Improving Patient Flow through Early Bed Requests at UNC Hospital ED: A Discrete-Event Simulation Study

Emily Riederer

Senior Honors Thesis Department of Statistics and Operations Research University of North Carolina, Chapel Hill

April 8, 2016

Approved:

Dr. Serhan Ziya, Thesis Advisor

Dr. Nilay Argon, Thesis Advisor

Dr. Debbie Travers

Acknowledgements

First and foremost, I would like to thank my advisors Dr. Serhan Ziya and Dr. Nilay Argon for their support, guidance, and insight. I would also like to thank Wanyi Chen, doctoral student in UNC's Department of Statistics and Operations Research (STOR), for all of her work estimating the input distributions and more generally for her advice and time working with me.

I would like to thank the rest of UNC Hospitals team collaborating on related research, particularly Dr. Debbie Travers and Dr. Ahbi Mehrotra, for welcoming me to the team and providing additional information about the ED and its operations.

Last but not least, I would like to thank Virginia Ahalt and Yan Hai, former Master of Science students in UNC's Department of Statistics and Operations Research (STOR). Their papers and models were instrumental in providing me with context on the UNC ED and its operations, and their models provided the foundation on which this simulation model was built.

Table of Figures, Charts, and Data Tables

1. Introduction

Emergency department (ED) overcrowding is a widely recognized problem in the United States due to numerous legal, economic, and operational factors [1]. This difficulty has motivated a significant body of clinical and academic interest in applying operations research (OR) techniques to improve patient flow in EDs [2] [3]. In particular, discrete event simulation has been extensively utilized to study each of the three steps of ED patient flow (into, within, and out of the ED) [2]. The flexibility of simulation makes it particularly useful for examining each stage and comparing alternatives to reduce overcrowding [3].

Many approaches to addressing ED overcrowding focus on ways to increase bed capacity for patients requiring service since beds are the bottleneck resource in many EDs [4]. The problem of bed capacity can be addressed at each stage of patient flow. For example, many works have considered the effect of physician-at-triage (PT) altering the flow within the ED by treating low acuity patients with low resource requirements in a separate clinic to preserve more beds for severe patients.

In contrast, early bed request attempts to improve flow out of the ED by reducing "bedblock", the utilization of ED beds by patients who have completed service but board in the ED until an in-patient bed is available. Early boarding is the process of identifying at the time of triage patients who will later be admitted as in-patients. The ED "calls ahead" to request a bed from the appropriate ward so that the in-patient bed is ready when, or soon after, the patient completes service in the ED. In theory, such a policy has great potential to deliver system-wide improvements since many studies recognize bed-block and patient flow out of the ED as major drivers in long LOS and wait durations and "one of the most well-known operational problems to afflict an ED" [2].

This paper describes the development and application of a simulation model for the UNC Hospitals ED to examine the effect of implementing an early bed request policy. Section 2 describes the patient flow at UNC Hospitals ED, the simulation model, the proposed early bed request policy, and how this policy is incorporated into the simulation model. Section 3 provides details on the data used, the estimation of input parameters, and validation of the model. Section 4 reports the results of simulation experiments on early bed request policies.

2. Patient Flow at UNC Hospitals Emergency Department and in Simulation Model 2.1. Background on the Emergency Department with Descriptive Statistics

The ED at UNC Hospitals is a certified Level 1 Trauma Center and is one of the largest referral centers in the state [5]. According to 2012 data, the ED has a throughput of nearly 70,000 patients per year (67,204 observations are recorded, excluding some patients who left without being seen or left against medical advisement), or nearly 200 patients per day (average of 184 in dataset). This section describes the flow of these patients through the ED.

Upon arrival, adult and pediatric patients are assessed by a triage nurse and assigned an Emergency Severity Index (ESI) score, a five-level triage categorization based on criteria of high acuity (ESI levels 1 and 2) and, for lower acuity patients, intensity of expected resource needs (ESI levels 3 to 5, in order of decreasing needs) [6]. The following table reports the aggregate distribution to provide a sense of the relative frequencies of patient types (although in both the actual and the simulated systems, this distribution is nonstationary).

1. ESI Distribution

Next, patients wait for an available bed in one of the ED wards. The ED is open continuously, but two of its four main wards close during non-peak hours. Team A and Team B comprise the main ED and are open 24 hours per day. Team D and Pediatrics closes from 2AM-9AM when patient volumes are low, and all patients can be treated in the main ED. To facilitate this closing, patients are not triaged to these wards after midnight.

All Adult ESI1 and 2 patients (and ESI3 patients identified as acute) are placed in the main ED. Two beds in the main ED are trauma beds are reserved specifically for ESI1.

Team D accommodates less severe adult patients (non-acute ESI3 and all ESI4-5) when it is open. Similarly, when open, lower ESI pediatrics patients can be served in the Pediatrics ward, although highly critical pediatric patients may also be seen in the main ED.

The approximate bed capacity of each ward is reported below.

2. Capacity and Schedule of UNC ED Wards

The rules for bed assignment are complex and subjective, relying on the expert judgements of hospital staff; however, they incorporate prioritization of higher acuity patients over lower acuity patients and a first-in-first-out rule for patients of similar acuity. The mean waiting times (from arrival to first encounter with a care provider) based on ESI groups are suggestive of this prioritization, and their values in the 2012 data are shown below. A small proportion of patients may opt not to wait and leave without being seen (LWBS). This constitutes 1.05% of adults and 0.426% of pediatric patients in the 2012 data, although the actual rates may be higher as some may leave before they are recorded at all.

3. Mean Waiting Time in Minutes by ESI Group

Once a patient is placed in a bed, the patient receives care. This may include many processes including but not limited to multiple interactions with doctors and nurses, consults with specialists, laboratory tests, radiology, and medication. The following table shows mean service times by ESI in the 2012 data.

ESI	Adults	Peds
	96.4	78.9
2	392.6	476.4
3	255.8	191.9
	135	120.9
	71.8	84.3

4. Mean Service Time in Minutes by ESI

After this period of service, a provider makes a decision regarding the patient's disposition: whether the patient will be discharged or admitted to an in-patient ward of the hospital. This decision is followed by another waiting period while the patient remains in his ED bed and either boards (waits for admission) or waits for discharge. Boarding occurs because admitted patients must wait for an in-patient bed to become available, and wait for discharge occurs as patients wait to receive discharge instructions from a provider and check out. The following tables shows the proportion of admitted patients by ESI and age and the mean boarding times by disposition in the 2012 data.

ESI	Adult	Peds
	86.8%	94.3%
2	70.9%	57.2%
3	34.2%	23.8%
	2.8%	2.3%
5	0.5%	0.5%

5. Proportion of Admitted Patients by ESI and Age

6. Mean Boarding Time in Minutes by ESI

Altogether, wait times, service times, and boarding times sum to a patient's total length of stay (LOS) in the ED. The following table summarizes mean LOS in the 2012 data.

7. Mean LOS in Minutes by ESI and Age

2.2. Overview of Simulation Model

The simulation model closely follows the major steps of patient flow at the UNC ED by modeling arrivals, triage and registration, waiting for an open bed, service, and boarding processes. These processes are modeled through a series of queues and stochastic delays, as illustrated in the flowchart below. The model is constructed and run in ARENA simulation software and builds upon the work of former Masters students in UNC's Department of Statistics and Operations Research (STOR), Virginia Ahalt (2013) [7] and Yan Hai (2014) [8].

8. Simulated Patient Flow

This section provide a qualitative description of each of the main steps depicted in the flow chart. The discussion of quantitative elements of the model, such as the estimation of input parameters, is postponed to Section 3.

Arrivals:

Patients arrive according to a nonstationary Poisson process with rates estimated separately for each combination of ESI, age (Adult/Peds), and hour of day. These arrival rates are thinned by the age-dependent proportion of LWBS patients.

Upon arrival, patients are assigned a disposition type (Admit/Discharge) based on the proportion in the data, conditional on ESI and age but not hour. ESI3 Adults are further classified as Acute or Non-Acute by a fixed proportion. This classification affects their bed placement.

Triage and Registration:

Upon arrival, patients enter a queue and wait for the next available of two triage nurses. They are then delayed a stochastic amount of time as they undergo triage.

In the actual ED, patients arriving by ambulance enter through a separate entry point and undergo triage with the charge nurse instead of a triage nurse. However, since the simulation does not include information on patient mode of arrival, this detail is omitted.

Waiting for a Bed:

After triage, patients enter a priority queue to wait for the next available bed. In the actual ED, the process of assigning patients to beds is complex and can depend upon many factors beyond the scope of the simulation model. The simulation uses prioritization rules and ward restrictions to define a reasonable abstraction of this process.

Priorities allow for the implementation of a priority queue in which higher priority entities have access to lower priority entities regardless of waiting time. Within a single priority, access to a resource is first-come-first-served.

Preferred orders specify the order in which an entity selects a resource if multiple are available. For example, a non-acute ESI3 patient may use either Team A, Team B, or Team D, but based on that preferred order, if both a Team B bed and a Team D bed is available, he will go to the Team B bed.

The following table summarizes these rules.

9. Bed Assignment Rules for Simulation

Once an appropriate bed becomes available, patients leave the queue, claim the bed, and occupy this bed until their departure from the system.

Service Process:

For the purpose of the simulation model, service begins when the patient enters a bed and consists of a stochastic delay from a distribution estimated from the data which depends upon ESI, age, and service start time. Specific resources used during service (e.g. doctors and nurses) or procedures done (e.g. labs or radiology) are not explicitly modeled.

Boarding & Waiting for Discharge Process:

In the real ED, the length of boarding depends on capacity and congestion of in-patient wards, which are far beyond the scope of the simulation. Hence, similar to the service process, the boarding process is modeled as a stochastic delay drawn from a distribution. Here, in addition to ESI, age, and boarding start time, the distribution also depends upon disposition.

2.3. Model Limitations and Assumptions

As with any simulation, modeling decisions introduce limitations and assumptions. This section elaborates on these assumptions regarding left without being seen patients, bed assignments, independent process times, and the interpretation of simulation results.

Left Without Being Seen (LWBS) Patients:

Thinning arrival rates by the proportion of LWBS patients assumes that the rate of LWBS is constant throughout the day and across ESIs and that the decision to leave is independent of conditions in the ED. This is unlikely as there would be no motivation for patients to arrive at the ED if they intend to leave without service regardless of their expected wait time.

Alternative approaches to modeling LWBS allow for the volumes to depend on ED conditions. One common approach is to estimate a "patience distribution" of the amount of time patients are willing to wait before leaving, using a variety of non-parametric [4], parametric [9], or proportional [10] methods. A previous model for the UNC ED employed a balking scheme in which patients leave based on the number of patients already waiting for a bed and used this number as a calibration parameter [8].

Treating LWBS patients by thinning the arrival rate was logical due to data limitations and the small proportion of LWBS patients in the dataset (1.05% adults; 0.426% pediatrics). Some LWBS patients are not captured in the dataset and patient covariates are limited, so estimating a model of patient behavior would be difficult. Additionally, thinning the arrival rates has computational advantages over removing the patients at a later point in time. Due to the small percentage, the effect on simulation output is negligible so these patients are excluded from the analysis.

Inflexible Bed Assignment Rules:

In the actual ED, bed assignment rules and restrictions of certain ESIs to certain wards might vary to address short-term capacity issues and assessments of individual patients' conditions. Since the simulation model does not contain this level of detail about patients, the rules described previously attempt to capture the approximate preferences of ED administrators.

A second limiting feature of bed assignment is the assumption that a bed is available for an incoming patient instantaneously upon it being vacated by a leaving patient. In the real system, beds might require preparation time between patients which might, in turn, depend on other resources such as staff to make these preparations.

Independence of Service Times and Boarding Times across Patients:

Service and boarding times are independent draws from distributions depending on patient type and time of day. Since these processes are modeled as delays based upon these distributions and not as processes which require and consume resources (such as the time of care providers), they do not explicitly vary with regard to occupancy or utilization of these latent resources in the model. These distributions implicitly reflect ED and hospital-wide congestion in their time-dependent nature, but they do this only under the assumption that the relationship between hour and congestion is not altered from the conditions under which these distributions were estimated. This issue is discussed further when considering the assumptions and limitations of the early bed request model.

ESI-Specific Effects:

As will be discussed in Section 3, some distributions in the model were estimated by grouping different ESI's together in order to obtain a better fit. Hence, the model's output is

more accurate when similar groupings are used than when different ESI's are analyzed separately.

2.4. Background on Early Bed Request at the UNC Hospitals ED

As described in Section 1, early bed request attempts to reduce "bed-block" by anticipating the need for future in-patient beds and requesting these beds before they are needed. Demand forecasts for in-patient beds can be predicted on a binary basis for each individual patient [11] or aggregated across all incoming patients over a time period [12]. The UNC Hospitals ED is currently interested in implementing the former approach with individual predictions based on a statistical prediction model.

Early boarding not only reduces boarding time and LOS for the selected patients, but, more importantly, aims to improve overall patient flow by reallocating ED beds occupied by boarding patients to patients awaiting service. Clearly, this policy could only improve (or, at worst, leave unchanged) boarding times, wait times, and ED capacity. However, practical implementation challenges require a careful analysis of the magnitude of these benefits relative to the costs.

This initiative requires coordination and cooperation between the ED and in-patient units of UNC Hospitals, so all stakeholders must be convinced of its benefits. Both qualitative and quantitative costs must be considered; for example, preparing an in-patient bed might mean prioritizing the discharge of another patient.

The level of cooperation required also highlights an important trade-off. A "false alarm" request for patients who are ultimately discharged would decrease good will and future buy-in from in-patient wards. Thus, misclassification must be kept at a minimum or avoided completely, even at the cost of not early bed request some highly likely admits. A simulation study can examine the benefits of early bed request even when the number of early-boarded patients is small.

For an example of the statistical prediction process, one model currently being developed is a generalized additive model (GAM) for ESI2 and ESI3 patients. ESI4 and ESI5 patients have such low admission rates that these patients would not be considered for early bed request. Conversely, the population of ESI1 patients is so small and such a high proportion is admitted that there is not sufficient data to estimate a model for ESI1 patients.

For each ESI2 and ESI3 patient, the model outputs a probability that the patient will be admitted. The ED could potentially decide to predict admission for patients whose predicted probability exceeds some cut-off value. A higher cut-off value is more conservative and requires greater certainty that the patient will be admitted. The proportions of correct classification and misclassification for the current GAM model, as measured by cross-validation, are given in the following table.

2.5. Overview of Simulation Model for Early Bed Request

Slight alterations adapt the simulation model described in Section 2.2 to facilitate experimentation with early bed request. The figure below (adapted from the first figure in Section 2.2) illustrates these changes which are described in this section.

11. Altered Simulated Patient Flow Admit/Discharge (A/D) Prediction at Triage

As in the original model, the patient's true disposition type (Admitted/Discharged) is assigned at the time of arrival, and the patient proceeds to the triage queue. Now, an additional step occurs at the end of triage. According to a conditional probability P(Flagged for Admission | True Admission Type; ESI), some patients are classified as being predicted for admission. For these patients, a boarding time is assigned from the time-dependent boarding distribution according to the current system time at triage.

Of course, in the real system, a patient's true disposition type is not known at triage, but this approach reflects the fact that this judgement is made based on patient attributes (e.g. major complaint) which are not in the model. Additionally, these conditional probabilities are a common and easily calculated method for assessing classification methods.

The process of waiting for a bed and receiving service precedes as before, as does the boarding process for non-flagged patients. For flagged patients, the model is altered at the beginning of the boarding process. If the time that has elapsed since a bed was requested exceeds the assigned early bed request time, the in-patient bed is considered to be ready, and the patient immediately departs from the system. Otherwise, the patient boards in the ED for residual time between the boarding time and the time that has elapsed since a bed was requested. In the case of misclassification, the patient waits for discharge as in the original model, and there is no benefit to the system from the reserved bed.

The following table summarizes boarding times for patients under the altered scheme.

12. Boarding Times by Patient Type under Early Bed Request Policy

2.6. Model Limitations and Assumptions related to Early Bed Request

This model makes two major assumptions regarding patient flow under early bed request. In addition, the manner in which the parameters were estimated lead to an additional limitation when interpreting model output.

Independence of Bed Reservations:

The first assumption is that bed reservations are independent, so that the reservation for one patient does not affect any other patient. For example, consider two patients A and B. A is flagged for admission and B is not, but both ultimately require admission. If B finishes service and A's inpatient bed is ready before A has finished service , the bed remains empty and idle waiting for A instead of going to B. Similarly, if patient C is flagged for admission but then discharged, C's reservation is wasted and does not reduce boarding time for some other patient.

The rationale for these assumption is that patients are admitted to many different wards of the hospital, so an in-patient bed reserved for one patient would often not be appropriate for another.

In a similar vein, early bed request may change the distribution of bed requests volume across time, but the model assumes that this does not affect the time-dependent boarding time distribution since the magnitude of any change is small and this is just one factor in a much larger and more complex system governing bed availability throughout the UNC Hospitals.

Impact of Policy on Other Parts of the ED System:

The second assumption is that shifts in the distribution of patients across processes (waiting for bed, service, and boarding) caused by early bed request will not congest other parts of the ED system. The analysis of early bed request patients in Section 4 will show that early bed request can increase the percentage of occupied ED beds containing patients who are receiving service versus boarding. However, the assumption implies that the service times will not increase as a result of greater occupancy of patients being served.

This is justifiable since boarding patients still require nursing, and the length of service time for patients depends on many factors beyond the capacity of doctors and nurses in the ED. For example, patients may need to undergo labs and other tests (which are conducted using hospital-wide resources) or may be waiting under observation.

Impact of Ward Restrictions of Analysis of Early Bed Request Effects:

As discussed previously, the decision to send certain patients to certain wards is more rigid in the simulation than is the case in the actual ED. This also means that the effect of early bed request on certain ESI groups is more limited. For example, reducing boarding times for ESI2 will only help patients in Teams A and B. In the formulation of our model, that means that this will strictly impact on ESI1, ESI2, and ESI3 Acute patients. There might be more of an impact in the case where ESI4 and ESI5 patients were sometimes placed in the main ED during the day.

Similarly, analyzing early bed request for ESI3 patients will impact the main ED (and thus ESI1, ESI2, and ESI3 Acute) with a probability equal to the proportion of ESI3 acute used in the model. Otherwise, it will affect Team D (ESI4-5 and ESI3 Non-acute). This proportion cannot be calculated by the available data and is instead determined by a combination an estimate provided by ED staff and calibration for the model.

One might also assume that ESI3 Acute patients are more likely to be admitted and more likely to predict for admission than ESI3 Non-acute patients. However, the model uses the same proportion for true admission and the same classification rate across all ESI3 patients.

3. System Data and Validation of Simulation Model

3.1. Actual Data and Simulation Inputs

3.1.1. Description of Data

One year of data was obtained from the UNC ED to support the estimation of parameters for this model and its validation. This dataset consists of observations from Jan. 1, 2012 to Dec. 31, 2012 (366 days) during which time the arrival of 67,204 patients was recorded.

The dataset includes information on patient characteristics (name, age, and gender), condition characteristics (chief complaints, acuity as given by ESI), disposition type (admitted, discharged, LWBS, etc.), and key timestamps (arrival, initial encounter with nurse, initial encounter with doctor, disposition time, and departure time).

3.1.2. Simulation Inputs

UNC STOR doctoral student Wanyi Chen conducted the statistical estimation of inputs for the simulation model. The following table summarizes the main inputs to the model and how they were estimated from the actual data. Tables detailing the arrival, service, and boarding input distributions are provided in the Appendix.

13. Main Simulation Inputs Estimated from Actual Data

3.2. Comparison of Outputs

Multiple methods were used to compare the simulation output to the actual system, including both formal tests of equality and ad-hoc numerical and graphical comparisons. Although each method exposes some imperfections, the results generally suggest that the simulation model provides a reasonable approximation of the actual system.

Mean LOS and wait times (from arrival to bed placement) were main performance metrics. Since boarding times and service times are governed by distributions directly computer from the systems data, LOS and wait time are the time quantities derived stochastically from the model. Additionally, these metrics are of particular interest when considering early bed request policies in Section 4. Occupancy level is also considered to further validate the patient flow.

The following table summarizes how the values of the validation metrics were estimated from the systems data for comparison with the simulation output.

14. Validation Metrics Estimated from Actual Data

3.2.1. Confidence Intervals

The following table shows the results of calculating a 95% paired-t confidence interval for the difference in mean times (LOS and wait time) between the real system data and the simulation output. LOS and Wait Time were the clear choices for time intervals to compare since boarding times and service times are governed by distributions which were directly calculated from the system data.

The 95% confidence intervals fail to reject equivalence between the systems based on aggregate length of stay or aggregate wait time. Additionally, slightly larger 99% confidence

intervals fail to reject equivalence of LOS for a partition of patient with by similar traits (the cross product of ESI groups and age classifications).

15. Confidence Intervals for Difference in Means of Simulated versus Actual System Data

Although the simulation passes the tests, the large mean differences for ESI1-2 Adults and ESI 4-5 Adults are still noteworthy. The simulation appears to systematically underestimate length of stay for the most severe patients and underestimate for the least severe.

One difficulty with validating based upon formal statistical testing is the limited data from the real world system. Observations from the actual and simulated systems were divided into 13 batches of 28 days. Both sets of batches exhibited lag-1 autocorrelation. The technical details are found in the Appendix, Section 7.1.

3.2.2. Summary Tables

The following table provides mean LOS and Wait Times from the actual data and simulated system based on 10 replications.

16. Comparison of Actual and Simulated LOS and Total Wait in Minutes

For both adult and pediatric patients, the lower ESIs suffer the worst percentage error in estimating length of stay. However, this is partially attributable to their shorter actual length of stay. Similarly, the highest ESIs with the worst percentage error in estimation of wait times since their true wait time is smaller.

3.2.3. Occupancy Levels by Process

In addition to comparing mean times experienced by patients in this system, it is also useful to consider the average patient volume at each major process (wait, service, and boarding) in the system. The following chart depicts this average occupancy level by hour of day for each of the major processes.

Occupancy levels for the actual distribution were computed by assuming zero patients were in the system at 12AM on January 1, 2012, and accounting for all arrival, bedding, disposition, and discharge timestamps throughout the year. For comparability, the simulation occupancy levels were computed based a single year-long replication with no truncation for the warm-up period.

^{17.} Average Hourly Occupancy Level by Process

From this chart, we observe that the simulation captures approximately the right shape for each occupancy distribution, but the total occupancy distribution appears slightly "shifted" to the right. That is, it follows the same trend as the actual distribution but as a slight lag in time.

The minima for both total occupancy and service occupancy are overestimated, and the maximum occupancy for boarding patients is slightly underestimated.

4. Early Bed Request Analysis 4.1. Experimental Conditions

The effect of early bed request is examined for a range of classification accuracies, i.e. $p = P$ (Flagged for Admission | Actual Admission). Only the case of perfect prediction (p=1) is considered for ESI1 because the volume of ESI1 patients is small and the likelihood of them being admitted is very high. For ESI2-3, scenarios in the following chart were considered. The case of perfect prediction is also reported for ESI2-3 to observe the theoretically best result that could be achieved by early bed request.

18. Classification rates for ESI2 (row) and ESI3 (col) considered

The early bed request policy is applied from 9AM-8PM. The question of the best time range to use depends upon the relative benefits and costs to different units of the hospital and is worthy of future consideration.

The next section analyzes the results of these different scenarios. These results were computed using the batch means method with twenty batches of 180 days (or roughly six months). Before recording data for these batches, the simulation runs for an eight-week warm-up period. Details on the warm-up analysis are found in the Appendix.

4.2. Results

4.2.1. System- and Patient-Wide Effect on LOS and Wait Times

The first consideration for early bed requests is the extent to which they can improve system-wide outcomes and not simply throughput times for individual patients. The following table summarizes the mean wait times (time from arrival until first treatment) and total length of stay for all patients under these different scenarios.

19. Mean Aggregate Wait and LOS across Experiments

We observe that in the case of theoretically perfect prediction accuracy, the maximum decrease in wait time at peak hours of ED congestion is 11 minutes. The results are similar if we consider the 75th percentile of LOS and wait (as shown in the following table) instead of the mean.

20. 75th Percentile Wait Time and LOS

These results are depicted graphically in the following charts which look at how wait time varies based on prediction rates and the number of predictions. It appears that the marginal benefit of additional predictions does not significantly diminish for high levels of prediction.

The main mechanism through which the length of stay decreases is by decreasing the time patients spend waiting for an available bed. To formalize this notion, the following table presents approximate 95% confidence intervals for mean wait times in aggregate and by ESI groups. The intervals highlighted in yellow are those that do not overlap (or come extremely close to not overlapping) the interval for the base model.

23. 95% Confidence Intervals for Wait Time under Experimental Conditions

Of the scenarios considered, prediction accuracy must be at least 50% for both ESI2 and ESI3 in order to realize a statistically significant decrease in wait time. However, much higher levels of prediction are required to achieve a clinically significant decrease.

In most cases, batches had less than 0.2 correlation. Complete information on correlation between batches, and which experiments would benefit most be a long run-length is contained in the Appendix.

After observing these system-wide effects on throughput time, it is also of interest to examine more specifically those patients who were predicted for admission. This is done in the following table which reports the average boarding time required by flagged patients and the average amount of boarding time saved. A slight minority of predicted patients still experience a positive wait time in the ED, but it is greatly reduced. The majority of predicted patients experience no wait.

p ₂	p ₃	Time Still Needed	Average Boarding Average Boarding Time Saved
0	25	137	38
25	0	120	60
25	25	130	47
25	50	134	43
50	25	129	49
50	50	136	41
50	75	136	39
75	50	132	43
75	75	134	41
75	100	136	39
100	75	133	44
100	100	136	41

24. Reduction in Boarding Times for Flagged Patients

25. Number of Patients Completely Avoiding Boarding

4.2.2. Occupancy Levels by Process

In addition to considering how early bed requests affects key time metrics, visualizing the occupancy distributions provides more insight into the mechanics underlying the policies. The following graph shows the occupancy level for waiting patients (from arrival until bed placement). Comparing the maximum (no early bed request) and minimum (early bed request with perfect prediction) by hour, this shows that at most the ED reduces the number of waiting patients by 2 at its peak hours.

26. Waiting Distribution by Hour under Different Boarding Scenarios

The following graph depicts the percentage of occupied ED beds which hold a patient receiving service, as opposed to boarding or awaiting discharge. As expected, increasingly aggressive policies to reduce bed-block increase this percentage. With theoretically perfect prediction, the maximum boost in this metric is about 3% which translates to approximately two more patients.

27. Proportion of Patients in ED Beds who are Receiving Service

4.2.3. Idle In-Patient Bed Times

The previously results suggest that early bed request has the potential to make possibly small but significant changes in patient throughput. However, this must be contextualized in terms of the cost. As discussed previously, many of these costs (e.g. cooperation) are qualitative in nature. One quantifiable metric is the amount of time that in-patient beds for which an early bed request was made, sit empty while waiting for a patient. The following table reveals that bed waste is not insignificant under the early bed request policy. On one hand, this might suggest that delaying until a later point in the patient's path through the ED might be beneficial; however, there is a trade-off in that delaying the bed request would also decrease the small positive benefit observed by implementing early bed request.

28. Analysis of Wasted In-Patient Bed Time: Time (Minutes) In-Patient Bed Ready ahead of Patient

4.3. Future Considerations

This simulation study suggests the promise of early bed request in improving systemwide outcomes. However, there are numerous further considerations.

Optimal Hours for Applying the Policy:

This study applies the early bed request policy from 2PM-8PM, during which time patient arrivals and ED congestion are growing steadily. However, both time specific hours of day and days of the week that the policy is implemented bear further consideration.

Optimal Time to Initiate an Early Bed Request:

The simulation always assumes that the bed request is based immediately at triage. However, if the expected wait and service time in the ED far exceeds the expected boarding time at triage, delaying the bed request might be a better strategy to reduce the time reserved inpatient beds sit idle in anticipation of an incoming ED patient.

A related issue is that more accurate predictions might be able to be made as a patient begins to undergo service and tests.

Diminishing Benefits:

The issue of whether the benefits from early bed request is not answered by this study since the impact of predicting ESI1-3 simultaneously is not considered. This is an important consideration when considering the costs and benefits of implements such a policy in the actual ED. Furthermore, if this policy were enacted in conjunction with discharge-related initiatives, the point at which benefits begin to diminish might also be reduced.

5. Conclusion

This paper provides an overview of a simulation model for the UNC Hospitals ED which was developed with the intention of including appropriate specificity to model different strategies to reduce bed-block. The simulation experiments suggest early bed request can be

effective and provide insight into manners in which this might be done conservatively. These include shortening the critical timeframe in which early bed request is used and postponing the initiation of an early bed request until later in the process.

6. Works Cited

- [1] W. J. Hopp and W. S. Lovejoy, "Emergency Department," in *Hospital Operations: Principles of High Efficiency Health Care*, New Jersey, FT Press, 2013, pp. 21-113.
- [2] S. Saghafian, G. Austin and S. J. Traub, "Operations research/management contributions to emergency deparment patient flow optimization," *IIE Transactions on Healthcare Systems Engineering,* vol. 5, no. 2, pp. 101-123, 2015.
- [3] S. A. Paul, M. C. Reddy and C. J. DeFlitch, "A Systematic Review of Simulation Studies Investigating Emergency Department Overcrowding," *Simulation,* vol. 86, no. 8-9, pp. 559- 571, 2010.
- [4] O. Yildiz, M. F. Kamali and T. Tezcan, "Analysis of Triage Systems in Emergency Departments," *Simon Graduate School of Business - Working Paper No. FR 15-20,* 2015.
- [5] U. S. o. Medicine, "Emergency Medicine," [Online]. Available: https://www.med.unc.edu/emergmed. [Accessed 15 March 2016].
- [6] N. Gilboy, P. Tanabe, D. Travers and A. Rosenau, "Emergency Severity Index (ESI): A Triage Tool for Emergency Department Care," Agency for Healthcare Research and Quality, 2012.
- [7] V. Ahalt, "Comparison of Emergency Department Crowding Scores: A Discrete-Event SImulation Approach," 2015.
- [8] Y. Hai, "A Discrete Event Simulation of UNC Emergency Department," *Paper for Satisfaction of Master of Science Degree Requirement,* 2014.
- [9] S. Saghafian, W. J. Hopp, M. P. Van Oyen, J. S. Desmond and K. L. Steven, "Patient Streaming as a Mechanism for Improving Responsiveness in Emergency Departments," *Operations Research,* vol. 60, no. 5, pp. 1080-1097, 2012.
- [10] G. Zayas-Caban, J. Xie, L. Green and M. E. Lewis, "Optimal Control of an Emergency Room Triage and Treatment Process," *Columbia Business School Research Paper No. 14- 51,* 2014.
- [11] S. Qiu, R. Babu, A. Murat, B. Batarse, H. Neemuchwala and W. Jordan, "A cost sensitive inparitent bed reservation approach to reduce emergency department boarding times," *Health Care Management Science,* vol. 18, no. 1, pp. 67-85, 2014.
- [12] J. S. Peck, J. C. Bennevan, D. J. Nightingale and S. A. Gaehde, "Predicting emergency department inpatient admissions to improve same-day patient flow," *Academic Emergency Medicine,* vol. 19, no. 9, pp. 1045-54, 2012.

7. Appendix 7.1. Hourly Arrival Rates

The following table shows the aggregate hourly arrival rates, as estimated from the data and used to model arrivals as a nonstationary Poisson process in the simulation.

29. Appendix: Table of Mean Arrivals per Hour by Day of Week

7.2. Service Time and Boarding Time Input Distributions

In estimating the service and boarding distributions, some ESIs groups were combined in order to obtain a better fit. The following table summarizes which patient categories were used in each estimation. These combinations are important to consider when analyzing the simulation output and attempting to disaggregate the effect of policy changes by specific ESIs.

30. Appendix: Patient Groups for Service and Boarding Estimations

The following tables show the best service and boarding distributions found to fit the data. In each of the tables, Hour denotes the hour in which the respective process was started.

Service Time Distributions

Adults

Peds

31. Appendix: Service Distributions by Age and ESI

Boarding Time Distributions Adults – Admitted

Adults – Discharged

Peds – Admitted

Peds Discharged

32. Appendix: Boarding Distributions by Age, ESI, and Disposition

7.3. Batch Correlations for Confidence Intervals (Validation)

Computation of the confidence intervals was complicated by the limited amount of data for the real-world system. To generate multiple observation for each metric of interest (LOS and wait time, in aggregate and in separated patient groups), the real data was divided into 13 batches of 28 days each (thus, an equal number of occurrences of each day of the week in the hope that they should be identically distributed). Means were computed for each batch.

To maximize comparability of the data sources, the simulation was also run for a single replication and divided into 13 28-day batches.

These batches fail to meet the independence assumption of the paired-t confidence interval. If a longer timeframe of real world data were available, this could potentially be corrected through the use of longer batches. However, with a single year of data, there are few solutions. The following table shows the lag-1 autocorrelation among batches for the actual and simulated data.

33. Appendix: Batch Correlations for Confidence Interval Construction (Validation)

7.4. Warm-Up Analysis

Warm-up analysis was conducted for LOS, wait time, and occupancy level. Occupancy level took the longest to reach a steady state, so it is the limiting factor in determining the truncation length. The chart below depicts the simulated occupancy level over 130 days in blue and a smoothed trend lines in orange (using a moving average with window size 5).

Steady state behavior appears to begin around day 50, or slightly over seven weeks into the simulation. Thus, analysis conservatively considered the warm-up period to be eight weeks.

34. Appendix: Warm-Up Analysis on Occupancy Level

7.5. Batch Correlations for Confidence Intervals (Analysis)

The following table shows the correlation of average wait time across batches in each of the simulation scenarios considered. . Values exceeding 0.2 in magnitude are given in red.

35. Appendix: Batch Correlations for Confidence Interval Construction (Analysis)