A Framework for Improving Access and Customer Service Times in Healthcare: Application and Analysis at the UCLA Medical Center

Catherine Duda, MPH, DrPH
Chief Administrative Officer
Anesthesiology & Emergency Medicine
Ronald Reagan UCLA Medical Center
Box 957115, 56-143 CHS
Los Angeles, CA 90095
Phone 310 206 0760 | Fax 310 825 2236
cduda@mednet.ucla.edu

Kumar Rajaram, PhD (Corresponding Author)
Professor of Operations and Technology Management
UCLA Anderson School of Management
Box 951481
Los Angeles, CA 90095
Phone 310 825 4154 | Fax 310 206 3337
krajaram@anderson.ucla.edu

Christiane Barz, PhD
Assistant Professor of Operations and Technology Management
UCLA Anderson School of Management
Box 951481
Los Angeles, CA 90095
Phone 310 825 0296 | Fax 310 206 3337
cbarz@anderson.ucla.edu

Thomas J. Rosenthal, MD
Chief Medical Officer and Professor
Ronald Reagan UCLA Medical Center
Box 957115, 56-143 CHS
Los Angeles, CA 90095
Phone 310 784 7700 | Fax 310 825 2236
Jrosenth@mednet.ucla.edu
A Framework for Improving Access and Customer Service Times in Healthcare: Application and Analysis at the UCLA Medical Center

Structured Abstract [172 words]

There has been an increasing emphasis on health care efficiency, costs and improving quality in health care settings such as hospitals or clinics. However, there has not been sufficient work on methods to improve access and customer service times in health care settings. The study develops a framework to improve access and customer service time for health care settings. In the framework, the operational concept of the bottleneck is synthesized with queuing theory to improve access and reduce customer service times without reduction in clinical quality. The framework is applied at the Ronald Reagan UCLA Medical Center to determine the drivers for access and customer service times and then provides guidelines on how to improve these drivers. Validation using simulation techniques shows significant potential to reduce customer service times and increase access at this institution. Finally, the study provides several practice implications that could be used to improve access and customer service times without reduction in clinical quality across a range of health care settings from large hospitals to small community clinics.

Keywords: hospital management, process analysis, access, customer service time
1. INTRODUCTION

There has been an increasing emphasis on health care efficiency, costs and improving quality at several health care settings such as hospitals and clinics. However, there has not been sufficient work on methods to improve access and customer service times defined as the sum of the processing and wait times the customer or patient experiences at the hospital. Understanding and improving access and customer service times is challenging as it requires a deep examination of an organization’s overall strategy, as well as the processes used to execute this strategy at several levels of the organization, including the corporate, business, and work process level. In addition, one needs to develop a comprehensive view of these processes, which involves understanding the customers, inputs, and process stages, and come up with the best tactics to utilize the process to effectively meet strategy.

Although there is vast literature available on the application of operations management in healthcare,\textsuperscript{1,2} none of the reviewed papers reported the use of operations models to understand the interdependence between hospital departments with the aim of improving access and customer service times. These are important aspects that if not managed effectively could lead to increasing numbers of refused admissions, longer waiting times for patients, decreased patient and staff satisfaction, wasted resources, and ultimately to decreased quality and increased mortality.\textsuperscript{3}

Therefore, a framework to improve access and reduce customer service times without reduction in clinical quality is presented. It is emphasized that in the framework, quality is at least maintained at current levels as potentially access and customer service times could be
improved if clinical quality standards are lowered. These situations are excluded. In this context, the framework will:

a. Map out the critical processes at each department in the hospital.

b. Identify the key sources of arrival and service variability at these processes.

c. Examine the patient flows through this process to determine the processing times at each step.

d. Calculate the capacity and utilization and identify potential bottlenecks at each department.

e. Identify how best to improve the performance of a department in terms of improving access and reducing customer service time.

f. Propose alternatives to increase access and reduce customer service times.

g. Validate the recommendations using simulation analysis.

The paper is organized as follows. In the next section, the framework for process analysis is described. Section 3 describes the application of the framework at the Ronald Reagan UCLA Medical Center. In Section 4, recommendations based upon the analysis are provided and simulation is used to validate these recommendations. In conclusion, Section 5 provides some key implications for practice.

2. **FRAMEWORK**

The method used to improve access and customer service times is based on the following steps, which is collectively referred to as the *framework for process analysis*.

**Step 1:** Draw a process flow diagram. This is typically a graphical and sequential representation of the inputs, stages, and outputs that make up the process.
Step 2: At each stage of the process, calculate the average processing times, define its range, and identify the sources of variability in processing times and arrivals that cause this range.

Step 3: Calculate the capacity, or output per unit time, of each stage using processing times. Define utilization as demand/capacity, and calculate the utilization at each stage.

Step 4: Identify the bottleneck, or the stage with the highest utilization. If the utilization of any stage is greater than 100 percent, then long run demand will not be met by this process.

Step 5: Consider changes to reduce variability of arrivals and service times in the system.

Step 6: Consider changes to shift the bottleneck to the most expensive stage (or the economic bottleneck) of the system.

Step 7: Consider changes to reduce the utilization of the bottleneck.

Step 8: Validate using simulation, evaluate changes, and implement the changes that lead to the highest improvement with the lowest cost.

While the steps outlined above are straightforward, there can be significant implementation challenges at several steps. For instance, when drawing a process flow diagram in Step 1, it can be difficult to decide which tasks to include in the analysis (i.e. the detail), how to combine tasks into stages (i.e. level of aggregation) and determining the best sequence of stages. In general, the detail, aggregation, and sequence should match the objective of the analysis, its intended use and also depends upon the specific analyst. However for successful implementation, there must be consensus between the analyst and user in terms of the detail, sequence, and degree of aggregation of the steps before the start of the other steps. In Step 2, data on processing times at each stage is often not available and requires the execution of a time-motion study. Further, one needs to develop a good understanding of the sources of
variability. In Step 3, the capacity of each stage should be calculated in isolation without accounting for constraints from the other stages. Such constraints will be imposed in Step 4. If many scenarios are given for processing times at a given stage, the worst case scenario should be used. This is done because if a stage is not the bottleneck under the worst case situation, it does not merit further managerial attention at this point. In Step 4, calculating demand to determine utilization can be challenging because when there are several types of patients, each type typically does not use each stage in a process equally. In Steps 5 through 7, care should be taken to identify the least expensive solutions that would have the highest impact. In Step 8, recommendations to improve system performance should be evaluated using discrete-event simulation. This allows an evaluation of the impact of recommended changes on patient flows and to investigate the complex relationships among different operational variables. Finally, note that this approach may not include all key parameters that influence departmental processes. Thus, one may need to make subjective assessments based on institutional knowledge, and may change over time. In this case, this framework should be reevaluated under different assumptions at different periods of time.

The developed framework for process analysis can be used to identify the bottleneck and increase capacity or access across the process. Indeed, activities similar to those described in the first four steps of this framework have been applied to increase process capacity in several contexts in the manufacturing and service industry. However, the contribution of this work lies in structuring and expanding these activities to include improvement in customer service times. This is achieved by employing concepts from queuing theory which have been increasingly used to achieve operational improvements in health care. The G/G/1 queuing
model is first used to identify the key drivers of customer service times. In the G/G/1 model, the first G represents a general distribution of patient or customer inter-arrival times, the second G represents a general distribution of processing times at the bottleneck, and the 1 represents the fact that process performance is primarily driven by the critical bottleneck resource. If this resource is composed of multiple servers in parallel, the effective capacity across these servers is used by assuming that these servers perform identical services and they are uniform in ability and quality. This is particularly relevant in this framework as no assumptions are made about the arrival process of customers and processing times at the bottleneck, and the time the customer spends at the hospital is mainly influenced by the bottleneck. In this model, average customer service times are a function of capacity, utilization, and variability, and which can be estimated by the following equation:

\[
\text{Average Customer Service Time} = \frac{1}{\mu} \left[ \frac{1}{1 - \rho} \right] \left[ \frac{C_a^2 + C_s^2}{2} \right]
\]

[Equation 1]

Here, \( \mu \) = service rate or capacity of the bottleneck stage in this process, \( \rho \) = number of arrivals per unit of time / \( \mu \), \( C_a \) = coefficient of variation in inter-arrival times, and \( C_s \) = coefficient of variation in processing times.

It is important to note that Equation (1) bases its estimate of average customer service times under the standard assumptions for the G/G/1 model where there is a first come first serve queue discipline and there is no customer balking. Observe from Equation (1) that customer service time is driven by three effects: the capacity effect, the utilization effect, and the variability effect represented by the first, second, and third terms respectively. The capacity effect reaffirms the intuition that the lower the capacity at the bottleneck, the longer customer service time is. The utilization effect emphasizes the fact that customer service times increase
dramatically if the bottleneck is overworked or over-utilized (i.e., utilization, $\rho$, gets closer to 1). If the utilization of any stage exceeds 100 percent, then the process is incapable of meeting even long run demand. The variability effect refers to the deviation between actual and expected inter-arrival and processing times. A common measure of variability is the Coefficient of Variation (CV), which represents the standard deviation of a parameter as a percentage of its mean. The variability effect implies that as the level of variability in the system increases due to the arrival of patients (measured as the CV of inter-arrivals and denoted by $C_a$) or due to how service procedures are conducted (measured as the CV in processing times and denoted by $C_s$), customer service times increase. Lack of capacity and high utilization are amplified by the variability effect because the term denoting the variability effect ($[C_a^2 + C_s^2]/2$), the capacity effect ($1/\mu$), and the utilization ($1/[1 - \rho]$) terms are all multiplicative in the above equation.

Equation (1) provides a conceptual framework to understand, and then attack, the drivers of customer service times. In particular, this suggests that customer service times are primarily driven by the capacity and utilization of the bottleneck, and by the degree of variability in arrivals and processing times at various stages in the process. Thus, once the sources of variability and the bottleneck have been identified in Steps 1 through 4, this equation provides the insight that customer service times can be reduced by increasing capacity at the bottleneck, reducing utilization at the bottleneck, and reducing variability in arrivals and service. This is exactly Steps 5 through 7 of the framework for process analysis.

The main contribution of this work is in synthesizing the concepts of bottlenecks with queuing theory by developing a framework for process analysis that can increase access and reduce customer service times. These aspects are very important in healthcare management.
and to the author’s best knowledge, this is the first framework to explicitly and jointly address these aspects. The next section describes the specific application of the framework for process analysis at the Ronald Reagan UCLA Medical Center (RRUCLA).

3. APPLICATION

The RRUCLA is a 456 bed acute care hospital located in Los Angeles, California. The institution’s mission is to deliver leading-edge patient care, research, and education (http://www.uclahealth.org/homepage_med.cfm). This is achieved by providing world-class medical treatment using cutting-edge technology in a patient-focused environment. RRUCLA has been consistently ranked one of the top five hospitals in the United States and has been rated the best hospital in the western United States for 20 consecutive years by US News and World Report.8

As a major, tertiary, academic medical center, the demand for health care services at the RRUCLA is high. This demand requires a high degree of process effectiveness to ensure that RRUCLA is able to see the largest number of patients with the highest possible quality and responsiveness. Responsiveness is measured by the average customer service times. This is the average of the sum of processing and wait times the patient or customer experiences across the hospital departments, with longer customer service times implying lower responsiveness. RRUCLA has found that capacity has been increasingly insufficient to meet growing patient demands. In addition, there are periodic fluctuations in patient volume that has been overwhelming the hospital’s capacity to respond. For the period from March 2009 to March 2010, the median inpatient occupancy was 98 percent in sharp contrast to existing guidelines of 85 percent.9 Hospitals operating at full capacity often “board” patients who need to be
admitted until inpatient beds become available, potentially causing safety and other problems. Average wait times for the period from July 2009 to February 2010, measured from the time of admission to placement in an inpatient bed was over eight hours. This was significantly larger than their targeted times of two hours. Such wait times can lead to dissatisfaction with medical care and a possible deterioration of patient’s health. Specific patient waits depend on processes within and across departments. For instance during a hospital stay, a patient may experience individual waits for beds, procedures, diagnostics, education, transportation, rehabilitation, and discharge-related processes. Since customer service times are the sum of processing and wait times, increasing wait times directly increases overall customer service times and reduces responsiveness.

There are several departments at the RRUCLA and patients can flow through many of the departments (Figure 1). The management team at each department is responsible for coordinating its processes and planning its staff to ensure smooth patient flow through the department. After extensive consultation with RRUCLA executive management, the following departments were chosen for detailed further study: [1] Emergency Department (ED); [2] Admissions; [3] Patient Transport; [4] Beds; [5] Operative Services; [6] Laboratory; [7] Radiology; and the [8] Pharmacy. These departments were selected based on the volume of patient flow as the management wanted to ensure that at least 50% of the hospital’s patients flowed through each department. The processes at these departments and their interactions with other departments are analyzed. It is important to take a holistic view to this analysis as a patient must go through several departments, and thus, several processes in order to obtain health care services. An individual department becomes a bottleneck and increases overall
customer service time when the ratio of demand to available service capacity is relatively high. As a result, hospital management faces the challenging decision to allocate limited resources effectively amongst competing departments.

**INSERT FIGURE 1 ABOUT HERE**

In this section, Steps 1 through 4 of the framework is executed on a department-by-department basis for the entire hospital. In order to understand overall hospital flow, data was collected in several phases. First, interviews were conducted with the top management at the hospital to understand the strategic objectives of each department. Interviews were also conducted with each department head to understand and describe current patient flow. The product of this phase was a series of departmental process flow diagrams. These were then submitted to the individual department management to ensure that the detail, level of aggregation and sequence of stages in these process flow diagrams were consistent with their expectations. It is essential to gain consensus on the process flow diagram if the recommendations based on its analysis had to be implemented at the appropriate departments. Data was then collected at the various stages of these process flow diagrams through software systems, interviews, or time motion studies as needed at each department. This data was also used to identify the key sources of variability at each department.

The next step involved estimation of demand (average and range) per day and of processing times (average and range) for each stage of a department’s process flow diagram. The ranges in demand and processing times provided an indication of the sources of arrival and service variability, respectively. To calculate average capacity at each stage of the process, data
was collected on the number of servers, hours of operation, and average processing times at each stage. This data was then used in the following equation:

\[ C_i = N_i T_i \left( \frac{60}{P_i} \right) \]  

[Equation 2]

Here: \( C_i \) = average capacity per day at stage \( i \), \( N_i \) = number of servers at stage \( i \), \( T_i \) = hours per day stage \( i \) is open, and \( P_i \) = average processing time in minutes per patient at stage \( i \).

The utilization of each stage is then calculated by dividing its average demand per day by its average capacity. The stage with the highest utilization is the bottleneck of the process. This analysis was performed across all departments. More details of this analysis including specific department process flow diagrams and the calculations for each stage in the department process flow diagram can be found in Duda.\textsuperscript{11} Based on this analysis, Table 1 summarizes the average demand, average processing times, average capacity, and utilization at the bottleneck at each of the analyzed departments. In addition, the range of demand and range of processing time for the bottlenecks in each department are detailed in Table 2.

**INSERT TABLES 1 AND 2 ABOUT HERE**

Using this analysis in the next section, recommendations are formulated for each department and the impact of the most important recommendations are validated using a simulation model developed in the process simulator software program, ProModel (Promodel Corporation, Orem, UT).

4. **RECOMENDATIONS**

In this section, Steps 5 through 7 of the framework are performed to provide recommendations to reduce variability, improve capacity, and reduce utilization at each department. As shown in Table 1, the RRUCLA has a number of opportunities for process
improvement specifically within the areas of operative services (98% utilization), emergency department (97% utilization), pharmacy (93% utilization), and hospital beds (91% utilization). As seen in Equation (1), such high levels of utilization lead to long customer service times. This can be further increased by large levels of variability in either arrival or processing times.

The bottlenecks in these departments are not necessarily caused by the inability of a single department to achieve maximum effectiveness. Instead, they are more likely caused by departments working in a semi-autonomous way to maximize departmental-specific patient flows without consideration for how such actions may affect the performance of other upstream or downstream processes. Identifying and managing system-level constraints (or interdependencies) is a better approach to achieving process effectiveness rather than improving each department in isolation. Access and customer service times can be improved by managing the bottlenecks of the departments with high utilization (by increasing capacity, or reducing utilization, or both) and by minimizing variability across all departments.

A series of recommendations to improve access and reduce customer service times are developed below. These recommendations were guided by the following principles. First, note from Equation (1) that the impact of a shortage of bottleneck capacity and overutilization of the bottleneck can be exacerbated by increased levels of variability. Therefore, it is critical that recommendations to change processes to reduce variability in arrivals and service at both the operational bottleneck and the department be aggressively pursued before improving the bottleneck and reducing its utilization. In particular, increasing the capacity of the bottleneck could increase access and reduce utilization. However, a large volume of patient flows associated with increased access along with current procedures can increase process variability
to the extent that the benefits of increased capacity and utilization are negated by increasing variability. Therefore, overall customer service times could actually increase.\textsuperscript{13} Second, the operational bottleneck should correspond to the economic bottleneck, or most expensive resource in the process. If the most expensive resource is not the bottleneck, then by definition it has slack capacity or idle time and one would like to minimize this at the most expensive resource to be cost effective. The economic bottleneck at each department was determined by utilizing an activity based cost accounting system developed by the RRUCLA. In this system, the fixed cost of equipment at each stage and variable cost of supplies and staff at each stage were used to calculate the unit cost as cost per unit patient per activity. Here, the dimension for activity was either set to a transaction or time depending on the nature of the stage. A stage with the highest unit cost represents the economic bottleneck at the appropriate department. Third, since small changes in utilization at higher levels of utilization can dramatically increase customer service times, recommendations to reduce utilization levels at the economic bottleneck should be identified. Finally, note that for cost effectiveness, utilization levels across the entire process can be managed by identifying stages that are particularly underutilized with respect to the economic bottleneck and aligning their utilizations with those of the economic bottleneck. However, it is critical that system variability first be reduced and the other two recommendations are first executed before one attempts this step as this may increase customer service times and negate the benefits of the previous steps.

These recommendations were developed by using the results of the process analysis to identify the bottleneck stage at each department. Recollect that Equation (1) helps identify the levers of customer service times as capacity of bottleneck, utilization of the bottleneck, and
variability in arrivals and service. This equation in turn helps focus and justify the recommendations based on which lever of customer service time is primarily affected by any particular recommendation. Table 3 summarizes the specific recommendations organized by department and the impacted lever of customer service times. Furthermore, within each department/lever category, the recommendations are listed in decreasing order of priority as needed. Deciding which specific recommendations to include and how to prioritize them in Table 3 was done in close consultation with the appropriate department heads and their team leaders. This aspect was crucial as this embeds the institutional knowledge of the workforce in understanding which idea would work in their organizational context. Such blending of expert judgment with process analysis is crucial for the successful implementation of this framework. In addition, as indicated in Table 3, several of the recommendations were consistent with prior research.

**INSERT TABLE 3 ABOUT HERE**

It is important to understand that once improvements are made at the bottleneck stage at a department, the hospital-wide bottleneck could potentially shift to the next department. For example, observe from Table 1 that any improvements in operative services that reduce utilization below 97% would make the ED the next bottleneck. To decide whether to continue to implement these recommendations, it is important to examine the process economics and business strategy of the organization. If for example the OR is not the most expensive resource or the economic bottleneck of the hospital, then the various recommendations are implemented until the economic bottleneck is reached. In case the OR is the economic bottleneck or the economic bottleneck is reached implementing the appropriate
recommendations in Table 3, the business strategy of the organization is revisited. If the strategy requires further improvements in access and customer service times, the target utilization is set based on these goals. The economic bottleneck and other subsequent bottlenecks are then improved to meet the target. In case such improvements are not prescribed by the strategy or the target utilization is met, the focus would be on managing by the economic bottleneck to ensure that it works effectively and that all other stages meet its requirements. In addition, it is important to make sure that variability in arrivals and service across all departments are reduced to the extent possible.

Finally, note that the recommendations provided in this Table 3 are specific to each department. However, they can also be used to develop some general insight into how to tackle the drivers of access (i.e., capacity of the bottleneck) and customer service times (i.e., variability in arrivals and service, capacity, and utilization of the bottleneck). These insights provide useful guidance to practitioners who apply the framework in other settings and are summarized in points 4 through 7 in Section 5.

4.1 Validation Using Simulation

Discrete Event Simulation (DES) is used to conduct Step 8 of the framework and validate the recommendations. DES has been increasingly used to analyze healthcare systems.\textsuperscript{3,14,15} The purpose of this simulation is twofold. First, note that Equation (1) is an approximation for calculating average customer service times in a multi-stage, dynamic setting. Therefore, it is important to validate the insights provided by this equation. Second, the purpose of this simulation is to identify in which departments process improvements would lead to the highest impact from a system wide or hospital perspective. To achieve these objectives, a simulation
model is developed to virtually analyze the impact of proposed system modifications from Section 4 on hospital access and customer service times and demonstrate the effects of: [1] decreasing variability of service times at the bottleneck; [2] increasing capacity at the departmental bottlenecks; and [3] reducing utilization by decreasing bottleneck processing time. Here, it is assumed that the changes in variability, capacity, and utilization can be achieved by following the detailed recommendations at the appropriate departments as summarized in Table 3.

The structure of the simulation model of the RRUCLA is shown in Figure 1. In this model, each department is represented by the bottleneck identified in Section 3. Since all practical aspects of the hospital cannot be simulated, it important to incorporate institutional knowledge at the highest possible level to decide what aspects to include in the development of the simulation model. This was done by ensuring that this model was thoroughly vetted by the department heads and team leaders. The model was constructed using the Promodel simulation software. Details on model formulation and validation are provided in Duda.11 This section describes the scenario analyses to determine how changes to service variability, capacity, and processing time affect access and customer service times. The range of the simulation parameters for the scenario analysis was chosen to cover a wide range of processing times in other hospitals gathered from surveys and appropriate publications.16,17 The goal was to understand the magnitude of change that could be expected if these scenarios were implemented in the actual hospital. Access is defined by hospital throughput, measured as the number of patients discharged from the hospital per unit time. Customer service times were defined by the Enterprise Length of Stay (ELOS),18 calculated as the sum of the various
Department Lengths of Stay (DLOS), including Admissions, Transport, OR, Lab, Radiology, and Pharmacy. The DLOS is measured as the sum of processing times and wait times at the appropriate department. Note that reductions in DLOS will improve the overall ELOS.

In a real situation, the implementation of these recommendations would commence at the largest bottlenecks or the departments with the highest utilization. Therefore, the simulation follows the same sequence and provides results by department organized in decreasing order of utilization as shown in Table 1. More details on the simulation results can be found in Duda.11

**Operative Services**

The operative services department has the highest utilization (98%) of all the hospital departments. First, the standard deviation of processing times at the Operating Room (OR) was altered in increments of 15 minutes from 0 to 120 minutes. The results show that as the standard deviation in OR processing time or service variability is effectively increased, the throughput performance of both the OR and the hospital are adversely affected. As variability in OR processing times increases to 120 minutes, hospital throughput decreases by 25 patients per week while ELOS increases by 5.4 days. This is because the DLOS of the other departments increase. Thus, this simulation provides quantified evidence that bottlenecks in one department impact upstream and downstream processes. Next, the capacity of the OR is increased in a stepwise fashion up to 200% from current levels. The results show that this would have the greatest impact on reducing utilization of the OR bottleneck. A 25% increase in OR capacity decreases OR bottleneck utilization from 99.6% to 82.6%, and shifts the hospital-wide bottleneck to the pharmacy. This capacity increase decreases ELOS by 2.5 days (26%) from
baseline, and increases hospital throughput by 10 patients per week, or 2.5%. Finally, OR bottleneck processing time was decreased in 5% decrements up to 50% of current levels. Results show that decreasing the bottleneck processing time at the OR from 220 to 187 minutes (15%) decreases ELOS by 2.5 days (26%). Furthermore, throughput would increase by 9 patients per week (2.3%), thereby increasing access to surgeries.

In light of this analysis, it is recommend that initiatives to reduce service time variability, increase OR capacity, and reduce processing times as described in Table 3 should be pursued in this department. The same analysis was conducted for the ED, which has an utilization of 97% with similar results. Therefore, the details are omitted and it is recommend that the initiatives to improve ED service variability, capacity, and process times as summarized by Table 3 should be implemented in this department.

**Pharmacy**

The pharmacy department has the third largest utilization (93%) at the hospital. First, scenarios were created to alter the standard deviation of processing times at the pharmacy in increments of 60 minutes from 260 to 740 minutes. The results show that an increase in the standard deviation of processing times or effectively service variability from 500 minutes to 740 minutes (48%) increases ELOS by 4 days and reduces hospital throughput by 12 patients per week (3.1%). Therefore, improvement efforts should begin by reducing variability of pharmacy processing time as suggested in Table 3. Next, the capacity of the pharmacy bottleneck was increased in individual increments up to 50% of current levels. However, such increases in pharmacy capacity do not significantly impact ELOS or hospital throughput. Finally, pharmacy processing time was reduced in 5% decrements down to 50% of current levels. A 20% decrease
in pharmacy processing time, from 625 minutes to 500 minutes, decreases ELOS by 0.8 days (7.9%). However, such reductions have no effect on hospital throughput. Therefore, this should not be the current focus of resources and managerial attention.

In light of these results, improvement efforts should be focused on reducing variability of pharmacy processing time as outlined in Table 3, as this would have the greatest impact on reducing access and customer service times across the entire hospital.

**Beds**

Total bed capacity was increased in 5% increments up to 50% of current levels. Results show that achieving only a 15% percent increase in capacity would reduce ELOS by 11 hours. However, such improvements have only minor effects on hospital throughput in this model. Nevertheless, such decreases in ELOS would improve patient satisfaction and reduce ED diversion rates. The traditional approach to increasing bed capacity within a department included adding more resources (beds and staff) and expediting discharges. However, the effectiveness of the solution can be increased by adopting a system wide perspective and pooling beds between the ED, ancillary departments, and inpatient areas, and by improving the processes of delivering care. Using such a system wide approach, the RRUCLA could dramatically improve bed availability without new capital expenditures. Such pooling of beds also reduces utilization at high utilization departments by distributing patient loads to underutilized beds in low utilization departments.

In addition to pooling of beds, utilization of beds can be reduced by minimizing preventable readmissions. In the year 2011, 3,772 or 16% of the adult patients were readmitted within 90 days after a previous discharge. Of these, emergency admissions were 2.7 times more
likely than elective admissions to be readmitted. Although not all readmissions are avoidable, some could be prevented by improving the quality of care. The added operational benefit is to improve access and customer service times for other patients. This benefit is not apparent and can be overlooked.

**Laboratory**

The standard deviation of processing time at the laboratory was altered in increments of 60 minutes from 168 to 648 minutes. The results show that an increase in the standard deviation of processing times or effectively service variability from 408 minutes to 648 minutes (59%) increases ELOS by 0.63 days, but has no effect on hospital throughput. Capacity of the laboratory bottleneck was then increased in increments from 0% to 50% of current levels. However, these improvements had minimal effect on hospital ELOS or throughput. Finally, lab processing time was reduced in 5% decrements down to 50% of current levels. However, these improvements also have a minimal effect on hospital ELOS and throughput.

The analysis indicates that reducing variability in service times, increasing capacity, and reducing processing times at this department does not significantly improve overall access and customer service times at the hospital. Thus, this department should not be the focus of managerial attention at this time. This is also consistent from the results of Table 1 that shows the laboratory has significantly lower utilization (71%) than the departments considered so far. This analysis also suggests that similar results can be expected from the remaining departments in Table 1 as they have lower utilization than the Laboratory. This is verified in the simulation and the details are omitted.
In summary, the simulation model can be used to evaluate the hospital-wide impact of changing service variability, capacity, and utilization at the bottleneck stage of each department. This model validates the intuition that increasing capacity can improve access, while reducing variability, increasing capacity, and reducing utilization can reduce customer service times as measured by the ELOS. The simulation is useful in understanding the complex relationship between these variables in a dynamic, multi-dependent setting and also in assessing the magnitude of the change. This in turn provides guidance on which department and drivers should be tackled to improve hospital-wide performance in access and customer service times. In particular, it provides the important insight that maximum improvement at the RRUCLA can be achieved by focusing on improving operative services, the emergency department, and the pharmacy versus any of the other departments. This insight is crucial for establishing management proprieties and would not have been validated without the simulation or Step 8 of the framework for process analysis.

5. PRACTICE IMPLICATIONS

There are several implications for practice that can be drawn from this study. The implications are listed below to encourage similar process improvement activities at other health care settings.

1. Process analysis can be used to identify the actual or operational bottleneck in a systematic and logical manner.

2. It is important to shift the operational bottleneck to the economic bottleneck or the costliest resource.
3. Customer service time can be managed by reducing variability in arrivals and service, increasing the capacity of the bottleneck and reducing the utilization of the bottleneck. It is important to first reduce variability in arrivals and service and then follow up by improving the capacity and utilization of the bottleneck.

4. Variability in arrivals can be controlled by following a queuing discipline, developing an appointment system, improving staff planning by cross-training workers to deal with peak periods, and reducing work batching at any particular department as they could create variability in patient flows at other departments.

5. Variability in service times can be controlled by identifying the best practices at each stage, training workers on the best practices, providing sufficient and timely access to information, using adequate automation, developing effective scheduling systems and finally reducing steps by either combining or eliminating steps.

6. Capacity of bottlenecks can be increased with improved scheduling, by employing more staff during peak periods, using better technology to reduce processing times and minimize down times, ensuring staff at bottleneck stages are used effectively, adding capacity by leasing equipment, and offloading demand to other stages.

7. Utilization of the bottleneck can be reduced by pooling resources, developing good information flows, offloading demand to other stages, and by effective scheduling so that parallel resources have the lowest possible utilization.

8. Simulation can be used to validate the recommendations of process analysis and to determine where the highest impact, from a hospital-wide perspective, can be achieved.
This provides management with priorities on which department to focus process improvement efforts.

The framework is particularly important given that timely access has been identified as one of the key elements of healthcare quality, and decreasing delays has become a focus of many healthcare institutions. However, there could be challenges implementing this framework at both the tactical and organizational level. At the tactical level, an important limitation would be the ability of the organization to collect accurate and timely data needed to conduct process analysis. While time motion studies were conducted as needed at the RRUCLA to collect this data, this approach may be costly, cumbersome, and disruptive to conduct on an ongoing basis. A potential solution for this limitation is in developing information systems using mobile or RFID technology to gather real-time data. This could be embedded in an automated decision support system. Developing such systems could be a very fruitful future area for research and business development. At the organizational level, a key limitation could be the appropriate alignment of incentives. This is particularly challenging as the costs and benefits of these improvements have different stakeholders such as hospitals, providers/employees, and patients with dissimilar and sometimes conflicting interests. For instance, an investment that increases access and reduces customer service times may have different financial consequences for the hospital in comparison to its patients. However, incentives to promote reductions in process variations such as those advocated in Step 5 of the framework will soon be directly encouraged by the federal government in the form of value-based purchasing (VBP), which scores providers based upon quality performance and patient satisfaction. Hospitals with the
highest scores will receive bonuses from a pool of dollars formed by withholding a portion of Medicare reimbursements across all providers. Recent research states that the best way for a hospital to improve its VBP score will be to reduce process variances at the departmental level. The framework helps to achieve such reductions in process variances.

In conclusion, the presented framework provides an effective way to increase access and improve customer service times across a range of health care settings from large hospitals to small community clinics.
References


### Table 1. RRUCLA Departmental Bottleneck Summary

<table>
<thead>
<tr>
<th>Stage</th>
<th>Department</th>
<th>Bottleneck Stage</th>
<th>Avg Demand (patients per day)</th>
<th>Avg Process Time of Bottleneck (min)</th>
<th>Avg Capacity (patients per day)</th>
<th>Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Emergency Rooms for Single Patient</td>
<td>125</td>
<td>300</td>
<td>130</td>
<td>97%</td>
<td></td>
</tr>
<tr>
<td>2a</td>
<td>Admissions – ED ED Admissions</td>
<td>34</td>
<td>15</td>
<td>192</td>
<td>18%</td>
<td></td>
</tr>
<tr>
<td>2b</td>
<td>Admissions – Elective Patient Interview</td>
<td>22</td>
<td>15</td>
<td>160</td>
<td>14%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Transport Request Assignment</td>
<td>700</td>
<td>10</td>
<td>1152</td>
<td>61%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Medicine Beds 7E Medicine</td>
<td>76</td>
<td>84</td>
<td>90%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Surgery Beds 8N Liver Transplant</td>
<td>137</td>
<td>150</td>
<td>91%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ICU Beds 7CCU Coronary Care Unit</td>
<td>98</td>
<td>108</td>
<td>91%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Operative Services OR Theaters</td>
<td>47</td>
<td>220</td>
<td>48</td>
<td>98%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Laboratory Specimen Preparation</td>
<td>8135</td>
<td>11109</td>
<td>73%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Radiology Attending Radiologist</td>
<td>388</td>
<td>576</td>
<td>67%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Pharmacy Review &amp; Verification</td>
<td>7000</td>
<td>7560</td>
<td>93%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 2. Range of Bottleneck Demand and Processing Time

<table>
<thead>
<tr>
<th>Stage</th>
<th>Department</th>
<th>Bottleneck Stage</th>
<th>Range of Demand (# of patients per day)</th>
<th>Range of Processing Time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Minimum</td>
<td>Maximum</td>
</tr>
<tr>
<td>1</td>
<td>Emergency Rooms for Single Patient</td>
<td>72</td>
<td>147</td>
<td>8</td>
</tr>
<tr>
<td>2a</td>
<td>Admissions – ED ED Admissions</td>
<td>4</td>
<td>54</td>
<td>10</td>
</tr>
<tr>
<td>2b</td>
<td>Admissions – Elective Patient Interview</td>
<td>1</td>
<td>53</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>Transport Request Assignment</td>
<td>168</td>
<td>936</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>Medicine Beds 7E Medicine</td>
<td>23</td>
<td>25</td>
<td>1 (day)</td>
</tr>
<tr>
<td>4</td>
<td>Surgery Beds 8N Liver Transplant</td>
<td>25</td>
<td>26</td>
<td>1 (day)</td>
</tr>
<tr>
<td>5</td>
<td>Operative Services OR Theaters</td>
<td>1</td>
<td>67</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>Laboratory Specimen Preparation</td>
<td>7000</td>
<td>9000</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>Radiology Attending Radiologist</td>
<td>245</td>
<td>492</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>Pharmacy Review &amp; Verification</td>
<td>6000</td>
<td>8000</td>
<td>2</td>
</tr>
</tbody>
</table>
Table 3. Summary of Recommendations

<table>
<thead>
<tr>
<th>Department</th>
<th>Variability</th>
<th>Capacity</th>
<th>Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating Services</td>
<td>Decrease variability in surgical case length by establishing best practices. 21</td>
<td>Manage sources of waste such as late arrivals by developing an appointment system and by cross training worker to deal with peak load demand. 21</td>
<td>Perform surgeries during off-peak hours. 21&lt;br&gt;Best scheduling of Operating Rooms. 23</td>
</tr>
<tr>
<td></td>
<td>Reduce variability in elective surgical caseload to avoid peaks and valleys in workload. 22</td>
<td></td>
<td>Measure/manage accuracy of estimated case times. 21&lt;br&gt;Effective scheduling to reduce utilization. 21&lt;br&gt;Shift outpatient surgical volume to an ambulatory surgery setting. 24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>信仰 among various hospital stakeholders to allow for accurate measurement and benchmarking. 24</td>
</tr>
</tbody>
</table>
Figure 1. RRUCLA Hospital Departments and Patient Flows

Outpatient
- Primary Care Clinics
- Urgent Care
- Emergency

Diagnostic & Treatment
- Operative Services
- Laboratories
- Radiology

Inpatient Beds
- Medical
- Surgical
- ICU

Post Hospital Care

Discharge