach is identical to the previous structure except the arrival rate λn must be zero whenever $n \geq K$, and there is no requirement that traffic intensity ρ is less than 1 (Gross et al., 2008, p. 76).

One useful model for the study of optimization in Queueing Theory comes from Erlang's Loss Formula (M/M/s/s), a special case of truncated queue, where no line is allowed to form (K = s) and which is valid for any M/G/s/s, independent of the form of the service time distribution. In other words, the steady-state system probabilities are only a function of the mean service time, not of the underlying CDF (cumulative distribution function) (Gross et al., 2008, p. 81-82). The length of stay (LOS) of an arriving patient is independent and identically distributed with the expectation μ. There is no waiting area, which means that an arriving patient who finds all beds occupied is blocked (Bruin et al., 2010).

According to Gross and others (2008), the M/G models consider a single-server queue with Poisson arrivals and general service distribution, where customers are served FCFS and all times and interarrival times are independent (p. 219).

The M/M/∞ model, with no limitations on service, has an interesting no-restriction characteristic for a steady-state solution to exist. The probability that any server is busy depends only on the mean service time and not on the form of the service-time distribution. This is also valid for any M/G/∞ model. The expected system size (L) is the expected number of customers in service (r), or the offered workload rate, so $L = r = \lambda/\mu$. As we have as many servers as customers in the system, the expected number of customers in line (Lq) is zero, as is the waiting time (Wq), so $Lq = 0 = Wq$. The average waiting time in the system is merely the average service time, so that $W = 1/\mu$, and the waiting time distribution function W(t) is identical to the service-time distribution, namely, exponential with mean $1/\mu$ (Gross et al., 2008, p. 84).

These assumptions do not fit for all patients. In the surgery and general unit, some patients have already been examined by their physicians previous to hospital admission, and the physicians have decided to send them to hospital. In this sense, this unit receives patients according to the availability of resources, such as physicians and surgery rooms. Although the availability of beds is not under the control of the physicians, it can be affected by the doctors, as many patients are submitted to a surgery. There is therefore in some cases a condition of dependence between the availability of beds and the arrival process.

3. Setting Description and Methodology

The healthcare sector is the world's largest service sector with total revenues of approximately US$ 2.8 trillion (Bhat et al., 2016). In Brazil, it represents 9.2% of GDP, approximately US$ 115 billion (IBGE, 2013). The healthcare system in Brazil is based around the Unified Health System (SUS – Sistema Único de Saúde, in Portuguese)

which offers free and universal coverage, and the private or public health insurance companies. SUS is a publicly funded healthcare system, organized around levels of complexity (primary, secondary, and tertiary), and all the administrative levels of the public sector (federal, state and municipality) fund the system in a shared funding system. The municipal level is in charge of the local administration of health system. The government income to provide for the healthcare system comes from taxation, goods tax, customs, financial services, or other sources.

Private or public insurance companies receive revenues from their insured clients. In September 2015, 25.9% of Brazilians were covered by private insurance (ANS, 2015). Even while representing a smaller proportion of the population, private sector spends were as much as the public sector.

The Hospital de Clinicas de Porto Alegre (HCPA), as a University hospital, provides services for SUS and the private and public insurance companies. Around 90% of inpatients are from the SUS, but thanks to the revenue obtained from the private system the hospital can acquire new equipment or renovate or build new facilities that benefit all patients. In this study, we worked with a surgery and general unit that receives private insurance patients.

Firstly, the correlation between LOS, time of care, gender, clinic, and age were explored, as patient characteristics and workload are related to capacity management. Secondly, descriptive statistics were identified and goodness of fit to arrival rate and service time were performed. Thirdly, queueing models were applied based on explanatory probability distributions, to obtain output parameters, and performance on different weekdays was performed. Finally, five scenarios were simulated by using the output parameters to verify the expected delay.

The stand-alone application of the EasyFit software distributed by MathWave Technologies was used to accomplish the appropriate probabilistic model. The computation of Queueing Theory models was carried out using QTSPlus Queueing Theory Software version 3, distributed by John Wiley & Sons, Inc. in connection with the book "Fundamentals of Queueing Theory," Fourth Edition, by Donald Gross, John Shortle, James Thompson and Carl Harris. Microsoft Excel and PASW Statistics 18 were used to compute statistical analysis.

4. The case study: General and Surgery Ward Unit

The general and surgery ward unit is dedicated to non-public healthcare system patients, where patients are usually assigned by their physicians, meaning the unit is not open to patients from the Emergency Department. Due to this very special situation, this unit was designated to take part in the development of a pilot information system, which implemented a complementary module of patient electronic records, by adding the care provided by the nurse, such as drug administration and procedures. These are registered electronically if performed by a nurse.

Operational data was obtained from the electronic patient records, such as the use of beds and activities performed by nurses. Some of these are recorded by the self-checking of nurses, if they performed the procedure (drug administration, nurse prescription and physician prescription), while others are recorded in the prescriptions ordered by physicians and nurses.

This unit receives general and surgery patients that have access to a wide range of hospital resources, such as physician specialties, nurse care, Intensive Care Units (ICU), Surgery Rooms (SR), Diagnostic Tests, and other professional specialists.

The database is from a sample of 525 inpatient admissions from October 2009 to March 2010, of which 58% were general clinical patients and 56% were male. This sample includes demographic data (age and gender), date/time of admissions, date/time of discharge, specialty, International Code Disease (ICD) and length of stay.

A subset from the previous sample adds data about the time spent by the nursing staff in caring for each patient in: drug administration, physician prescriptions, nurse prescriptions, blood transfusion, internal transport to the surgery room, to outpatient visits and to the diagnostic department. This second database has 471 patients, 58% general clinical, 57.7% female, and a mean age of 57.5 years.

In this period, the characteristics and resources of the ward were 25 beds, 72.2% bed utilization, 3285.7 patient days, 4550 bed days available (operational capacity), seven Registered Nurses (RN), and 21 Licensed Practical Nurses (LPN).

4.1. Understanding correlations between variables

In order to understand the main variables and to identify possible correlations, this section will consider the existence or otherwise of correlations between the variables in order to find possible explanations for the use of hospital resources, such as nurses or beds, by the admitted patients, based on the subset database. Diwas and Terwiesch (2009) found that resources in hospitals are sensitive to the levels of burden and that health service workers can adapt to system needs by expending more effort to increase the service rate as required.

Table 1 provides some descriptive statistics on the variables considered for the model, together with summary graphs for Total Care Time (Figures 1 and 2), where the existence of some differences can be noted, especially in terms of gender and clinic. Males tend to consume more Total Care Time and LOS than females in both clinics.

Table 1 around here

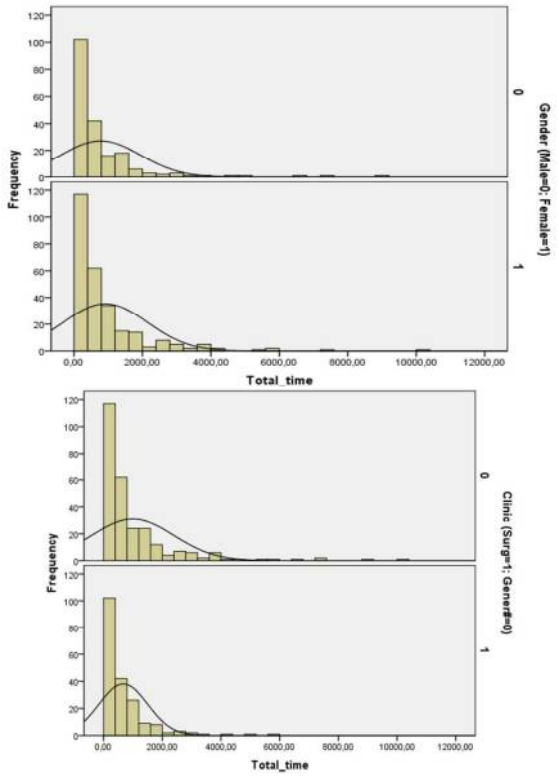

Figure 1: Histogram for Total Care Time (all patients) by gender and clinic.

We computed the correlation coefficient (r_{xy}) between Total Care Time, LOS and Age for all patients and under the two different clinical settings (Table 2). There is a similar correlation between LOS and Total Care Time in these three situations. Considering all the patients, there was a significant correlation between Age and Total Care Time, and no significant correlation between Age and LOS. However, if segregated by clinic, these correlations are only significant for the surgery clinic.

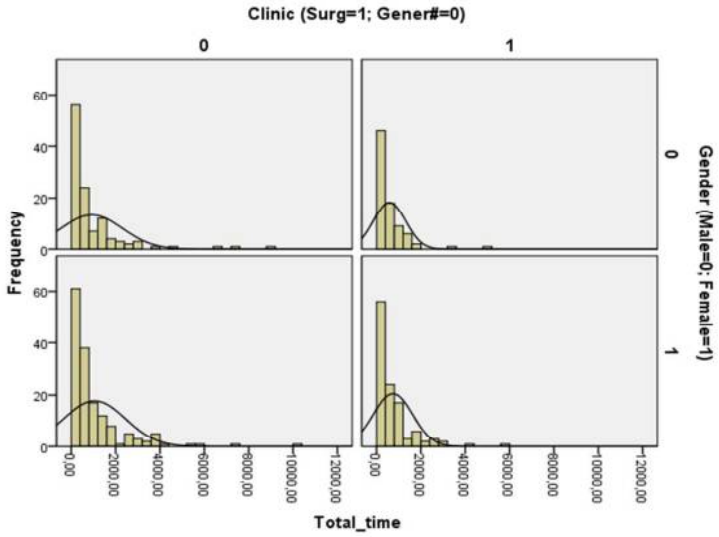

Figure 2: Histogram for Total care time (split by clinic) by gender.

Table 2 around here

The same analysis was performed by gender (Table 3). It can be seen that older patients receive more nurse care time, for both genders, a result that was significant for female patients. The same situation was identified for Age and LOS, and again was significant for females. There was a correlation between LOS and total nurse care time, following the pattern above.

Table 3 around here

4.2. Queue parameters

The arrival and service rates analysis was based on the database of 525 admitted patients. This database contains admission and discharge date/time in the unit, so from the frequency of admissions per day we defined categories of arrivals and verified the number of days in each category. At the same time, the service rate was obtained by the time between those dates or the length of stay (LOS) of each patient.

Descriptive statistics of arrivals indicated that the mean was 3.24 arrivals per day, the standard deviation was 1.94, and the coefficient of variation (Cv = standard deviation/mean) was 0.60, meaning this distribution was not as variable as it used to be in the healthcare process, where the Cv was very close to one or more (Green and Nguyen, 2001, p. 426).

The goodness of fit was performed by the Kolmogorov-Smirnov method in the EasyFit program. For a statistic value (or critical value) of 0.19307, and a p-value (α) of $9.2144*10^{-6}$, we can conclude that there is no evidence to claim that the data do not have a Poisson distribution, with $\lambda = 3.24$ arrivals per day. It is worth mentioning that

such characteristics are not expected as most of the patients are scheduled (Bruin et al., 2010).

For LOS, descriptive statistics indicate that the mean was 6.50 days and the standard deviation was 8.13, so the Coefficient of Variation (Cv) is 1.25. This distribution has more variability than the previous category and is similar to most healthcare process according to Green and Nguyen (2001). The skewness of 3.48 shows a distribution skewed to the right.

The goodness of fit for service time distribution, applying the Kolmogorov-Smirnov method, ranked the Log-Logistic probability model as the first alternative. However, the Exponential distribution model was also not rejected as a possible explanation for the empirical distribution, as with a statistical value of 0.07695 and a p-value of 0.02006 there is no evidence to claim that the data do arise from an exponential probability model, with μ equal to 0.1538 hospitalizations per day.

As the empirical distributions for the arrivals and service rates can be explained by the Poisson and Exponential probability distributions, respectively, but also by another probability distribution, especially in the service rate, we applied queueing formulations to identifying parameters in the surgery and general ward unit.

4.3. Planning a ward unit based on queueing models

We can verify that the combined service-completion rate, or the total number of hospitalizations per day ($s*\mu$: 25*0.1538 = 3.85), is greater than $\lambda = 3.24$, conforming to the condition for the existence of a steady-state solution. The utilization rate (ρ = 84.26%) is higher than the ward occupation rate (75.0%). In fact, there are different approaches to computing these: the utilization rate formula considers the whole period in hospital of the patient as LOS, while the ward occupation rate considers the days in the months.

According to the model, the mean number of patients in the system (L) was 22.74, and the mean number of patients in the queue (Lq) was 1.68. In the empirical data, hospitalized patients ranged from 6 to 25, an average of 18.5. This difference may occur as the M/M/s assumes all patients are unscheduled. The mean wait time in the queue (Wq) was 0.52 days or 12h:28m, but the delay probability of a customer arriving in the queue [1-Wq(0)] was 31.46%, which is slightly high for this occupation rate.

If we observe the number of admissions to the ward per day of the week, there is a significant reduction on Sunday and Saturday, despite bed capacity availability (Table 4). This trend may not be regarded as caused by disease, but due to the convenience of the hospital and staff. Occupancy phenomenon has also been shown to be considerably lower on weekends and over holiday periods (Tierney and Conroy, 2014).

Table 4 around here

Using different numbers of servers (beds) but assuming the same service time and varying the arrival rate, in order to achieve the same utilization rate in this strip, we obtained the expected delays in days for each scenario (Figure 3). The graphic shows that the higher the utilization rates, the higher is the expected delay, which is logical. The results presented in Green and Nguyen (2001) are very similar, for instance: in the present study, the maximum expected delay is 4.91 days for 95% and 20 beds, and 1.64 days for 95% and 50 beds, while in Green and Nguyen study, where the unit is a surgery clinic, the maximum value is 4.6 and 1.6 days, respectively.

In the 30-bed scenario, an increase of just 0.23 in the arrival rate, or 6.9 patients per month, when the utilization rate is 85%, is sufficient to increase the expected delay by 14h:14m. A possible explanation is the mix of general and surgery patients in the same ward unit. As was described at the beginning of section 4, there are significant differences between the LOS of general and surgery patients, as well as the total care time demanded by such individuals. This means patients are more susceptible to delays, according to the occupation rates, due to the variability in the LOS associated with random arrivals.

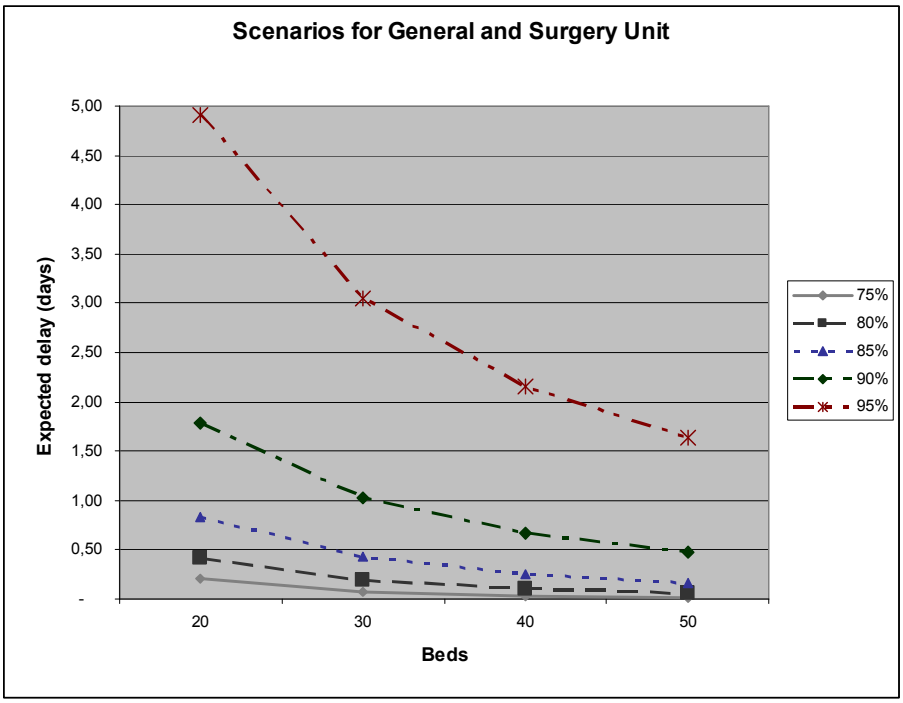

Figure 3: Expected delay for five different utilization rates.

If this unit had just surgery patients, considering the LOS for these patients (6.05 days), the unit could admit 84 more patients in the period, with 11h:37m of delay for the same utilization rate. If the unit had just general patients (LOS=9.09 days), it would have a decrease of 28 patients in a month, with 17h:27m of delay. According to literature, one of the focuses of a hospital on similar processes involves the need to separate different flows of patients, work, and goods, enabling each to move according to their own logic and pace (Rechel et al., 2010).

On three weekdays (Monday, Tuesday, and Wednesday), the admissions were between 19-20% (the peak was on Wednesday with 20%), while they were 6% and 7% on Sunday and Saturday, respectively.

Decisions about how many patients to admit from each clinic are usually related to physician and surgery room availability, but could also consider the level of service, or target expected delay. For this ward unit, a new policy toward equal distribution of admissions over the week would decrease expected delay and increase patient treatments. Lantz and Rosén (2016) pointed out the relevance of where and when

additional resources will be delivered, and how much should be delivered, to ensure a certain level of service.

Bearing in mind the correlations found in 4.1, it is expected that the LOS of a general patient will be longer and demand more total care time than a surgery clinical patient, but also that there will be a predictable LOS for clinical and surgery patients. This information might be useful in an admission schedule to decide on which date a type of patient should be admitted, to balance demand and resource availability.

After admitting a patient to the hospital, it is useful to improve patient flow as a feature in the bed management optimization program. An indication there is room for improvement comes from the M/G/s/s model, which provides the optimal number of servers in a system based on a smooth service time. This model identifies the minimum number of parallel servers necessary to guarantee that the overflow probability is no higher than a target threshold. In our case, we supposed the actual probability of delay (31.46%), which resulted in an optimal number of 17 beds. This is a strict model, however, and so we can apply another to evaluate the results.

The M/G/∞ model accepts any form of service-time distribution. Therefore, the expected system size is the expected number of customers in the service. In this case, the expected number of beds is 21, which may be related to the ward unit under study. This result is also valid for the M/M/∞ model.

5. Conclusions

In this paper, we discussed how Queueing Theory can contribute to hospital operations management. Implementing solutions for healthcare systems requires the integration and optimization of resources, and queueing models can help achieve this goal by demonstrating the effect of variability in patient mix for delays, as well as the optimal number of beds and the target service level.

The application of the M/M/s model found the system to be busier than it really was, as the number of patients computed in the system (L) was 22.82 and the hospitalized patients ranged from 12 to 23, an average of 18.5. Even without a high occupation rate (72.8%), the unit had a utilization rate of 84.4%, and 31.89% probability of delay for arriving customers.

Applying the M/G/s/s model for the optimal number of servers in the system, it was found that the number of beds would be 17 instead of 25, with 31.89% probability of delay.

The M/G/∞ model revealed that the required capacity was 21 beds. Regardless of which queueing model was applied, the findings make it clear that there is room for improvement in capacity management. Such decisions might involve finding the right mix of permanent versus temporary workers to balance supply and demand (Roth and Menor, 2003).

Balancing admissions and the types of patient over the week are possible solutions to optimizing bed and nurse utilization. An economically viable occupancy rate can be achieved by merging departments (bed pooling) or mixing patient flows (Liu and D'Aunno, 2012), augmenting bed capacity, both at an acceptable service level (Bruin et al., 2010).

Our findings demonstrate that the patient mix results in a highly sensitive delay rate, due to LOS variability, as we varied the utilization rate and the number of beds. Decisions about how many patients from each clinic to admit are usually related to

physician and surgery room availability, but could also consider the level of service, or a target expected delay.

The throughput at which services are performed does not always mean they are performed more quickly. Instead, a system can work with efficient coordination and control of the flow of all operations – including patients, personnel, and other resources.

Balancing the demand for beds can allow heavily used hospitals such as that in the present study to increase throughput or be used to decrease beds in a less heavily used units. One paradox of the healthcare system in Brazil is the hospital occupation rate, which is on average 46% across the entire country, but 62% in state capitals, while in OCDE countries this rate is 81% (Marinho et al., 2001). On the other hand, some hospitals have occupation rates of over 100% (Souza et al., 2010). In both situations, there is a need to adequately use hospital resources without affecting quality of service.

The Queueing Theory created a quantitative portrait of a day-by-day reality and is a powerful tool for hospital management.

Further analysis that was not performed in this study should be carried out, such as analyzing hourly arrival rates to optimize staff levels, considering the differences in demand between weekdays and weekends, and the improvement of support processes.

We investigated a number of opportunities for applying Queueing Theory in a public University hospital in Brazil. Though these findings are from a unique hospital, they can be expanded to other organizations, as the healthcare system is a significant presence in our society.

This field is a huge setting for operations management researchers and practitioners to implement theories. Complex solutions, the limited budgets of the organizations and people with limited options in terms of access and availability of services are the main characteristics of the system. Despite many advances in hospital management, there is room for further progress, notably those aimed at diminishing burden in many population clusters.

References

ANS (Agência Nacional de Saúde Suplementar). (2015), "Dados Gerais - Taxa de cobertura (%) por planos privados de saúde (Brasil - 2005-2015)", available at: http://www.ans.gov.br/perfil-do-setor/dados-gerais (accessed 26 December 2015).

Anupindi, R., Chopra, S., Deshmukh, S. Mieghem, J.A. and Zemel, E. (2012) *Managing Business Process Flows – principles of Operations Management*. Pearson, Prentice Hall, 3rd. Edition, Upper Saddle River, NJ.

Bamford, D. and Chatziaslan, E. (2009), "Healthcare capacity measurement", *International Journal of Productivity and Performance Management*, Vol. 58, No. 8, pp. 748-766.

Barasa, E., Molyneux, S., English, M. and Cleary, S. (2015) "Setting healthcare priorities in hospitals: a review of empirical studies", *Health Policy and Planning* Vol. 30, Nr. 3, pp. 386–396.

Ben-Gal, I., Wangenheim, M. and Shtub, A. (2010), "A new standardization model for physician staffing at hospitals", *International Journal of Productivity and Performance Management*, Vol. 59, No. 8, pp. 769-791.

Bhat, S., Gijo, E.V. and Jnanesh, N.A. (2016), "Productivity and performance improvement in the medical records department of a hospital", *International Journal of Productivity and Performance Management*, Vol. 65, No. 1, pp. 98–125.

Bruin, A. M., Rossum, A. C., Visser, M. C. and Koole, G. M. (2007), "Modelling the emergency cardiac in-patient flow: an application of Queueing theory", *Health Care Management Science*, Vol. 10, No. 2, pp. 125-137.

Bruin, A.M., Bekker, R., Zanten, L. And Koole, G.M. (2010), "Dimensioning hospital wards using the Erlang loss model", Ann Oper Res, Vol. 178, No. 1, pp. 23–43.

Budnick, F., McLeavey, D. and Mojena, R. (1988), *Principles of Operations Research for Management*. Irwin, Illinois.

Cochran, J. K. and Roche, K. T. (2009), "A multi-class Queueing network analysis methodology for improving hospital emergency department performance", *Computers & Operations Research*, Vol. 36, No. 5, pp. 1497-1512.

Derlet, R. W. and Richards, J. R. (2000), "Overcrowding in the nation's emergency departments: Complex causes and disturbing effects", *Annals of Emergency Medicine*, Vol. 35, No. 1, pp. 63-68.

Diwas, S. and Terwiesch, C. (2009), "Impact of Workload on Service Time and Patient Safety: An Econometric Analysis of Hospital Operations", *Management Science,* Vol. 55, No. 9, pp. 1486–1498.

Edwards, N. and Harrison, A. (1999), "The hospital of the future: planning hospitals with limited evidence: a research and policy problem", *BMJ*, Vol. 319, pp. 1262-1264.

Green, L. V. (2003), "How Many Hospitals Beds?" *Inquiry*, Vol. 39, No. 4, pp. 400-412.

Green, L. V. and Nguyen, V. (2001), "Strategies for Cutting Hospital Beds: The impact on patient service". *HSR: Health Services Research*, Vol. 36, No. 2, pp. 421-442.

Green, L. V., Kolesar, P. J. and Whitt, W. (2007), "Coping with Time-Varying Demand When Setting Staffing Requirements for service system", *Production and Operations Management*, Vol. 16, No. 1, pp. 13-39.

Green, L. V., Soares, J., Giglio, J. F. and Green, R. (2006), "Using Queueing Theory to Increase the Effectiveness of Emergency Department Provider Staffing", *Academy of Emergency Medicine*, Vol. 13, No. 1, pp. 61-68.

Gross, D., Shortle, J. F., Thompson, J. M. And Harris, C. M. (2008), *Fundamentals of Queueing Theory*. John Wiley & Sons, Inc., Hoboken, NJ.

Harper, P. and Shahani, A. (2002), "Modelling for the planning and management of bed capacities in hospitals", *Journal of the Operational Research Society*, Vol. 53, No. 1, pp. 11-18, 2002.

IBGE. (2006), "Estatísticas da saúde: assistência médico-sanitária". Rio de Janeiro, Centro de Documentação e Disseminação de Informações – CDDI.

IBGE. (2008), "Economia da saúde: uma perspectiva macroeconômica 2000 – 2005". Rio de Janeiro, Centro de Documentação e Disseminação de Informações – CDDI.

IBGE. (2013), "Conta-Satélite de Saúde Brasil". Rio de Janeiro, Centro de Documentação e Disseminação de Informações – CDDI.

Karakusevic, S. (2010), "Designing an Integrated Health Care System – What are the Key Features?" *Journal of Integrated Care,* Vol. 18, No. 4, pp. 36-42.

Lakshmi, C. and Iyer, S.A. (2013), "Application of Queueing theory in health care: A literature review", *Operations Research for Health Care*, Vol. 2, No. 1-2 , pp. 25–39.

Langabeer, J. R. II. (2008), *Healthcare Operations Management: A Quantitative Approach to Business and Logistics*. Jones and Bartlett Publishing, Boston, Massachusetts.

Lantz, B. and Rosén, P. (2016), "Measuring effective capacity in an emergency department", *Journal of Health Organization and Management*, Vol. 30, No. 1, pp. 73-84.

Liu, N. and D'Aunno, T. (2012), "The Productivity and Cost-Efficiency of Models for Involving Nurse Practitioners in Primary Care: A Perspective from Queueing Analysis", *HSR: Health Services Research*, Vol. 7, No. 2, pp. 594-613.

McManus, M. L., Long, M. C., Cooper, A., Mandell, J., Berwick, D. M., Pagano, M., & Litvak, E. (2003). Variability in surgical caseload and access to intensive care services. Anesthesiology: The Journal of the American Society of Anesthesiologists, 98(6), 1491-1496.

Marinho, A., Moreno, A. B. and Cavalini, L. T. (2001), *Avaliação descritiva da rede hospitalar do Sistema Único de Saúde (SUS) - Texto para Discussão*, No. 848, available at:
http://www.ipea.gov.br/portal/index.php?option=com_content&view=article&id=4098 (accessed 26 December 2015).

Martens, P. and Huynen, M. (2003), "A future without health? Health dimension in global scenario studies", *Bulletin of the World Health Organization*, Vol.81, No. 12, pp. 896-901.

Morton, A. and Cornwell, J. (2009), "What's the difference between a hospital and a bottling factory?", *BMJ*, Vol. 339, No. b2727, pp. 428-430.

Neter, J., Kutner, M., Nachtsheim, C. and Wasserman, W. (1990), *Applied Lineal Statistical Models*. Irwin, Chicago, IL.

Noon, C. E., Hankins, C. T., Cote, M. J., & Lieb, M. (2003). Understanding the impact of variation in the delivery of healthcare services/practitioner application. Journal of Healthcare Management, 48(2), 82.

Olsson, O., & Aronsson, H. (2015). Managing a variable acute patient flow–categorising the strategies. Supply Chain Management: An International Journal, 20(2), 113-127.

Rechel, B., Wright, S., Barlow, J. and McKeea, M. (2010), "Hospital capacity planning: from measuring stocks to modelling flows", *Bulletin of the World Health Organization*, Vol. 88, No. 8, pp. 632–636.

Reid, P., Compton, W.D., Grossman, J.H. and Fanjiang, G. (2005), *Building a Better Delivery System: A New Engineering/Health Care Partnership*, The national academies press, Washington, D.C.

Roemeling, O., Land, M., Ahaus, K. (2017). Does lean cure variability in health care?. International Journal of Operations & Production Management, 37(9), 1229-1245.

Roth, A. and Menor, L. (2003) "Insights into service operations management: a research agenda", *Production and Operations Management*, Vol. 12, No. 2, pp.145-164.

Souza, A. A., Lara, C. O., Neves, A. P. and Moreira, D. R. (2010), "Indicadores de desempenho para hospitais: análise a partir dos dados divulgados para o público em geral", *7º Congresso USP de Iniciação Científica em Contabilidade*, 2010.

Sztrik, J. (2016), *Basic Queueing Theory, Foundations of System Performance Modelling*. GlobeEdit, Saarbrücken, Germany.

Terwiesch, C., Diwas, K.C. and Kahn, J. M. (2011), "Working with capacity limitations: operations management in critical care", *Critical Care*, Vol. 15, No. 4, pp. 308-314.

Tierney, L. T. and Conroy, K. M. (2014), "Optimal occupancy in the ICU: A literature review", *Australian Critical Care*, Vol. 27 , No. 2, pp. 77-84.

Yuancheng Zhao, Qingjin Peng, Trevor Strome, Erin Weldon, Michael Zhang, Alecs Chochinov, (2015) "Bottleneck detection for improvement of Emergency Department efficiency", Business Process Management Journal, Vol. 21 Issue: 3, pp.564-585.

Zachariah, R. and others (2009), "Operational research in low-income countries: what, why, and how?", *Lancet Infection Disease*, Vol. 9, No. 11. pp. 711–717.

Table 1: Descriptive statistics

Category	Total care time (min.)		LOS (days)		Age	N
	Mean	Std. Dev.	Mean	Std. Dev.		
Male	778.79	1170.20	7.45	7.79	57.19	199
Female	918.15	1227.00	8.08	8.51	57.81	272
General	994.42	1401.21	9.09	9.27	58.99	273
Surgery	672.91	829.24	6.05	6.05	55.54	198
Male – Surgery	582.10	726.54	5.72	5.21	54.10	83
Female – Surgery	738.45	893.45	6.30	6.61	56.58	115
Male – General	919.52	1390.49	8.69	9.02	59.40	116
Female – General	1049.77	1410.96	9.38	9.48	58.70	157

Table 2: Correlation Coefficient (r_{xy}) – All patients | Surgery | General clinic

Correlation Coefficient r_{xy}	Total care time	LOS	Age
Total care time	1	0.80^{**} | 0.82^{**} | 0.79^{**}	0.12^{**} | 0.25^{**} | 0.06
LOS		1	0.11 | 0.34^{**} | -0.02
Age			1

** $p \leq 0.01$

Table 3: Correlation Coefficient (r_{xy}) – male | female

Correlation Coefficient r_{xy}	Total care time	LOS	Age
Total care time	1	0.79^{**} | 0.81^{**}	0.11 | 0.13^{*}
LOS		1	0.09 | 0.12^{*}
Age			1

** $p \leq 0.01$; * $p \leq 0.05$

Table 4: Admissions in surgery and clinic ward unit during the whole period.

Week-day	Admissions	Percent.
Sunday	28	6%
Monday	90	19%
Tuesday	88	19%
Wednesday	96	20%
Thursday	67	14%
Friday	68	14%
Saturday	34	7%