The Health Care Manager
Volume 32, Number 3, pp. 212–226

A Framework for Improving Access and Customer Service Times in Health Care

Application and Analysis at the UCLA Medical Center

Catherine Duda, DrPH; Kumar Rajaram, PhD; Christiane Barz, Dr rer pol; J. Thomas Rosenthal, MD

There has been an increasing emphasis on health care efficiency and costs and on improving quality in health care settings such as hospitals or clinics. However, there has not been sufficient work on methods of improving access and customer service times in health care settings. The study develops a framework for improving 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 for reducing customer service times and increasing 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. Key words: *access, customer service time, hospital management, process analysis*

THERE HAS BEEN an increasing emphasis on health care efficiency and costs and on improving quality at health care settings such as hospitals and clinics. However, there has not been sufficient work on methods of improving 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 are challenging as they require 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 health care,[1,2] none of the reviewed articles 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.[3]

Author Affiliations: Ronald Reagan UCLA Medical Center (Drs Duda and Rosenthal) and UCLA Anderson School of Management (Drs Rajaram and Barz), Los Angeles, California.

The authors have no conflicts of interest.

Correspondence: Kumar Rajaram, PhD, UCLA Anderson School of Management, Box 951481Los Angeles, CA 90095(krajaram@anderson.ucla.edu).

DOI: 10.1097/HCM.0b013e31829d7313

Therefore, a framework for improving access and reducing 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:

- map out the critical processes at each department in the hospital,
- identify the key sources of arrival and service variability at these processes,
- examine the patient flows through this process to determine the processing times at each step,
- calculate the capacity and utilization and identify potential bottlenecks at each department,
- identify how best to improve the performance of a department in terms of improving access and reducing customer service time,
- propose alternatives to increase access and reduce customer service times, and
- validate the recommendations using simulation analysis.

The article is organized as follows. In the next section, the framework for process analysis is described. The third section describes the application of the framework at the Ronald Reagan UCLA Medical Center (RRUCLA). In the fourth section, recommendations based on the analysis are provided, and simulation is used to validate these recommendations. The concluding section provides some key implications for practice.

FRAMEWORK

The method used to improve access and customer service times is based on the following steps, 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%, 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.

Although the steps outlined above are straightforward, there can be significant implementation challenges at several steps. For instance, when drawing a process flow diagram, it can be difficult to decide which tasks to include in the analysis (ie, the detail), how to combine tasks into stages (ie, level of aggregation), and determining the best sequence of stages. In general, the detail, aggregation, and sequence should match the objective of the analysis and 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 are often not available and require the execution of a time-and-motion study. Furthermore, 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 greatest impact. In step 8, recommendations for improving 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 these may change over time. In this case, this framework should be reevaluated under different assumptions at different periods.

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 4 steps of this framework have been applied to increase process capacity in several contexts in the manufacturing and service industry.[4] However, the contribution of this work lies in structuring and expanding these activities to include improvement in customer service times. This is achieved by using concepts from queuing theory, which have been increasingly used to achieve operational improvements in health care.[5–7] The G/G/1 queuing model[4] 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 interarrival times, the second G represents a general distribution of processing times at the bottleneck, and 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 that 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 can be estimated using the following equation:

Average customer service time
$=[1/\mu][1/(1-\rho)][(C_a^2+C_s^2)/2]$(Equation 1)

Here, μ is service rate or capacity of the bottleneck stage in this process; ρ, number of arrivals per unit of time / μ; C_a, coefficient of variation in interarrival 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-served queue discipline and there is no customer balking. Observe from Equation 1 that customer service time is driven by 3 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 the customer service time. The utilization effect emphasizes the fact that customer service times increase dramatically if the bottleneck is overworked or overutilized (ie, utilization, ρ, gets closer to 1). If the utilization of any stage exceeds 100%, then the process is incapable of meeting even long-run demand. The variability effect refers to the deviation between actual and expected interarrival and processing times. A common measure of variability is the coefficient of variation (CV), which represents the Standard Deviation (SD) of a parameter as a percentage of its mean. The variability effect implies that as the level of variability in the system increases because of the arrival of patients (measured as the CV of interarrivals and denoted by C_a) or because of 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 for understanding, and then attacking, 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 bottle neck 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 health care management, and to the authors' 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 RRUCLA.

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. Ronald Reagan UCLA Medical Center has been consistently within the top 5 hospitals in the United States and has been rated the best hospital in the western United States for 20 consecutive years by the *US News & 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. The 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%, in sharp contrast to existing guidelines of 85%.[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 more than 8 hours. This was significantly larger than their targeted times of 2 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. Because 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). 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 further detailed 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 hospitals 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.[10] As a result, hospital management faces the challenging decision to allocate limited resources effectively among competing departments.

In this section, steps 1 through 4 of the framework are executed on a department-by-department basis for the entire hospital. In order to understand overall hospital flow, data were 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 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 were then collected at the various stages of these process flow diagrams through software systems, interviews, or time-and-motion studies as needed at each department. These

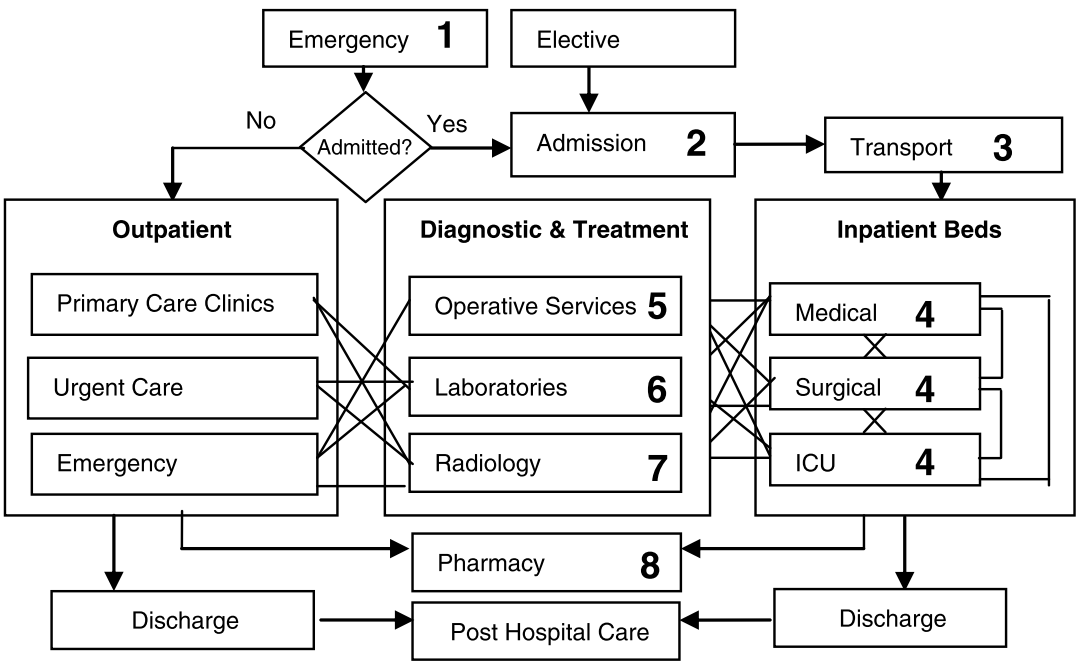

Figure. Ronald Reagan UCLA Medical Center hospital departments and patient flows.

data were 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 were collected on the number of servers, hours of operation, and average processing times at each stage. These data were then used in the following equation:

$$C_i = N_i T_i (60/P_i) \text{(Equation 2)}$$

Here, C_i is 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.[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.

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

Table 1. RRUCLA Departmental Bottleneck Summary

Stage	Department	Bottleneck Stage	Average Demand, No. of Patients/d	Average Process Time of Bottleneck, min	Average Capacity, No. of Patients/d	Utilization, %
1	Emergency	Rooms for single patient	125	300	130	97
2a	Admissions—emergency department (ED)	ED admissions	34	15	192	18
2b	Admissions—elective	Patient interview	22	15	160	14
3	Transport	Request assignment	700	10	1152	61
4	Medicine beds	7E medicine	76	9936	84	90
	Surgery beds	8N liver transplant	137		150	91
	Intensive care unit beds	7CCU coronary care unit	98		108	91
5	Operative services	Operating room theaters	47	220	48	98
6	Laboratory	Specimen preparation	8135	7	11109	73
7	Radiology	Attending radiologist	388	15	576	67
8	Pharmacy	Review and verification	7000	5	7560	93

Abbreviation: CCU, coronary care unit.

RECOMMENDATIONS

In this section, steps 5 through 7 of the framework are performed to provide recommendations to reduce variability, improve

Table 2. Range of Bottleneck Demand and Processing Time

Stage	Department	Bottleneck Stage	Range of Demand (No. of Patients/d)		Range of Processing Time, min	
			Minimum	Maximum	Minimum	Maximum
1	Emergency	Rooms for single patient	72	147	8	2500
2a	Admissions—emergency department (ED)	ED admissions	4	54	10	20
2b	Admissions—elective	Patient interview	1	53	7	23
3	Transport	Request assignment	168	936	5	15
4	Medicine beds	7E Medicine	23	25	1 (d)	158 (d)
4	Surgery beds	8N Liver transplant	25	26	1 (d)	258 (d)
5	Operative services	Operating room theaters	1	67	17	1050
6	Laboratory	Specimen preparation	7000	9000	4	10
7	Radiology	Attending radiologist	245	492	1	30
8	Pharmacy	Review and verification	6000	8000	2	7

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), ED (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 semiautonomous 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) are a better approach to achieving process effectiveness rather than improving each department in isolation.[12] 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 and are shown in Table 3. 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

Table 3. Summary of Recommendations

Department	Variability		Capacity	Utilization
	Service	Arrivals		
Operating services	Decrease variability in surgical case length by establishing best practices[13] Reduce variability in elective surgical case load to avoid peaks and alleys in workload[15]	Manage sources of waste such as late arrivals by developing an appointment system and by cross training worker todeal with peak load demand[13]	Perform surgeries during off-peak hours[14] Better scheduling of operating rooms[14]	Measure/manage accuracy of estimated case times Effective scheduling to reduce utilization[13] Shift outpatient surgical volume to an ambulatory surgery setting[16]
Emergency department (ED)	Establish standard time frames for specialty consultation eg, dermatology, cardiology, rheumatology) to reduce length of stay in the ED[17] Decrease variability in elective surgery scheduling to improve ED output and congestion[15]	Initiate a queuing discipline such as a rapid admission protocols to reduce ED boarding times and time to admission[18]	Increase availability of patient care appointments in the primary care setting to avoid inappropriate use of the ED[19]	Decrease use of ED for nonurgent health problems by directing minor, appropriate cases to an urgent care center[19]
Pharmacy	Increase pharmacy automation to improve the level and quality of process control[20] Reduce order processing steps by eliminating or reducing steps	Decrease batching of pharmacy orders to improve inpatient and discharge process flows[21] Examine division of labor between pharmacists and technicians to ensure that highly trained pharmacists are used to the fullest extent of their degree[22]	Reduce interruptions of pharmacist by hiring appropriate support staff at peak periods	Reduce unnecessary demand at the main pharmacy by informing patients about pharmacy sub-branches within the hospital

(continues)

Table 3. Summary of Recommendations, Continued

	Variability			
Department	**Service**	**Arrivals**	**Capacity**	**Utilization**
Beds	Provide accurate and timely access to information to consulting doctors by developing the required information systems Implement training programs tobetter manage inpatient resources to reduce length of stay[24]	Reduce variability of bed demand for elective surgical cases/scheduled admissions[15]	Increase availability of discharge to communityresources, such as skilled nursing facilities, home health, hospice, and palliative care, to hasten discharges[23]	Pool beds across departments to the extent possible Decrease inappropriate utilization by having admission decisions reviewed by senior physician/bed director[25] Reduce unplanned, inappropriate readmissions to both medicine and surgical units[25]
Laboratory	Ensure clear turnaround time definitions, including acceptable and unacceptable performance based on clinical evidence, among various hospital stakeholders to allow for accurate measurement and benchmarking[26]	Deliver specimens via pneumatic tube system versus courier to decrease batching[27]	Better technology to reduce processing times and minimize down times Lease additional processing equipment, particularly centrifuges to reduce turnaround time for results[28]	Reduce duplicate laboratory orders by better information flows using effective data transmission systems Reduce unnecessary demand at the main laboratory by informing patients about pharmacy sub-branches within the hospital

negated by increasing variability. Therefore, overall customer service times could actually increase.[29] 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 set to either 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, because 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 2 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.

It is important to understand that once improvements are made at the bottleneck stage in 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 less than 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 operating room (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 their requirements. In addition, it is important to make sure that

variability in arrivals and service across all departments is reduced to the extent possible.

Finally, note that the recommendations provided in 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 (ie, capacity of the bottleneck) and customer service times (ie, 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 the Practice Implications section.

Validation Using Simulation

Discrete-event simulation is used to conduct step 8 of the framework and validate the recommendations. Discrete-event simulation has been increasingly used to analyze health care systems.[3,30,31] The purpose of this simulation is 2-fold. First, note that Equation 1 is an approximation for calculating average customer service times in a multistage, 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 the Figure. In this model, each department is represented by the bottleneck identified in the Application section. Because all practical aspects of the hospital cannot be simulated, it is 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.[32,33] 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),[34] calculated as the sum of the various department lengths of stay (DLOS), including admissions, transport, OR, laboratory, 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 SD of processing times at the OR was altered in increments of 15 minutes from 0 to 120 minutes. The results show that as the SD in OR processing time or service variability is effectively increased, the throughput performance of both the OR and the

hospital is adversely affected. As variability in OR processing times increases to 120 minutes, hospital throughput decreases by 25 patients per week, whereas ELOS increases by 5.4 days. This is because the DLOS of the other departments increases. 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 recommended that initiatives to reduce service time variability, increase OR capacity, and reduce processing times as described in Table 3 be pursued in this department. The same analysis was conducted for the ED, which has a 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 be implemented in this department.

Pharmacy

The pharmacy department has the third largest utilization (93%) in the hospital. First, scenarios were created to alter the SD of processing times at the pharmacy in increments of 60 minutes from 260 to 740 minutes. The results show that an increase in the SD of processing times or effective 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% 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, 3772 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 SD 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 SD of processing times or effective 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, laboratory 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 in this department do 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, whereas 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, multidependent setting and also in assessing the magnitude of the change. This in turn provides guidance on which department and drivers should be tackled to improves 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 ED, 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.

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 using 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 off-loading demand to other stages.
7. Utilization of the bottleneck can be reduced by pooling resources, developing good information flows, off-loading 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 health care quality,[35] and decreasing delays has become a focus of many health care 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. Although time-motion studies were conducted as needed at the RRUCLA to collect these 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 (radiofrequency identification) 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, which scores providers based on 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 value-based purchasing score will be to reduce process variances at the departmental level.[36] 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

1. Cayirli T, Veral E. Outpatient scheduling in healthcare: a review of literature. *Prod Oper Manage*. 2003; 12(4):519-549.
2. Yih Y, ed. *Handbook of Healthcare Delivery Systems*. Boca Raton, FL: CRC Press; 2010.
3. Hall RW. *Patient Flow: Reducing Delay in Healthcare Delivery*. New York, NY: Springer Science; 2006.
4. Anupindi R, Chopra S, Van Mieghem JA, Zemel E. *Managing Business Process Flows: Principles of*

Operations Management. 2nd ed. Upper Saddle River, NJ: Prentice Hall; 2005.
5. McManus ML, Long MC, Cooper A, et al. Variability in surgical caseload and access to intensive care services. *Anesthesiology*. 2003;98(6):1491-1496.
6. McManus ML, Long MC, Cooper A, Litvak E. Queuing theory accurately models the need for critical care resources. *Anesthesiology*. 2004;100(5):1271-1276.
7. Litvak E, Buerhaus PI, Davidoff F, Long MC, McManus ML, Berwick DM. Managing unnecessary variability in patient demand to reduce nursing stress and improve patient safety. *Jt Comm J Qual Patient Saf*. 2005;31(6):330-338.
8. US News Best Hospitals 2010-2011. US News & World Report Web site. 2010. http://www.usnews.com/besthospitals. Accessed October 6, 2010.
9. Hemesath M, Pope GC. Linking Medicare capital payments to hospital occupancy rates. *Health Aff*. 1989;8(3):104-116.
10. Fomundam S, Herrmann J. *A Survey of Queuing Theory Applications in Healthcare [The Institute for Systems Research Technical Report 2007-24]*. College Park, MD: University Of Maryland; 2007.
11. Duda C. *Applying Process Analysis and Discrete Event Simulation to Increase Access and Customer Service Times at the Ronal Reagan UCLA Medical Center* [doctoral dissertation]. Los Angeles, CA: University of California Los Angeles; 2011.
12. Goldratt EM, Cox J. *The Goal: A Process of Ongoing Improvement*. Great Barrington, MA: The North River Press; 1992.
13. Tyler DC, Pasquariello CA, Chen CH. Determining optimum operating room utilization. *Anesth Analg*. 2003;96(4):1114-1121.
14. Chessare J. Creating non-block scheduling to increase volume, revenue, and surgeon satisfaction. Urgent Matters Web site. Published November 2004. http://urgentmatters.org/e-newsletter/318807/318808/318811. Accessed October 6, 2010.
15. Litvak E, Prenney B, Fuda KK, Long MC, Levtzion-Korach O, McGlinchey P. *Improving Patient Flow and Throughput in California Hospitals Operating Room Services*. Boston, MA: Boston University Health Policy Institute; 2006.
16. Kutscher B. Outpatient care takes the inside track. *Mod Healthc*. 2012;42(32):24-26.
17. McHugh M, Van Dyke K, McClelland M, Moss D. *Improving Patient Flow and Reducing Emergency Department Crowding: A Guide for Hospitals* [publication no. 11(12)-0094]. Rockville, MD: Agency for Healthcare Research and Quality; 2011.
18. Amarasingham R, Swanson TS, Treichler DB, Amarasingham SN, Reed WG. A rapid admission protocol to reduce emergency department boarding times. *BMJ Qual Saf*. 2010;19(3):200-204.
19. *A Matter Of Urgency: Reducing Emergency Department Overuse* [research brief]. Cambridge, MA: New England Healthcare Institute; 2010.
20. Tan WS, Chau SL, Yong KW, Wu TS. Impact of pharmacy automation on patient waiting time: an application of computer simulation. *Ann Acad Med Singapore*. 2009;38(6):501-507.
21. Richards S. *Improving Cost Efficiencies Related to the Discharge Policy: An Analysis of the Retail Pharmacy at Moffitt Cancer Center and Research Institute* [paper 73]. Tampa, FL: University of South Florida; 2011.
22. Schommer J. Pharmacist workload and time management. *Drug Top*. 2001;145(4):45-54.
23. Brewster LR, Felland LE. Emergency department diversions: hospital and community strategies alleviate the crisis. *Issue Brief (Center for Studying Health System Change)*. 2004;78:1-4.
24. Rachoin JS, Skaf J, Cerceo E, et al. The impact of hospitalists on length of stay and costs: systematic review and meta-analysis. *Am J Manag Care*. 2012; 18(1):e23-e30.
25. Payne SM. Identifying and managing inappropriate hospital utilization: a policy synthesis. *Health Serv Res*. 1987;22(5):709-769.
26. Hawkins RC. Laboratory turnaround time. *Clin Biochem Rev*. 2007;28(4):179-194.
27. Fernandes CMB, Worster A, Eva K, Hill S, McCallum C. Pneumatic tube delivery system for blood samples reduces turnaround times without affecting sample quality. *J Emerg Nurs*. 2006;32(2):139-143.
28. Lippi G, Salvagno GL, Montagnana M, Guidi GC. Preparation of a quality sample: effect of centrifugation time on stat clinical chemistry testing. *Lab Med*. 2007;38(3):172-176.
29. Kc D, Terwiesch C. Impact of workload on service time and patient safety: an econometric analysis of hospital operations. *Manage Sci*. 2009;55(9):1486-1498.
30. Jun JB, Jacobson SH, Swisher JR. Application of discrete-event simulation in health care clinics: a survey. *J Oper Res Soc*. 1999;50(2):109-123.
31. Jacobson SH, Hall SN, Swisher JR. Discrete-event simulation of health care systems. *Int Ser Oper Res Manage Sci*. 2006;91:211-252.
32. Khare RK, Powell ES, Reinhardt G, Lucenti M. Adding more beds to the emergency department or reducing admitted patient boarding times: which has a more significant influence on emergency department congestion. *Ann Emerg Med*. 2009;53(5):575-585.
33. Ozcan YA, Tanfani E, Testi A. A simulation-based modeling framework to deal with clinical pathways. In: *Proceedings of the 2011 Winter Simulation Conference*. 2011;1190-1201.
34. Jackson R. *Average Length of Stay: It's Time for a New Metric*. Alpharetta, GA: Jackson Healthcare; 2008.
35. Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: The National Academies Press; 2001.
36. Andrews H, Wessels G. Healthcare reformers are focusing on value: are you? *Healthc Financ Manage Assoc*. 2009;63(80):44-52.