The Flexible Patient Flow Simulation Framework

Jonathan E. Helm, Shervin AhmadBeygi, Mark P. Van Oyen
Department of Industrial and Operations Engineering
University of Michigan, Ann Arbor, Michigan 48109-2117, USA

Abstract

Hospital congestion and the many associated healthcare delivery cost and quality issues are caused or exacerbated by decisions made without understanding system-level patient flow. Without patient flow models, decisions that affect the entire system are made considering only the effect on a small subsection of the system. This paper presents a flexible simulation framework as a powerful tool for modeling patient flow. Using the Patient Flow Simulation Framework on inpatient admissions, we show that leveraging controls to stabilize workflow through the hospital can mitigate congestion related problems by dramatically reducing variability in hospital occupancy. This yields higher quality delivery at a lower cost.

Keywords

Simulation Framework, Patient Flow, Hospital Admission Control, Stabilizing Hospital Occupancy

1. Introduction

In the United States process knowledge in health care lags significantly behind the manufacturing sector. One consequence of this is that hospital care services are subject to significant, unnecessary and detrimental fluctuations in patient census and associated workload. Indeed every care process in the hospital is negatively impacted by variability in admissions that is propagated to downstream hospital services. This variability in patient census has been linked specifically to congestion and chaos in the Emergency Department (ED), excessive radiology backlogs, strains on nurse and ancillary staff, and overcrowding in the Intensive Care Unit (ICU) and Post Acute Care Unit (PACU). System wide congestion results in compromised quality of care, emergency patient diversions for lack of beds, excessive patient Length of Stay (LOS), and significant understaffing and overstaffing costs [7,9,10]**.**

Though hospitals have little control over variability in emergency arrivals, they can control elective patient admissions. In most hospitals elective surgeries are scheduled without considering the downstream resources needed to care for the patient post surgery. As a result hospitals experience overcrowding of critical resources like the ICU and PACU and unnecessary and unpredictable swings in hospital occupancy. To understand the consequences of decisions, such as elective surgery scheduling decisions, on the hospital as an entire system, one must understand the path a patient follows through the system from the decision point onward. We call this patient flow.

To give the reader a sense of the magnitude of the opportunity lost because of variability in hospital workload, we present a hypothetical analysis of a typical community hospital. This is not intended to be a precise savings estimate, but merely an illustration of the significance of occupancy variability. According to data from the United States Department of Health and Human Services [1], the "typical" hospital in the US has 160 staffed beds with an average occupancy of 57.7 %, or 92 occupied beds. Assume that this typical hospital consistently allocates resources to support peak occupancy levels during the week. This is not an unreasonable assumption because the high occupancy variability in most hospitals makes predicting peak workload difficult.

Now suppose the variability in occupancy were reduced enough so that one could assert with confidence that the census will rarely exceed 120% of the average, or 111 occupied beds. This means high variability caused the hospital to allocate resources for 49 additional beds to meet the same average daily demand. Reducing the variability in hospital occupancy essentially frees up 49 bed-units of hospital resources. The resources freed by this increased efficiency can be used at the hospital's discretion for such improvements as increased throughput, reduced cost or stored surge capacity for emergency preparedness.

To quantify the value of these 49 bed-units of resource, we consider the per diem cost of a hospital bed. The average per diem for a hospital bed is $1,817 according to data from the Department of Health and Human Services [1]. As a conservative estimate, assume that the fixed cost of a bed accounts for 40% of the per diem. Assume also that the reduced variability only affects weekdays since most hospitals have reduced resources on weekends. The annual value estimate becomes: 49 beds * 260 days * 1817 dollar per diem * 60% = $13,889,148 annually.

In this paper we consider how patient flow modeling can help a hospital reduce this expensive variability in occupancy. We begin first with high level conceptual patient flow modeling. Then we present the flexible Patient Flow Simulation Framework (PFSF) that contains the components necessary to transform a conceptual patient flow model into a stochastic simulation model. In the final section we focus on a specific application of the simulation framework to patient flow through a network of hospital bed units. Using the simulation framework we evaluate implementable policies to reduce detrimental variability in hospital occupancy.

2. Patient Flow Modeling

Patient flow modeling is extremely important for understanding system-level consequences of hospital decisions and policies. Currently many important hospital decisions are made independently, without considering the workload strain and costs those decisions place on downstream hospital resources. Take the decision to schedule a patient for an elective surgery, for example. After surgery, a patient may spend a day in the intensive care unit and then be transferred to acute care for one more day and finally transferred to a surgical bed to recover for the next three days before being discharged. This places a workload on several critical hospital resources including the ICU, ACU, hospital bed, nursing staff and any testing services that may be required.

Without an understanding of patient flow and the workload a surgery places on the rest of the hospital resources, scheduling occurs only according to the surgeon's time preference, leaving the hospital to accommodate the elective inflow. With each surgeon scheduling independently, the number of patients scheduled can vary widely from day to day. This scheduling practice can make elective admissions even more variable and unpredictable than emergency admissions. The high variability in elective admissions causes congestion that blocks ED from moving to an inpatient bed. As the ED becomes overcrowded with patients waiting for a bed, the quality of patient care is negatively impacted leading to increased LOS and/or worsening patient disposition [3]. In extreme cases, ED congestion can lead to an increased mortality rate [10]. ED congestion can be reduced by a patient flow management system that uses the hospital occupancy level in decision making [7].

Figure 1 illustrates a patient flow model for a hospital with 3 units (Medicine, Surgery, and ICU). Emergent patients arrive to the units with rates λ_1, λ_2, and λ_3 and are transferred to another unit or discharged with probability p_{ij} after completing the stay in unit i. Scheduled patients enter the system through a controlled mechanism, as do patients with urgent care needs.

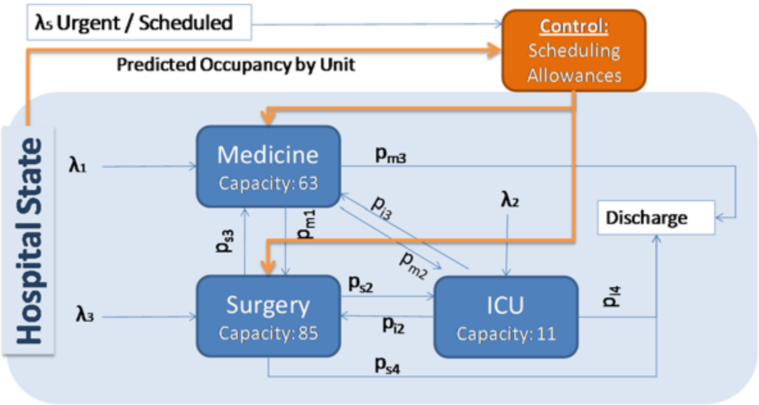

Figure 1 – Conceptual Patient Flow Model with 3 Units

A significant amount of research uses patient flow modeling to link hospital admission policies with census. [2] developed early stochastic models to link admissions decisions with hospital occupancy. In a more recent study, [6] used a multistage stochastic approach to establish that variability in daily hospital occupancy in combination with high occupancy levels can increase the risk of hospital overflows. These papers establish the importance of employing census information to make admission decisions and of reducing occupancy variability.

There has also been extensive research linking admission policies to census using simulation models. [5] developed an inpatient admissions scheduling and control system to maximize average occupancy subject to constraints on the number of cancellations and emergent diversions. According to [8], an effective patient flow model can enable high patient throughput, low patient wait times, short LOS, and low clinic overtime.

The models in the literature demonstrate the impact patient flow modeling can have on healthcare delivery system functioning. Each model above, however, serves to analyze a specific patient flow problem. In Section 3 we present a simulation framework that has the flexibility to model multiple classes of patient flow problems.

3. The Patient Flow Simulation Framework

As demonstrated in Section 2, patient flow models are powerful tools for improving healthcare delivery systems. There are many tools for modeling and analyzing patient flow systems; however most of them suffer from scalability in that every flow system must be rebuilt from scratch. Fortunately, most patient flow systems share a set of common building blocks. To take advantage of these common elements, we designed a flexible patient flow modeling framework that can be used to efficiently build simulations of a vast majority of patient flow systems.

The Patient Flow Simulation Framework (PFSF) can model elective and emergent patients flowing through a network of operating rooms and recovery units, patients and tests being routed through diagnostic imaging and ancillary services, and patients flowing through outpatient clinics to name a few applications. The PFSF can also model larger enterprise level systems, including care pathways of patients who flow through a hospital or the entire care pathways of patients from the time they first seek medical care until their treatment is completed. It can even be applied to networks of hospitals to take advantage of shared resources. In the Section 4 we demonstrate how the PFSF can be used to model and analyze patient flow through a hospital from the point of entry to the time of discharge, but first we detail the design and core functionality of the PFSF.

The PFSF was developed in C++ for several reasons. First, the object-oriented nature of C++ makes the modularization of patient flow structures very convenient. Secondly, C++ interfaces well with optimization toolkits, enabling the framework to develop both descriptive and prescriptive models. Finally, C++ is a flexible and widely familiar program language, and, unlike applications developed with off-the-shelf simulation software, products designed with C++ do not require a software license and can run on any computer. This reduces barriers to patient flow research acceptance and use in hospitals by making patient flow simulation more accessible.

Bearing this in mind, the PFSF is structured as a set of C++ library files with a class structure for each major patient flow component. The components currently represented include a patient record, several patient arrival components, a unit input component, a patient output and routing component, and a component for collecting simulation statistics.

During the simulation new patient structures are created for each new arrival to the hospital. These patients are represented by the patient record that contains data regarding the patient's condition. This record contains a link to the input component that determines the length and resources required for the initial segment of treatment. The record then travels with the patient and whenever the patient finishes a segment of their treatment, the patient record contains a link to the output control associated with the patient's condition to determine the next treatment segment. Essentially, the patient record encodes the probabilistic care pathway of the patient's condition which is transformed into an actual sample path as the patient transitions through different segments of care during the simulation.

The arrival components allow for stochastic and deterministic arrivals, and the queueing of patients awaiting admission. The arrival component object contains a switch that indicates whether the arrival is random or deterministic. In the random case, the component contains functions to generate a random number of each patient type. If the arrival control object is set to deterministic, the distribution contains an array from which it generates successive non-random numbers of patients at chosen simulation time points. Some of the policies investigated in section 4 rely on a class of on-demand patient type (e.g. expedited patients that require admission before a given deadline) that does not seek immediate admission to the hospital, but instead queues up and waits for a signal to seek admission. To implement these policies, the arrival control component also contains queue control functionality.

With each hospital resource unit (e.g. the Surgery Unit in Figure 2), there are associated input controls and output controls for each patient type that can use the resource unit. When a patient needs service from a resource unit, the patient’s record links to the appropriate input control object for their condition to generate the patient’s service time for that resource. Similarly, when the patient finishes a treatment segment, the output control object associated with a particular patient’s condition generates the set of resources required for the next segment of the patient’s treatment.

Associated with each input control object is an alternate action object. This data structure encodes what alternative resource combinations could be used to satisfy a segment of patient treatment in the event that the patient’s preferred resource is full. For example, if a hospital fills its surgical beds, a surgical patient may be placed on a medical unit. Should one encounter a healthcare delivery system with structures beyond those in the existing framework, PFSF can be easily expanded to incorporate new structures.

Figure 2 demonstrates how the patient flow model in Figure 1 can be translated into a simulation using the PFSF. First, the PFSF reads in a data file containing hospital historical data and summary statistics and then generates the simulation automatically from the building blocks as shown in Figure 2. The applications of the PFSF include demonstrating the effect of decisions and policies on a larger system, testing new policies, comparing polices, and optimizing an objective under a specified set of constraints. In the next section we use the PFSF to generate a simulation model of a hospital similar to the one shown in Figure 1 and Figure 2 to demonstrate how hospital admission policies can stabilize hospital occupancy by reducing the variability in elective patient scheduling to improve quality and reduce the cost of care delivery.

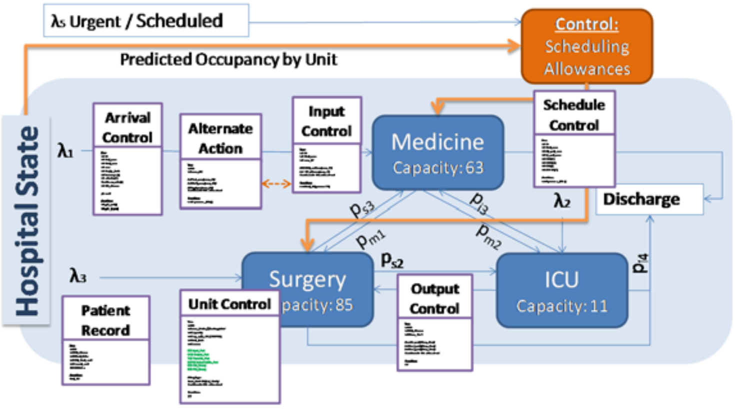

Figure 2 – PFSF Components Overlaid on a Conceptual Patient Flow Model

4. Applications of the Framework

To illustrate the power of the PFSF we return to the motivating problem from the introduction to investigate the effect of variability in occupancy on the functioning of the hospital as a whole. Several papers have investigated hospital admission policies that reduce census variability and improve key hospital metrics. Here we use the PFSF to demonstrate how combining two complementary hospital admission control policies from [4] and [7] can greatly improve the operational functioning of a hospital. [4] focuses only on elective admissions while [7] complements this by proving the optimality of using a double threshold policy based on hospital occupancy levels to regulate the daily admissions. The two thresholds divide the hospital occupancy level into three regions: full, neutral, and empty. Based on which region the occupancy falls into, the hospital will cancel surgeries, do nothing, or call in expedited patients from a queue. The policy we test uses both a level loaded schedule, and a daily cancel/call-in decision point.

To make the example broadly applicable, we consider two identical typical community hospitals, Hospital A and Hospital B, of 160 beds and three main units as in Figure 1 above. We use the national average length of stay of 4.4 days [1]. Emergent arrivals comprise around 47% of inpatient admissions. The data used to generate the two hospitals is based loosely on historical data from an actual hospital. The only difference between Hospital A and B is that Hospital A uses a typical front loaded scheduling practice with no daily control thresholds, and Hospital B uses level loaded scheduling and call-in and cancellation thresholds from the two papers mentioned above. To determine the optimal threshold levels for Hospital B, we used a neighborhood search as described in [7]. This demonstrates the capabilities of the PFSF to perform prescriptive optimization in addition to comparison and descriptive modeling. The simulation was run for 1,000 days and replicated 100 times to ensure that the standard error on the estimators was sufficiently small. The results below are striking when one considers that the admission scheduling and control policies used in Hospital B are quite simple and require very little IT support to implement.

In Figure 3, Hospital B shows a reduction of more than 80% in the number of cancellations and diversions, and the 27% reduction in the variability in hospital occupancy means the daily workload is more predictable. This ensures that staffing levels in all hospital subsystems can be more effectively matched to workload / patient demand, thus reducing both overstaffing and understaffing. More predictable workload and staffing can result in cost savings, improved quality of patient care and less chaos for healthcare professionals.

Concerning patient safety, quality of care and ease of care delivery, consider the number of patients placed off-unit. When a patient is placed in a location other than their preferred unit because that unit is full, caring for that patient becomes more difficult. That is because an off-unit patient is not receiving care from staff that specializes in their particular condition. You may for example have a medical focused staff caring for a surgical patient who has been placed off-unit when the surgery ward was full. Since off-unit patients are physically distanced from similar patients, it makes it problematic for doctors to visit those patients during rounds and wastes the doctors valuable time. As seen in Figure 3, Hospital B has significantly fewer patients off-unit, which translates into better, more efficient and easier care delivery.

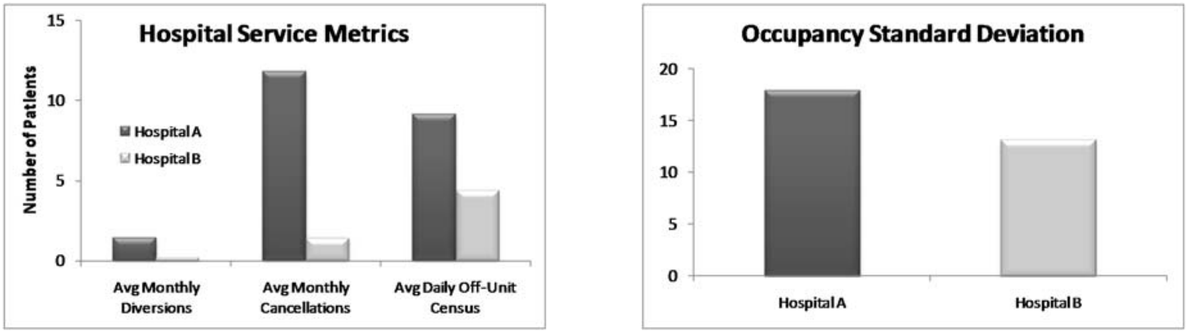

Figure 3 – Comparison of Service & Quality Metrics between Hospital A and Hospital B

Figure 4 compares the average daily census by day of week for Hospitals A and B. It is important to note that Hospital A spends the entire middle of the work week at or near peak occupancy, creating an impression of an overloaded hospital. The reality is that the demand for service in Hospital A is the same as in Hospital B. In fact Hospital B has higher throughput due to fewer cancellations and diversions. By using the patient flow model in admission decision making, Hospital B rarely reaches peak occupancy levels. This leveling of hospital occupancy in Hospital B could be improved further by optimizing the scheduled admissions since a weekday level loaded schedule does not fully level the occupancy when there are few elective admissions on the weekend. As might be expected, the cancellations and diversions peak in the middle of the week for Hospital A in concert with the midweek congestion and significantly outweigh the cancellations and diversions faced by Hospital B. This leveling of the inpatient census facilitated by call-in and cancellation thresholds and a level elective admission schedule drives the improved operational metrics seen in Figures 3 and 4.

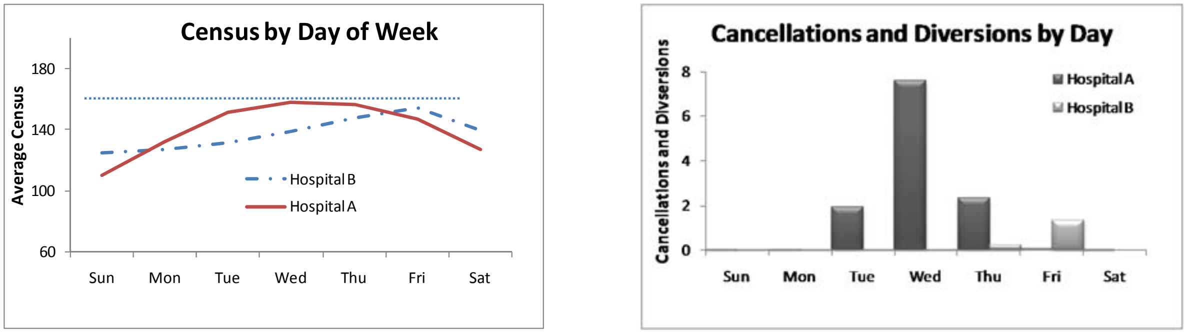

Figure 4 – Comparison Day of Week Metrics between Hospital A and B

In contrast with the point-of-use centered decision making, the management and control of inpatient admissions underscores the value of utilizing the entire patient flow pathway through the hospital. Just as in this illustrative example, any hospital with historical patient flow data can be analyzed using the PFSF. The patient flow simulation can be built with data on the number of hospital units and beds in each unit, daily arrivals (both emergent and scheduled), transfers between units within the hospital, length of stay subject to patient type and what actions are taken when a patient needs admission to a full unit.

5. Conclusion

Patient flow modeling is critical to understanding the effect of decisions on healthcare delivery system functioning. The application of the Patient Flow Simulation Framework to hospital admission control demonstrates the need to consider patient flow in decision making and provides a means to analyze, evaluate and develop policies from a patient flow perspective. The benefits of patient flow informed decision making range from cost savings, to increased throughput and access, to improved quality of care, to greater patient and staff satisfaction.

The power of the PFSF lies in its flexibility, its portability, and its extendibility. Without rewriting the simulation, the PFSF can model almost any kind of system in which patients arrive, receive service, get transferred and leave the system. It can also interface with a variety of optimization software and can be tailored across platforms for large or small-scale implementations. Note, however, that hospital congestion relief using the approach detailed in this paper is just the starting point for a complete healthcare system transformation through patient flow modeling.

The authors believe that all powerful actors in a hospital can benefit from using the PFSF and the admission control policies detailed above to contain their occupancy variability. While nudging surgery schedules in the way we propose must be approached with sensitivity to the internal personnel dynamics of hospitals, the increased throughput from managing census variability allows doctors to do more surgery. The stable census levels simplifies nurse staff planning and creates a more stable work environment by reducing the frequency with which the nurse staff becomes overloaded and underloaded. Hospital administrators appreciate the financial benefits associated with predictable workload. The PFSF can facilitate this and other win-win situations for a healthcare system that needs to contain the ever growing cost of care delivery and high personnel turnover rates.

Once the hospital occupancy is stabilized, the transformation continues with systematic flow modeling to support decision making in the emergency department, operating rooms, nurse staffing, ancillary services, transporters, urgent care facilities and outpatient clinics. At every level in this patient flow healthcare delivery transformation, the PFSF has vast potential to reduce cost and improve quality and access throughout the healthcare delivery process.

Acknowledgements

Here we gratefully acknowledge the support of Professor Walton M. Hancock and a seed grant from the College of Engineering at the University of Michigan. The work of the first author was supported by a National Science Foundation Graduate Research Fellowship.

References

1. Agency for Healthcare Research and Quality. HCUPnet - Healthcare Cost and Utilization Project. January 2009. http://hcupnet.ahrq.gov/ (accessed January 26, 2009).
2. Connors, M.M., 1970, "A stochastic elective admissions scheduling algorithm," Health Services Research, 5(4), 308-319.
3. Forster, A.J.,Stiell, I., Wells, G., et al., 2003, "The effect of hospital occupancy on emergency," Academic Emergency Medicine, 10(2), 127-133.
4. Gallivan, S., Utley, M., 2005, "Modelling admissions booking of elective in-patients into a treatment centre," IMA J Management Math, 16(3), 305-315.
5. Hancock, W.M., Walter, P.F., 1979, "The use of computer simulation to develop hospital systems," SIGSIM Simul. Dig., 10(4), 28-32.
6. Harrison, G.W., Shafer, A., and Mackay, M. 2005. "Modelling variability in hospital bed occupancy," Health Care Management Science, 8(4), 325-334.
7. Helm, J., AhmadBeygi, A., Van Oyen, M., 2009. "Design and Analysis of Hospital Admission Control for Operational Effectiveness," Production and Operations Management Society (submitted).
8. Jun, J.B., Jacobson, S.H., and Swisher, J.R., 1999, "Application of discrete-event simulation in health care clinics: A survey," The Journal of the Operational Research Society, 50(2), 109.
9. Machlin, K., Carper, S.R., 2007, "Expenses for hospital inpatient stays: 2004," Statistical Brief, 164.
10. Sprivulis, P.C., J. Da Silva, I.G. Jacobs, et al., 2006, "The association between hospital overcrowding and mortality among patients admitted via Western Australian emergency departments," Medical Journal of Australia, 208-212.