The Impact of Delays on Service Times in the Intensive Care Unit

Carri W. Chan
Decision, Risk, and Operations, Columbia Business School, New York, NY 10027, cwchan@columbia.edu

Vivek F. Farias
Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA 02139, vivekf@mit.edu

Gabriel J. Escobar
Division of Research, Kaiser Permanente, Oakland, CA 94612, gabriel.escobar@kp.org

Mainstream queueing models are frequently employed in modeling healthcare delivery in a number of settings, and further are used in making operational decisions for the same. The vast majority of these queueing models ignore the effects of delay experienced by a patient awaiting care. However, long delays may have adverse effects on patient outcomes and can potentially lead to longer lengths of stay (LOS) when the patient ultimately does receive care. This work sets out to understand these delay issues from an operational perspective. Using data of over 57,000 Emergency Department (ED) visits, we use an instrumental variable approach to empirically measure the impact of delays in ICU admission, i.e. ED boarding, on the patient's ICU LOS for multiple patient types.

Capturing these empirically observed effects in a queueing model is challenging as the effect introduces potentially long range correlations in service and inter-arrival times. We propose a queueing model which incorporates these measured delay effects and characterize approximations to the expected work in the system when the service time of a job is adversely impacted by the delay experienced by that job. Our approximation demonstrates an effect of system load on work which grows much faster than the traditional $1/(1-\rho)$ relationship seen in most queueing systems. As such, it is imperative that the relationship of delays on LOS be better understood by hospital managers so that they can make capacity decisions that prevent even seemingly moderate delays from causing dire operational consequences.

Key words: Delay effects, queueing, Healthcare

1. Introduction

Delays arise routinely in various healthcare settings: they are a consequence of the inherent, highly variable requirements of healthcare services and the overwhelming demand for these services. It is natural to conjecture that delays in receiving care can result in a variety of adverse outcomes – and indeed, there is some support in the medical literature for such conjectures (e.g. Chalfin et al. (2007), Renaud et al. (2009), de Luca et al. (2004)). This paper proposes to study one such adverse outcome in the intensive care setting: delays in receiving intensive care can result in longer lengths of stay (LOS) in the Intensive Care Unit (ICU). From an operational perspective, this effect has two consequences. The first, of course, is the immediate impact on the delayed patient. The second, *systemic* impact is the increased congestion caused by the increased care requirements for the delayed patient. In particular, the increased ICU LOS can result

in delays to *other* patients requiring the same ICU resources, which in turn results in longer LOS for those patients, and so forth. This paper will (empirically) study the extent of this phenomenon across multiple patient types. We then propose to modify extant queueing models (that are frequently used to model such systems) to account for the phenomenon and present a theoretical analysis for the same.

Delays and the ED-ICU Interface: Patients who arrive to a hospital via the Emergency Department (ED) first under go assessment and stabilization. If a decision is made to admit a patient into the hospital, this patient may 'board' in the ED while waiting to be admitted. Such delays occur for patients of all severities and is often due to unavailability of inpatient beds (Shi et al. 2015). It is particularly troubling when delays occur for the most critical patients–those destined for the ICU. ICUs provide the highest level of care and are very expensive to operate. As such, these units tend to be small, resulting in frequent delays in ICU admission.

Hospitals have adopted a number of approaches to deal with ICU congestion. For instance, ICU congestion can result in discharging current patients preemptively (Chalfin 2005, Dobson et al. 2010, Kc and Terwiesch 2012, Chan et al. 2012), blocking new patients via ambulance diversion (Allon et al. 2013) or rerouting patients to different units (Thompson et al. 2009, Kim et al. 2015). In this work, we focus on a frequent symptom of this congestion: admission delays. Indeed, congestion in the ICU often forces patients to wait in the ED until an ICU bed becomes available (see Litvak et al. (2001)). With an increase in critical care usage (Halpern and Pastores 2010) and a relatively stagnant supply of ICU beds, it is no wonder that delays for patients awaiting ICU admission are growing. In fact, there exists a shortage of ICU beds, which is projected to persist (Green 2003).

This paper will focus on the flow of patients from the ED into the ICU. In particular, we will examine the 'boarding' delay experienced by these patients and the impact of this delay on the length of the patients' stay in the ICU. The effect we study is in contrast to the previously studied phenomenon of 'speeding up' current patients (e.g. Kc and Terwiesch (2012))]. Because we see little evidence of this effect in our patient cohort, we aim to gain a better understanding of the impact of congestion on ICU admission delays, rather than its impact on ICU discharges.

Standard Queueing Models Fall Short: Queueing models are often used to model and analyze patient flows in hospital settings. These models are predictive and can provide valuable insight into the impact of changing demand scenarios as well as staffing or, more generally, capacity provisioning alternatives. See Green (2006) for an overview of how queueing models have been used in healthcare applications. The vast majority of these queueing models assume that the service requirement of a job is independent of the state of the queue upon its arrival. In a healthcare setting, this assumption is equivalent to ignoring the effects of delay on LOS experienced by a patient awaiting care. As we show in this paper, this is not a tenable assumption. In addition, there have been various condition specific studies in the medical community demonstrating that delays can result in an increase in mortality (de Luca et al. 2004, Chan et al.

2008, Buist et al. 2002, Yankovic et al. 2010) and/or extend patient LOS (Chalfin et al. 2007, Renaud et al. 2009, Rivers et al. 2001). We will explore this phenomenon for a variety of patient types.

As we shall see, even in the simplest settings, the underlying queueing process exhibits long range dependencies and, consequently, Markovian models of the same are high dimensional. This is not surprising, since capturing the delay effect creates long-run correlations between service times and inter-arrival times–bursts in arrivals will correlate with longer service times–and very little can be said about such systems. While such models may still be beneficial in simulation, the queueing phenomena made transparent by simple $M/M/s$ type models is obscured. As such, an important component of this paper is a simple set of closed-form approximations to a key performance metric for such systems.

Contributions: While physicians recognize that delays are detrimental for an individual patient, our analysis provides insight into the impact such delays may have on increasing overall congestion and reducing access to care for other critical patients. This work is the first to rigourously analyze the impact of delays on LOS. In particular, we make the following contributions:

1. Using retrospective data of over 57,000 patients from a large hospital network, we empirically estimate the impact of delays in transfers from the ED to the ICU on LOS for *multiple types* of critically ill patients. Our empirical study is granular and characterizes the magnitude of this effect for a variety of patient primary conditions. We estimate a Heckman selection model with an endogenous regressor and find strong evidence that increased ED boarding times are associated with longer ICU lengths of stay for a number of patient conditions. Loosely, for some primary conditions (such as Vascular), a single additional hour of boarding delay (relative to mean delay) is associated with an approximately 11.37% increase in ICU LOS.

2. Next, we examine the implications of this delay effect when considering queueing models often used to model hospital systems. We develop an $M/M(f)/s$ queueing model as an analogue of an $M/M/s$ queueing model, where service times are exponentially distributed with mean which increases with congestion according to a growth function f. We present a rigorous, analytically tractable approximation to such models that, in addition to being quite accurate, provides a simple, transparent view of the impact of congestion on the amount of work in the system in the *presence* of the delay effect. We find a relationship between system load and expected work which grows much faster than the $1/(1-\rho)$ relationship seen in most queueing systems. We view the simplicity of these approximations as surprising since queueing systems with long-range correlations in service and inter-arrival times are known to be notoriously difficult to analyze.

3. We use numeric and simulation results to demonstrate that, due to the relationship exhibited by our queueing model with delay effects, it is imperative for hospital managers to carefully characterize the delay phenomenon for their patient cohorts. Ignoring the impact of delays on LOS when making operational decisions can result in persistent over-crowding, where delays can spiral out of control much faster than anticipated.

1.1. Related Literature

Our work is related to three main bodies of research: empirical work looking at the effect of delays and congestion on patient outcomes as well as other empirical work focusing on estimation methodologies; queueing models with congestion-based dynamics; and, queueing models in healthcare.

The medical community has invested significant effort into measuring the detrimental impact of delays on patient outcomes. The majority of this work has focused on a binary notion of delay: was a patient delayed or not? For instance, a transfer from the Emergency Department (ED) to the Intensive Care Unit (ICU) was labeled as 'delayed' if it was greater than 6 hours (Chalfin et al. 2007); however, there was no distinguishing between 6 and 20 hours of delay. They find that the median hospital length of stay (inclusive of ICU and general medical ward stay) is 1 full day longer and the in-hospital mortality rate was 35% higher for patients who were boarded more than 6 hours. The definition of delay can be on the order of minutes as in the case of cardiac patients (de Luca et al. 2004, Buist et al. 2002, Yankovic et al. 2010, Chan et al. 2008) or up to 5 days for burn-injured patients (Sheridan et al. 1999). All of these works focus on a single patient condition in a single hospital and may lead one to conjecture that the delay effect is isolated to a narrow section of the patient population that visits the ICU. We verify instead that the delay-effect is prevalent across multiple hospitals and ailments. For some conditions, we do not find evidence of a delay effect, suggesting hospital administrators must be prudent about the composition of their patient population when making operational decisions.

Our empirical approach leverages fluctuations in congestion of inpatient units. Kc and Terwiesch (2009, 2012) and Anderson et al. (2011) consider how high load impacts ICU LOS following surgery. These works find that high occupancy levels can result in *shorter* patient length-of-stay (LOS) due to a need to accommodate new, more critical patients. Moreover, such reductions in LOS can increase risks for readmission and death. In contrast, our work considers the *admission*, instead of discharge, process which is altogether a fundamentally different medical decision. In particular, we examine how the occupancy level in the unit which a patient should be admitted can *increase* LOS in the current and subsequent unit. Notice the delay we consider and the speedup effect seen in these prior works actually work in opposition. We find that for the patient population we consider, speedup seems to have little, if any, effect. Kim et al. (2015) also considers the impact of the occupancy levels of downstream hospital units; however, the focus is on how high occupancy levels can affect patient routing and subsequently, patient outcomes. In the present work, we focus on the ICU and how congestion impacts delays rather than the routing of patients to a potentially less desirable recovery unit. That said, the findings of Kim et al. (2015) are evidence of potential sample selection issues which may arise if one only considers patients who are admitted to the ICU. Our setting has a number of econometric challenges: sample selection and endogenous regressors (sicker patients have priority for admission, so have shorter boarding times, but also are more likely to have longer LOS). As such,

we leverage the methodology established in Heckman (1979), Meijer and Wansbeek (2007) to estimate our model.

Shi et al. (2015) also consider ED boarding, but focuses on the impact of hospital discharge policies on patient boarding. Similar to our work, they consider empirical analysis to motivate stochastic models. Using simulation models, they approximate inpatient operations in a hospital in Singapore. In our work, we aim to provide analytic approximations to the impact of ED boarding on system dynamics such as average number of patient hours in the system, i.e. what is the aggregate number of hours all the patients currently in the system will spend in the ICU.

Motivated by our empirical findings, we consider how to incorporate the measured delay effect into our queueing models. Powell and Schultz (2004), Ata and Shnerson (2006), George and Harrison (2001) all consider queueing systems where service times can be increased or decreased depending on congestion. In general, they find that service rates should *increase* with congestion. In a similar vein, Anand et al. (2010) examines the quality-speed tradeoff in an M/M/1 queue where service times can be reduced at the expense of service quality while reducing delay costs and find that the equilibrium behavior is starkly different than in traditional queueing models. We also compare the impact of congestion-dependent service times to traditional queueing models; however, in contrast to these papers, we study a system where the service rate is not controlled but a function of the system's history and tackle the long range correlations which arise from these effects.

Whitt (1990) and Boxma and Vlasiou (2007) examine a G/G/1 queue with service times and interarrival times which depend linearly on delays. Under very special conditions–e.g. the workload must decay over time, or interarrival times must increase as service rates decrease–stability conditions and approximations to the waiting times can be derived. While both of these works consider workload that may increase with delay, the dynamics of our system are very different. In particular, our interarrival times are not a function of service rates, which is required for the results in Whitt (1990) and Boxma and Vlasiou (2007). Consequently, the workload in our system will never decay as it must in the aforementioned works.

In recent work, Dong et al. (2015) attempt to model the queueing phenomenon at hand by having the instantaneous service rate decrease with congestion, rather than having congestion impact individual jobs (patients). They analyze this system in a heavy traffic regime, ignoring the granular modeling we undertake here, and additionally assuming that an abandonment process regulates the system. While such a model is potentially quite useful to understand phenomena such as diversion to other units, it is unclear how their results apply here (i.e. without abandonment), and for finite sized systems.

While there has been important work focusing on state-dependent queueing systems, they are unable to fully capture the healthcare specific dynamics which are estimated from real hospital data and presented in this paper. Our goal is to develop a framework which accounts for the type of delay effect which can appear in a healthcare setting. In doing so, we hope to expand the way queueing models can be used in such a

setting. Queueing theory has been a useful tool to estimate performance measures, such as waiting times, and to provide support in operational decision making, such as determining staffing levels. For instance, Yankovic and Green (2011) consider a variable finite-source queueing model to determine the impact of nurse staffing on overcrowding in the Emergency Department. In a related vein, de Véricourt and Jennings (2011) consider an M/M/s/n queue to estimate the impact of nurse-to-patient ratio constraints on patient delay. Green et al. (2006) modified the traditional M/M/s queueing model to develop time-varying staffing levels for the Emergency Department. To the best of our knowledge, despite the ever-present delay effect in healthcare applications, no other works have explicitly taken it into account.

2. Empirical Motivation: Model and Analysis

In this section, we empirically examine the impact of delays for patients being transferred from the ED to the ICU. We find that delayed transfers from the ED to the ICU are associated with increases in ICU LOS. These findings have significant implications for capacity planning and resource allocation in the ICU. We will estimate a reduced form model that relates patient physiological factors and ED boarding time to ICU LOS. We examine the impact of boarding time across different patient categories.

2.1. Data

We analyze a large patient data set collected from 19 facilities within a single hospital network covering urban and suburban locales for a total of 212,063 patient admissions over the course of 18 months. The largest hospital had a maximum ICU occupancy of 44 patients, while the average ICU size was 19 beds. These ICUs have an average occupancy level of 70%. This data includes patient level characteristics such as age, sex, primary condition for admission (i.e. congestive heart failure or pneumonia), and four separate severity scores based on lab tests and comorbidities. It also includes operational data which tracks each patient through each unit, marking time and dates of admission and discharge. Hospital units were classified into six broad categories including Emergency Department (ED), General Medical Ward, Transitional Care Unit (TCU), Intensive Care Unit (ICU), Operation Room (OR), and Post Anesthesia Recovery Unit. This allowed us to calculate the hourly occupancy level in each hospital unit. In order to avoid censored occupancy levels, we restricted our analysis to patients who were admitted during the middle 12 months of the study. As this was an *inpatient* dataset, the captured time in the ED is the time difference between the order to admit to an inpatient unit and when the patient actually left the emergency department. Hence, this captures the *ED boarding time* and is measured as the time from when the admit order was placed until the patient is physically admitted to an inpatient unit. Note that this does not include the time for triage, stabilization, and assessment, all of which will typically be activities that occur prior to the request for an inpatient bed.

110,574 patients were admitted via the ED. We consider patients whose admission was classified as ‘ED, medical’, i.e. their admission was via the ED and their ailment was not considered surgical (the flow of

surgical patients is rather different and governed by surgical schedules, so such patients were excluded). Similarly, we excluded patients who were admitted to the Operating Room (OR) directly from the ED[1]. To understand the impact of delay on different patient types, we classify patients based on over 16,000 ICD9 admission diagnosis codes into 10 broad groups of ailments based on the types of specialists who treat them: Cancer, Catastrophic, Cardiac, Fluid&Hematologic, Infectious, Metabolic, Renal, Respiratory, Skeletal, and Vascular (Escobar et al. 2008). While there are some patients who do not fall into one of these categories, we focus on these main groupings which the majority of patients fall under.

Severity scores in the data were determined at the time of hospital admission and capture the severity of the patients at the time the request for an inpatient bed was made. In order to use these scores for risk adjustment, we excluded all patients who were admitted to an inpatient bed more than 48 hours after hospital admission since it is unlikely the scores will accurately measure the severity of patients after that. These scores are used for the over 3 million patients in this hospital network and have similar predictive power as the APACHE and SAPS scores with c statistic in the 0.88 range (Zimmerman et al. 2006, Moreno et al. 2005). See Escobar et al. (2008), Chan et al. (2012), Kim et al. (2015) for further description and use of these severity scores. We also restricted our analysis to patients whose hospital stay was less than 60 days. Patients who stayed longer are outliers with LOS more than 6 standard deviations greater than the mean and are unlikely to be representative of the general patient population.

2,930 patients were removed from the sample because they died. This is common practice in the medical community because various factors, such as Do-not-resuscitate orders, can bias LOS estimates for patients who die (Norton et al. 2007, Rapoport et al. 1996). We note that we verified the robustness of our empirical analysis by also including patients who died and find our results are quite similar. When determining occupancy levels, all patients who are treated in the hospital are included.

The data cleaning process is depicted in Figure 9 in Appendix A. The final dataset consisted of 57,063 patients. 5,996 of these patients were admitted to the ICU from the ED. Table 1 summarizes the statistics for the different patient categories.

We wish to understand how delays to ICU admission impacts patient ICU LOS and whether ICU LOS is increasing in ED boarding time. While such a relationship is natural to conjecture, the significance this phenomenon can play in capacity management (as we will see in the subsequent sections) merits that we establish its veracity rigorously. In addition, the empirical study in this section will also allow us to quantify the magnitude of the delay effect for different classes of patients.

2.2. Estimation

We now describe our model which forms the basis for our estimate of the impact of boarding delay on ICU LOS. To test our hypothesis, we consider the ICU LOS, $ICULOS$, and ED boarding time of each patient,

[1] Note that patients admitted to a medical service can go into the OR for surgery.

Condition Category	Non-ICU admits N	Non-ICU admits ED boarding (hours)	Non-ICU admits Age (years)	ICU admits N	ICU admits ED boarding (hours)	ICU admits Age (years)	ICU admits ICU LOS (hours)
Cancer	507	3.54 ± 5.02	65.93 ± 14.32	27	4.28 ± 4.99	52.50 ± 36.96	64.89 ± 11.01
Cardiac	17772	3.46 ± 4.28	68.12 ± 14.68	2203	3.58 ± 4.22	37.75 ± 36.59	66.09 ± 14.32
Catastrophic	1278	3.88 ± 4.93	69.38 ± 19.65	685	2.77 ± 3.91	87.15 ± 83.70	62.20 ± 18.37
Fluid&Hem.	2900	3.54 ± 4.58	68.47 ± 18.11	164	4.30 ± 5.24	45.78 ± 47.79	64.70 ± 16.10
Infectious	11379	3.97 ± 5.10	66.92 ± 19.11	1012	3.85 ± 4.71	65.73 ± 16.86	74.75 ± 84.08
Metabolic	2979	3.75 ± 4.68	63.05 ± 19.62	650	2.87 ± 3.30	51.70 ± 57.21	48.64 ± 19.92
Renal	1753	3.62 ± 4.86	67.11 ± 17.77	123	3.49 ± 4.62	64.04 ± 63.75	60.67 ± 16.44
Respiratory	6487	3.90 ± 5.01	68.45 ± 15.91	741	3.32 ± 4.05	65.50 ± 75.96	66.32 ± 15.62
Skeletal	2727	3.45 ± 4.45	69.34 ± 18.33	98	4.83 ± 5.78	52.70 ± 55.09	66.00 ± 18.70
Vascular	3285	3.45 ± 4.37	71.10 ± 14.23	293	3.28 ± 3.92	53.01 ± 42.01	69.70 ± 13.71

Table 1 Mean $\pm$ Standard deviations are reported for 10 patient categories.

$EDBOARD$. Due to the long tails in ICU LOS, we take the logarithm of ICU LOS. Further, let Z be a matrix of control variables for each patient which includes various physiologic and operational factors which may affect ICU LOS, such as patient severity, age, primary condition, day of admission, and hospital where care is received for each patient. See Table 3 in Appendix A for more details. Our model is then:

$$\log(ICULOS) = \hat{\beta}^T Z + \delta EDBOARD + u \tag{1}$$

The zero-mean noise term u is assumed to be uncorrelated with Z. The coefficient δ may be interpreted as measuring how each additional hour of ED Boarding increases expected ICU LOS: $\delta > 0$ would support our hypothesis. We will run separate analyses for each patient ailment group (e.g. Cardiac versus Cancer patients) to see if and how the delay effect, δ, varies.

2.2.1. Econometric Challenges Due to ethical and practical concerns, it is not possible to run a randomized experiment to see how delay affects patients. Hence, we focus on using retrospective data to estimate this effect. This introduces a number of challenges in our estimation.

Endogeneous regressors: There may be unobservable factors which impact both ED Boarding and ICU LOS. For instance, a very severe patient may be given priority and transferred to the ICU earlier than other, less severe patients. Because he is severe, he will also have a longer ICU LOS due to the increased time required for recovery. If these severity factors are unobservable, they could make our estimate of δ biased.

Selection bias: In this work, we only consider the impact of delay on patients who are admitted to the ICU. However, only 11% of patients admitted to the hospital via the ED are admitted to the ICU. Kim et al. (2015) finds that when the ICU is busy, fewer patients are admitted into the unit so that only the very severe patients receive ICU care. This would result in longer boarding times (due to ICU congestion), but also longer ICU LOS (due to the increased severity of admitted patients). Because the selection of patients is non-random (it depends on patient severity which may not be completely observable in the data), our estimates of δ may be biased.

2.2.2. Estimation Approach For ease of notation, throughout this section, we will let $Z_1, Z_2,$ and Z_3 denote matrices which contain control variables as in Z from Eqn. (1).

In isolation, each challenge can be addressed via established tools which have been used extensively in the econometrics literature. An instrumental variable approach (e.g. two-stage least squares) is a common approach in estimating models with endogenous regressors (Wooldridge 2002). Note that this approach requires a valid instrument, which is uncorrelated with the unobservable noise, but influences the outcome (ICU LOS) through its impact on the endogenous regressor (ED LOS). For instance, the following equations can be used to model such a system:

$$\log(ICULOS) = Z_1\beta_1 + \delta EDBOARD + u_1 \tag{2}$$

$$EDBOARD = Z_2\beta_2 + u_2 \tag{3}$$

where, by assumption the u_i's are 0 mean and independent of Z. Note that u_1 may be correlated with $EDBOARD$, so that a valid identification strategy requires Z_2 to contain an instrument for $EDBOARD$. Replacing $EDBOARD$ in Eqn. (2) with the predicted value from Eqn. (3) allows one to identify the impact of ED boarding on ICU LOS. However, such an approach can become problematic when there is sample selection. In particular, if $ICULOS$ is only observed for a subset of the patients, the assumption of uncorrelated noise in the first equation may not hold:

$$\text{e.g. } E[u_1|Z_1, ICUADM = 1] \neq 0 \text{ or } E[u_1|Z_2, ICUADM = 1] \neq 0$$

where $ICUADM$ is a variable which indicates admission of the patient into the ICU. To account for the potential biases introduced by sample selection, the selection of patients into ICU treatment can be modeled via a Probit model and the Heckman model introduces a correction factor, the Inverse Mills Ratio, to account for the potential biases due to the fact that data only exist for selected observations (Heckman 1979, Wooldridge 2002). Note that this approach requires data from both the selected and unselected samples to estimate the selection model. The selection model is thus:

$$ICUADM = 1(Z_3\beta_3 + u_3 > 0) \tag{4}$$

where u_3 is an error term which assumed to be a standard normal random variable which is uncorrelated with Z_3. Additionally, we assume that (u_1, u_3) is independent of Z and satisfies the following parametric relationship[2]:

$$E[u_1|u_3] = \gamma u_3 \tag{5}$$

[2] Note that other relationships between the error terms can be considered, but this particular assumption is standard as seen in Heckman (1979).

Consider the following steps of algebra, which are used to modify the $\log(ICULOS)$ equation in (2):

$$\begin{aligned}\log(ICULOS) &= Z_1\beta_1 + \delta EDBOARD + E[u_1|Z, ICUADM = 1] - E[u_1|Z, ICUADM = 1] + u_1 \\ &= Z_1\beta_1 + \delta EDBOARD + E[u_1|Z, ICUADM = 1] + e \end{aligned} \tag{6}$$

where $Z = [Z_1, Z_2, Z_3]$ and $e = u_1 - E[u_1|Z, ICUADM = 1]$. It is easy to see that, by construction, $E[e|Z, ICUADM = 1] = 0$.

Now we consider the term $E[u_1|Z, ICUADM]$:

$$\begin{aligned}E[u_1|Z, ICUADM] &= E[E[u_1|Z, u_3, ICUADM]|Z, ICUADM] \\ &= E[E[u_1|Z, u_3]|Z, ICUADM] \\ &= \gamma E[u_3|Z, ICUADM]\end{aligned} \tag{7}$$

where the first equality comes from iterated expectations, the second equality comes from the fact that (Z, u_3) uniquely defines $(Z, u_3, ICUADM)$ as given by equation (4), and the last equality follows from parametric assumption (5) and the assumption that u_1 and u_2 are independent of Z. Then, by our selection model in (4), we have that $E[u_3|Z, ICUADM = 1] = E[u_3|u_3 > -Z_3\beta_3] = \lambda(Z_3\beta_3)$, where $\lambda(\cdot) = \phi(\cdot)/\Phi(\cdot)$ is the Inverse Mills Ratio and ϕ and Φ are the pdf and cdf of a standard normal, respectively. Inserting this into (6) results in the following reduced form model for ICU LOS for the sample of patients who are admitted to the ICU:

$$\log(ICULOS) = Z_1\beta_1 + \delta EDBOARD + \gamma\lambda(Z_3\beta_3) + e$$

Since we potentially have both a selection and endogenous regressor issue, our problem can be cast as a Heckman selection model with an endogenous regressor (Schwiebert 2012, Wooldridge 2002). Our model is thus:

$$\begin{aligned}\log(ICULOS) &= Z_1\beta_1 + \delta EDBOARD + \gamma\lambda(Z_3\beta_3) + e \\ EDBOARD &= Z_2\beta_2 + u_2 \\ ICUADM &= 1(Z_3\beta_3 + u_3 > 0)\end{aligned} \tag{8}$$

where Z_1, Z_2, Z_3 are exogenous controls that they are uncorrelated with e, u_2, u_3. e is zero mean conditional on Z and $ICUADM = 1$, while u_2 and u_3 are zero mean conditional on Z for the whole population. $EDBOARD$ is potentially endogenous, so may be correlated with u_1, and subsequently, e. We assume the noise term in the $ICUADM$ model, u_3, is vector of independent and identically distributed zero mean standard normal random variables, so that the model is consistent with the ICU admission model estimated in Kim et al. (2015). Additionally, we assume that $(Z_2, EDBOARD)$ and $(Z_3, ICUADM)$ are always

observed and $ICULOS$ is observed only when $ICUADM = 1$. Note that we can observe $EDBOARD$ for all patients even if they are not admitted to the ICU. As such, our model satisfies Assumption 17.2 in Wooldridge (2002). Moreover, by Theorem 17.1 in Wooldridge (2002), a two-stage least-squares approach will result in consistent estimates for δ under sample selection with the Inverse Mills Ratio included as an exogenous regressor.

In order to estimate δ, we follow Procedure 17.2 in Wooldridge (2002). We first estimate the selection and ED boarding models. We then use the estimates for the IMR ($\hat{\lambda}(Z_3\beta_3)$) and ED boarding time ($\hat{EDBOARD}(Z_2\beta_2)$) to estimate the ICU LOS model. What remains is to calculate the correct standard errors. As indicated in Wansbeek and Meijer (2000), such a multi-stage regression approach will lead to consistent estimates and a GMM approach can be used to calculate standard errors. This is the approach taken in Allon et al. (2013), Meijer and Wansbeek (2007). What differentiates our model is that we actually have observations of the endogenous regressor (ED boarding times) even when the patient is not admitted to the ICU. As such, we can utilize these observations to increase the estimation power. Due to the large sample size of patients *not* selected for ICU admission, performing the matrix inversion necessary to estimate the standard errors via GMM is numerically challenging. As such, we use non-parametric bootstrapping with replacement over 1000 samples to do so (Wooldridge 2002, Schwiebert 2012). We do this separately for each of the 10 patient categories.

Z_1, Z_2, Z_3 all contain physiologic and operational factors which are available for all patients, such as severity, age, day of admission, and hospital where care is received. Because $EDBOARD$ is potentially endogenous, Z_2 also includes an instrument which influences $ICULOS$ only through its relationship to $EDBOARD$. Similar to Kim et al. (2015), we use the congestion of the patient's first inpatient unit (i.e. ICU congestion for ICU patients and ward congestion for the general ward patients) as an instrument. In this case, we consider the average hourly occupancy during the time the patient is boarding in the ED. We define the next unit as 'busy' if the occupancy level is greater than 80% of the maximum patient census over the course of the year. This binary measure of ICU congestion is similar to the approaches taken in McClellan et al. (1994), Kc and Terwiesch (2012) and Kim et al. (2015) among others. Note that we examined other measures of busy including different thresholds and times at which the occupancy was measured and found similar results. Finally, we include the congestion of the ICU and nonICU units at the time of inpatient unit admission into Z_3. We note that while these various measures of congestion are related, their correlation is typically around 20% and no more than 50%.

2.3. Empirical Results

We first consider the impact of busy inpatient units on ED Boarding. Note that we look at the congestion in the next unit the patient visits, so for ICU patients, we consider the congestion in the ICU, while for non-ICU patients, we consider congestion in the non-ICU units. With $p < .001$, ED Boarding time increases by

1.8-2.5 hours when the occupancy level of the next impatient unit is greater than 80%. This result supports our intuition that inpatient unit congestion increases boarding time. Consistent with Kim et al. (2015), we find that when the ICU is busy, the likelihood of ICU admission decreases.

We now consider the impact of ED Boarding on ICU LOS. As a measure of model robustness, we consider two models: the first does not use any instrumental variables and the second uses the Heckman approach with endogenous regressor as discussed earlier. Table 2 summarizes the delay effects for the 10 primary condition categories of interest. We also provide the coefficient, γ, on the Inverse Mills ratio. Statistically significant results of γ suggests evidence of selection bias in some patient categories. We see evidence of estimation bias, especially in the case of Renal patients, where using traditional Ordinary Least Squares suggests that increased boarding time actually reduces ICU LOS. This goes against medical knowledge and intuition. We can see that our approach is able to adjust for this bias. When we control for the endogeneity of ED boarding and the potential selection bias, all statistically significant coefficients in this case are positive.

We can see that for patient categories: Fluid & Hematologic, Renal, and Vascular the delay effect is statistically significant ($p < .10$). For these ailments, 1 additional hour in ED delay is associated with a 11.37%-38.21% increase in ICU LOS. As we will see in our analysis of queueing systems with delay-dependent service times, this impact can be substantial.

The regressions for Cancer would not converge, so we could not achieve coefficient estimates. We also do not see any statistically significant results for patient conditions Cardiac, Catastrophic, Infectious, Metabolic, Respiratory, and Skeletal. Cancer and Skeletal are the patient conditions with the fewest number of ICU patients, so the lack of statistically significant results may be attributed to the small sample size. There are 650 samples of Metabolic patients, yet it seems that delays may have little impact on ICU LOS. This may be because Metabolic corresponds to chronic conditions including diabetes, immune disorders, end stage renal disease, etc. Subsequently, these patients may be more delay tolerant. While the patients are considered severe (they still need ICU care), there is likely to be less urgency when the patient's primary condition for admission is chronic. Finally, Skeletal refers to conditions such as broken hips, which may be susceptible to infection if left untreated; however, their urgency is likely to be lower than other patients such as those who had a stroke (Vascular).

2.3.1. Robustness Checks and Discussion While Table 2 presents our main empirical results, we also performed a number of additional regressions to test the robustness of our results.

First, we considered various measures of busy-ness of the ICU and non-ICU units. We considered thresholds of 75% and 85% occupancy, as well as linear and quadratic specifications. The main insights from our results did not vary drastically with these different specifications.

	(i) OLS		(ii) Model (8)		
	δ	R^2	δ	γ	R^2
Cancer	0.0775	1.0000	.	.	.
	(.)		(.)	(.)	
Cardiac	-0.0040	0.1638	0.0007	0.4618***	0.1683
	(0.0049)		(0.0340)	(0.1727)	
Catastrophic	-0.0137	0.1850	0.0427	-0.5002	0.1847
	(0.0093)		(0.0389)	(0.4365)	
Fluid&Hematologic	-0.0101	0.3047	0.3237**	0.2800	0.3439
	(0.0150)		(0.1325)	(0.4937)	
Infectious	-0.0117	0.1280	0.0259	1.0398***	0.1414
	(0.0078)		(0.0444)	(0.3997)	
Metabolic	-0.0260**	0.1580	0.0332	0.0088	0.1565
	(0.0110)		(0.0450)	(0.3165)	
Renal	-0.0736***	0.5224	0.3821***	-2.6919**	0.5402
	(0.0223)		(0.1334)	(1.0964)	
Respiratory	-0.0063	0.1312	0.0079	0.3458	0.1457
	(0.0099)		(0.0576)	(0.4737)	
Skeletal	0.0110	0.3346	0.0158	0.9692	0.3494
	(0.0328)		(0.2560)	(1.1587)	
Vascular	-0.0296**	0.2457	0.1137*	1.0365	0.2559
	(0.0133)		(0.0651)	(0.6440)	

Standard errors in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$

Table 2 log(ICU LOS) regression results: (i) ordinary least squares without instrumental variables; (ii) uses ICU Occupancy $> 80\%$ at ICU admission time as an instrumental variable.

We note that prior work has demonstrated that when the ICU is busy, patient LOS may decrease (Kc and Terwiesch 2012). In their work, they focus on a single cardiac ICU where patients are cared for following cardiac surgery. In our case, we do not consider surgical patients. We focus on ED medical patients. Kim et al. (2015) shows that scheduled surgical patients are most likely to experience speedup when the ICU is busy, while ED medical patients do not seem to experience speedup when the ICU becomes congested. Our data is consistent with these findings.

A number of works have shown that congestion during an ICU visit can result in worse outcomes (e.g. Chalfin (2005), Kc and Terwiesch (2012)), so we also included a measure of ICU congestion in Z_1[3]. In particular, we consider the average hourly ICU congestion during a patient's ICU stay. This is similar to the

[3] This has also been demonstrated in non-ICU settings (Kuntz et al. 2014).

approach in Kim et al. (2015), which considered the average daily congestion. When excluding this measure or using congestion in the first 24 or 48 hours of ICU admission, we find the delay effect still exists, though the statistical significance is sometimes weaker.

Note that our estimation approach relies on a number of assumptions and if any of these are violated, it raises questions to the reliability of our results. For instance, we assume the noise term in the selection model is normally distributed. We also assume a specific parametric relationship between this noise term and that of the $\log(ICULOS)$ model, which is stated in Eqn. (5). If either of these assumptions do not hold, it could invalidate our results. We use congestion in various inpatient units as instrumental variables. While we tested and found that these measures of congestion are uncorrelated with *observable* measures of severity, it is impossible to check this with respect to *unobservable* measures of severity. If this were not true, it would invalidate the IV estimation approach. Despite these caveats, we find substantial evidence that, for a large group of patients, delays in ICU admission are associated with increases in ICU LOS. As expected from the medical literature, the impact of delays varies across different patient conditions. We next devote our attention to understanding the implications of this delay effect on traditional queueing insights. While we notice the delay effect can vary across different types of patients, our models will focus on a single class system in order to develop focused insight into the delay effect.

3. Incorporating the Delay Effect: M/M(f)/s Model

Motivated by our empirical analysis, we turn our attention to developing queueing models which incorporate the delay effect. Such analysis allows one to measure the impact of ignoring the delay effect when using conventional queueing approaches. To do this, we introduce an $M/M/s$-like queueing system which has jobs with delay-dependent service times. Our analysis assumes a single patient class in order to focus on the impact of the delay effect. Such an assumption is reasonable in hospitals with specialized ICUs. For instance, some large hospitals have dedicated cardiac ICUs where non-surgical cardiac patients are given priority. At a higher level, this modeling assumption, which is necessary to allow for analysis, also provides a first step in understanding the effect of delay-dependent service times on queueing phenomena.

We consider a model with Poisson arrivals and exponential service times. However, the service rate of the standard exponential random variable now depends on the delay of the job; we denote this dependence by $M(f)$ where f is an 'inflation' function that we will define shortly. Hence, the service time (equivalently, LOS) of a job is inflated from some nominal value by a quantity which depends on the number of jobs in the queue upon the job's arrival. Such a model is able to capture the dynamics estimated from the patient data in the previous section.

We now formally introduce our delay-dependent queueing system. Consider an s server queueing system described as follows: Jobs arrive according to a Poisson process at rate λ and are served in First-Come-First-Served (FCFS) fashion. We let N_t denote the number of jobs in the system at time t. Job i arrives at

time t_i and its service time is exponentially distributed with mean $1 + f\left(N_{t_i^-}\right)$ where $f(\cdot)$ is a function, referred to as a *growth function*, which takes values in a finite set and satisfies the following requirements:

1. $f(m) = 0$ for $m = 0$.
2. $f(\cdot)$ is bounded and non-decreasing.

In what follows, we will examine the behavior of this system and the impact of the growth function, $f(m)$. We will refer to such a system as a queueing system with delay dependent workload, and abbreviate it with the notation $M/M(f)/s$.

Remark 1 *Note that service times depend on the number of jobs in the system upon arrival, rather than the realized waiting time of the job. This is primarily for tractability and we find that even with this assumption, analysis of the model is still not straight forward. That said, we will see in Section 6, this simplification does not alter the insights substantially, as N_t is a very good proxy for wait times.*

3.1. Stability of an $M/M(f)/s$ System

We first begin our analysis of our queueing system with delay-dependent service times by considering the stability for such a system. While the stability condition, and consequently the throughput of an $M/M(f)/s$ system, is a relatively coarse performance benchmark, it provides interesting insight into the behavior of such systems. We have that:

Proposition 1 *An $M/M(f)/s$ system is weakly stable, i.e.*

$$\lim_{t\to\infty} \frac{W_t}{t} = 0$$

if and only if

$$\frac{\lambda}{s} \leq \frac{1}{1 + f_{\sup}}$$

where W_t is the work in the system at time t and $f_{\sup} = \sup_m f(m)$ is the supremum of f.

The proof of this result can be found in the appendix. To provide some intuition of this result, if a burst of jobs arrive, they will all experience some delay and an increase in service requirement. If a particularly bad burst of jobs arrive in sequence, the system will quickly deteriorate to the point where all jobs are delayed and require maximal service time. Hence, the stability requirement is based on the maximum possible job requirement. We see that short term behavior (bursts) can have lasting effects which impact long-run average behavior (stability). We note that these dynamics highlight the challenges associated with analyzing such a system. We can see a complex correlation structure between the service times and interarrival times arises. Periodic and evenly spaced interarrival times corresponds to no service requirement inflation; however, bursts and short interarrival times will lead to large service times. While the question of stability reduces

to the standard stability characterization under the worse-case scenario of all jobs inflating maximally, the system dynamics are more nuanced.

Our stability analysis demonstrates how substantial the delay effect can be. We consider a simple example to further illustrate this effect. Consider a system with a daily arrival rate of 10 Renal patients whose mean LOS is 60.67 hours. Our empirical analysis suggests that a single hour of delay would result in a 38.21% increase in LOS (see Table 2). In order to maintain stability with a maximal 38.21% increase in LOS, the system would need at least 35 beds, while only 26 beds are needed in a system without delay effects. Stability is the most basic service requirement. If one wished to ensure high quality service one could use traditional $M/M/s$ analysis to verify that an ICU with 33 beds would guarantee that no more than 5% of patients would have admission delays of more than 6 hours[4]. However, even if a 38.21% increase in LOS were the *maximum* increase in LOS, this ICU would be unstable, resulting in extraordinarily poor system performance. Thus, the system without delay effect would provide a high service quality guarantee, while the system with delay effect would be unstable. Using a simple simulation model which increases LOS by 38.21% if a patient must wait, we see that 37 beds (4 more) would be needed to ensure the same high service quality guarantee. Indeed, ignoring the delay effect can result in poor capacity management decisions.

In some instances, our delay-dependent queueing system can be represented as a multi-dimensional Markov Chain, which we formally discuss in Appendix C. The transition matrix for this Markov Chain has a block diagonal structure. However, the size of the blocks can be arbitrarily large depending on the nature of the function f. While one may be able to solve for the steady-state dynamics numerically for special cases, it does not provide much insight for the general model. Moreover, this approach quickly becomes intractable with general f functions. Despite starting from the innocuous $M/M/s$ queueing model, the introduction of the delay effect makes the resulting system far too difficult to permit an exact analysis. As such we focus on producing approximations by constructing suitable upper bounding systems. This analysis provides some insight into how the issues above might impact nominal predictions that do not account for the impact of delay on service time.

4. Approximating the Workload Process

This section will be concerned with establishing a simple approximation to the long run average workload of an $M/M(f)/s$ system. We focus on the workload as it is a common accounting metric used in the hospital setting[5]. It can also be used as a surrogate for delays. Finally, from a technical point of view, we are able to establish tractable bounds for the workload.

Let us denote by W_t and N_t respectively, the workload and number in system processes in this system, where we define the workload as the total amount of work in the system based on realized service times.

[4] As mentioned in Chalfin et al. (2007), delays of more than 6 hours are associated with worse outcomes.

[5] The number of patient days specifies how many days patients, in aggregate, stayed in a hospital or unit.

Consider also, an $M/M/s$ system with arrival rate λ and service rate $\frac{1}{1+f_{\sup}}$ where $f_{\sup} = \sup_m f(m)$. Assume the service discipline for this system is FCFS. We denote by $\overline{W}_t$ and $\overline{N}_t$ respectively, the workload and number in system processes in this system. We will frequently refer to the former system (the system we are interested in analyzing) as system 1 and the latter system (which will have value in our producing bounds) as system 2. Finally, we denote by $\underline{W}_t$, the workload process in an $M/M/s$ system with arrival rate λ and service rate 1, i.e. a system *without* any delay-effect or relationship to the growth function $f(m)$. We will refer to this system as the baseline, delay-independent system and use its behavior as a comparison benchmark for our $M/M(f)/s$ system and the corresponding bound we will establish. We let $E[W], E[\overline{W}]$, and $E[\underline{W}]$ denote the expected work in each system. That is, if we start the systems according to their respective stationary distributions, then these correspond to the expected work in each system at time 0: $E[W] = E[W_0]$, $E[\overline{W}] = E[\overline{W}_0]$, and $E[\underline{W}] = E[\underline{W}_0]$

4.1. An Upper bound for a Step Function

In order to provide more insight into the bound we will derive, we start by examining a special case of the delay-growth function, f. In particular, we focus on the case where jobs have nominal service requirement of mean 1 which increases to $1 + k$ if there are N^* or more jobs in the system upon arrival:

$$f(m) = \begin{cases} 0, \ m < N^*; \\ k, \ m \geq N^*. \end{cases}$$

Such a delay growth function captures the increased service time required by jobs (patients) who arrive to a congested system (i.e., $m \geq N^*$). Such a growth function bears similarities to some of the medical literature which examines the increase in workload of delayed versus not delayed patients (Chalfin et al. 2007, Renaud et al. 2009). Moreover, we consider the case where the service times are exponentially distributed. We can establish the following upper bound:

Theorem 1 *Assume that $f(\cdot)$ is defined according to $f(m) = k$ for $m \geq N^*$, and $f(m) = 0$ otherwise. We have that the expected workload, $E[W]$, satisfies*

$$E[W] \leq E[\overline{W}] - \lambda(2k + k^2)P(\overline{N} < N^*)$$

where $\overline{W}$ and $\overline{N}$ denote the workload and number of jobs in a traditional $M/M/s$ system with arrival rate λ and service rate $1/(1 + k)$.

The upper bound consists of the amount of work in the system if *all* jobs were inflated, which is then corrected according to the second term in the bound. In particular, when considering the expected workload in the system, we can look at the total aggregate work in the system up to time t (i.e. integrate $\int_0^t W_t dt$) and divide by t. With this approach, we start by looking at the contribution of an individual job to the integral component. To provide some intuition of the correction term, let's consider the case where $N^* = s$ and

examine the amount of work contributed by an arbitrary job, i. We let σ_i be the realized service requirement for job i in our $M/M(f)/s$ system and $\overline{\sigma}_i$ is the amount of work the ith job brings in system 2, where all jobs are inflated to expected service time of $1+k$. We note that we correct for the extra amount of work that is introduced whenever a job does not have to wait upon arrival, i.e. $\overline{N}_{t_i^-} < s$. A job that immediately begins service contributes a total of $\frac{1}{2}\sigma_i^2$ work, i.e. it brings work σ_i that is depleted at constant rate 1 until it completes service. The total contribution is then the area of the right triangle with width and height equal to σ_i. But this job does not have to wait, so the amount of work that is actually contributed is $\frac{\overline{\sigma}_i^2}{2(1+k)^2}$, which accounts for the artificial inflation of the work to expected size $1+k$. Therefore, to account for the actual amount of work introduced by a job who does not have to wait, we subtract the amount of work contributed by the inflated job $\frac{1}{2}\overline{\sigma}_i^2$ and add the amount of work by the correct mean 1 sized job: $\frac{\overline{\sigma}_i^2}{2(1+k)^2}$. See Figure 1 for an illustration of accounting to correct for the excess work introduced. Recognizing that the second moment of an exponential random variable with mean μ is $2\mu^2$, we derive the desired result.

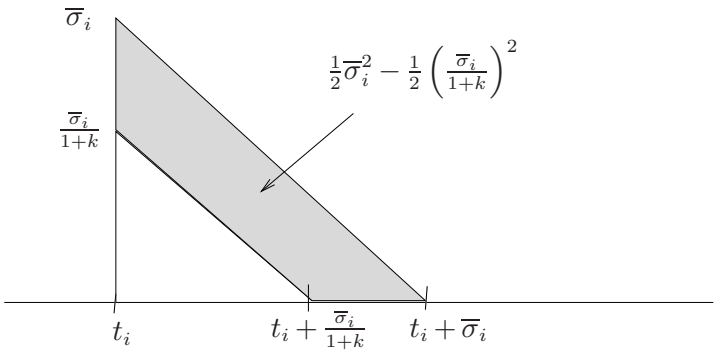


Figure 1 **Due to the inflation of all jobs, each job which experiences zero delay contributes excess work which is shaded in gray.**

Note, that for $k=0$, our bound is tight for a queueing system without delay-dependent service times: the upper bound is equivalent to the classical results of an $M/M/s$ queue. Additionally, the bound is tight as $\rho \to 0$ and $\rho \to 1$, with the $M/M(f)/s$ system reverting back to an $M/M/s$ queue with $\mu = 1$ or $\mu = 1/(1+k)$, respectively. The first expression in the upper bound corresponds to a system where *all* jobs have their service time increased, irrespective of the amount of delay experienced. However, the workload does not unilaterally increase with the load. The second part of the expression represents the correction for over inflating the workload for jobs which do not experience excess congestion. We note that this is an upper bounding system because, while we account for the correct workload if a job is not delayed, we do not correct for the propagation effect of its inflated workload on delays for future jobs. Still, as we will see later, the upper bound is quite accurate for systems with various growth factors, k, and number of servers, s.

We observe that the upper bound in Theorem 1 admits a simple analytical expression. This allows us to generate a clean understanding of the impact of delay on the workload process akin to our understanding of the role factors such as utilization play in a traditional $M/M/s$ system, which we will explore in Section 5.

4.2. A General Upper Bound for an $M/M(f)/s$ System

As we saw in Section 2, the delay effect can be gradual. Thus, we now generalize our result from Theorem 1 to other delay-growth functions. Consider any growth function $f(\cdot)$ with a countable number of discontinuities. Let $0 = M_0 < M_1 < M_2 < \cdots < M_{J-1} < M_J = \infty$ be break points in the function f, so that if the number of jobs in the system upon arrival of a new job satisfies $M_j \leq \overline{N} < M_{j+1}$, the service rate of that job is $1/(1+k_j)$, where $k_j \leq k_{j+1}$. Hence,

$$f(m) = k_j, \text{ if } M_j \leq m < M_{j+1}$$

Thus, f is an arbitrary non-decreasing piece-wise constant function. Most non-decreasing functions can be reasonably approximated within the framework of piece-wise constant functions.

As we have described before, W and N are defined as the workload and jobs processes for this delay-dependent queueing system. Similarly, let $\overline{W}$ and $\overline{N}$ be the workload and jobs processes for an $M/M/s$ system with arrival rate λ and service rate $1/(1+k_J)$, where $k_J = \max_j k_j$. We can then establish the following upper bound to our $M/M(f)/s$ system:

Theorem 2 *If f is a non-decreasing piece-wise constant function with $f(m) = k_j$ if $M_j \leq m < M_{j+1}$, we have that the workload process, W, satisfies*

$$E[W] \leq E[\overline{W}] - \sum_{j=0}^{J} \left[\lambda(2k_J + k_J^2 - 2k_j - k_j^2) P(M_j \leq \overline{N} < M_{j+1})\right]$$

The proof of this result requires a coupling argument and can be found in Appendix D. To provide some insight into the interpretation of this bound, we parse through the two expressions which compose the upper bound:

1. The first term corresponds to the expected work in the system if all job are inflated maximally to mean service time $1 + k_J = 1 + f_{\sup}$. Thus, it corresponds to the expected work in an $M/M/s$ system with $\rho = \lambda(1+k_J)/s$. However, most jobs will not be inflated to the maximum size, which brings us to the second term.

2. The second term corresponds to the correction necessary for overinflating the workload of jobs with moderate or no wait upon arrival. If this occurs, the work that each new job brings is a factor of $\frac{1+k_j}{1+k_J}$ less than the amount of work that arrives in the $\overline{W}$ system. Removing this extra work results in the multiplier of the last expression.

Note that the only time we rely on the exponential service times is to make the algebraic simplification in Proposition 6 to establish the closed form expression for the correction term. Hence, the bound can be extended to general service times, but may not result in as clean expressions.

Remark 2 *A lower bound for the expected workload of an $M/M(f)/s$ system can be derived using a similar approach to that used for the upper bound here. However, we find that such a lower bound is very loose. One of the biggest issues with the delay effect is the negative externalities on other jobs when a job is delayed and its service requirement increased. In our upper bounding system, our correction factor does not correct for the propagation effect. However, a corresponding lower bounding system does not include the propagation effect. We find that this propagation of delays is a primary driver in the dynamics of an $M/M(f)/s$ system; thus, the lower bound is not very accurate.*

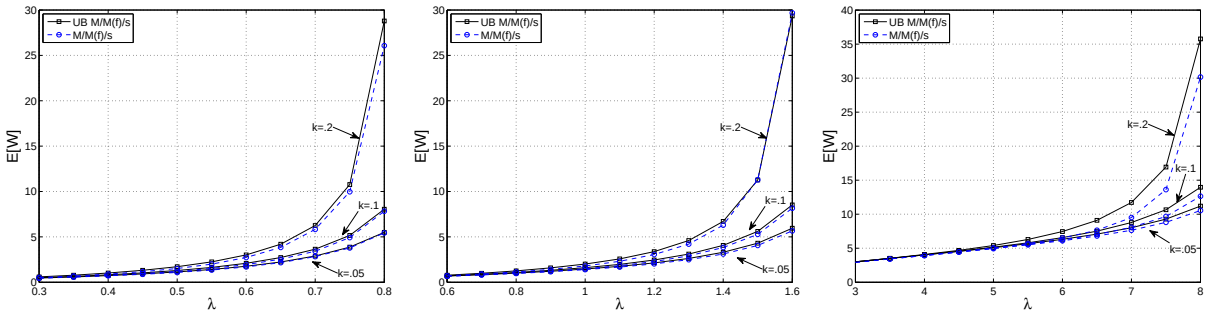
5. Sensitivity Analysis of the Bound

Traditional queueing models have been used to guide operational decisions in healthcare, such as staffing levels and numbers of beds (e.g. McManus et al. (2004), Green (2006), Yankovic and Green (2011), de Véricourt and Jennings (2011) among many others). None of these models account for the delay effect. We now consider how the upper bound derived in Section 4 can be used to gain a better understanding of a queueing system with delay dependent service times. We will see that the system behavior is very different under the presence of a delay effect and this can result in potentially very different operational decisions. As such, it is important to account for the effect appropriately. To see this, we focus on the result of Theorem 1, where the growth is represented by a step-function.

We start by examining the behavior of the workload as the magnitude of the delay effect increases in Figure 2. The first aspect to notice is the accuracy of the derived upper bound in comparison to the simulated workload of the $M/M(f)/s$ system. This allows us to utilize our upper bound to derive more insights into the behavior of our delay dependent system.

What is most striking is the change in workload as the system utilization increases (recall that service times have been normalized to 1). It is well known that traditional $M/M/s$ systems are sensitive to changes in arrival rate λ and the number of servers s, so that as $\rho \to 1$, the workload increases rapidly according to the relationship $1/(1-\rho)$. The introduction of the delay effect results in much more rapid increases. That said, the bound is looser as the number of servers increases. Because of pooling, the impact of the system load decreases as the number of servers increases. We explore these phenomena more precisely via explicit evaluation of our bound.

For any number of servers, s, it is possible to compose exact expressions for our upper bound when the growth function is a step function as in Theorem 1. To demonstrate this process, we now explicitly evaluate our bound in two cases: a single server and two servers. While such a small system may not be generally



(a) 1 server (b) 2 servers (c) 10 servers

Figure 2 **Comparison of expected workload in a simulated $M/M(f)/s$ system versus the derived upper bound for $s = 1, 2,$ and 10. Inflation is given by a step function: $f(m) = k1_{\{m \geq s\}}$ with $k = .05, .1,$ and $.2$.**

applicable to an ICU setting, there are specialized ICUs which can be very small. For instance, in California, the smallest number of licensed Medical/Surgical ICU beds amongst hospitals with such an ICU is 2 and three hospitals have a 3 bed ICU (State of California Office of Statewide Health Planning & Development 2010-2011). More generally, there are other service settings which include a delay effect and have few servers. For instance, Primary Care may be one such setting (though the delay effect is likely much smaller than in the ED to ICU setting which we are considering here). In our evaluation of explicit expressions, we consider $N^* = s$, so that the workload increases for any job which is delayed.

The Single Server Case $M/M(f)/1$: We want to compare the behavior of the $M/M(f)/1$ system to a regular $M/M/1$ system which does not have any delay effect. We denote the workload in an $M/M/1$ system with arrival rate λ and service rate 1 as $\underline{W}$ and note that $E[\underline{W}] = \frac{\rho}{1-\rho}$ for $\rho = \lambda$. We denote by W^{UB} the upper bound derived in Theorem 1. Using this upper bound in conjunction with a Taylor series approximation, we have that

$$\frac{W^{UB}}{E[\underline{W}]} = \frac{1-\rho}{1-(1+k)\rho} \approx 1 + E[\underline{W}]k$$

so that the workload in our $M/M(f)/1$ system grows quadratically with the expected work in a traditional $M/M/1$ system. Considering that the work grows according to $1/(1-\rho)$ for a traditional $M/M/1$ system, we see that in our new system with delay dependent service times, the work will grow much more rapidly with ρ (i.e., the additive term grows like $1/(1-\rho)^2$).

The Two Server Case $M/M(f)/2$: We now consider a similar analysis to the single server case when there are two servers. Because there are two servers, we now define the system load $\rho = \lambda/2$ and maintain this definition in what follows. For our $M/M(f)/2$ system, we have:

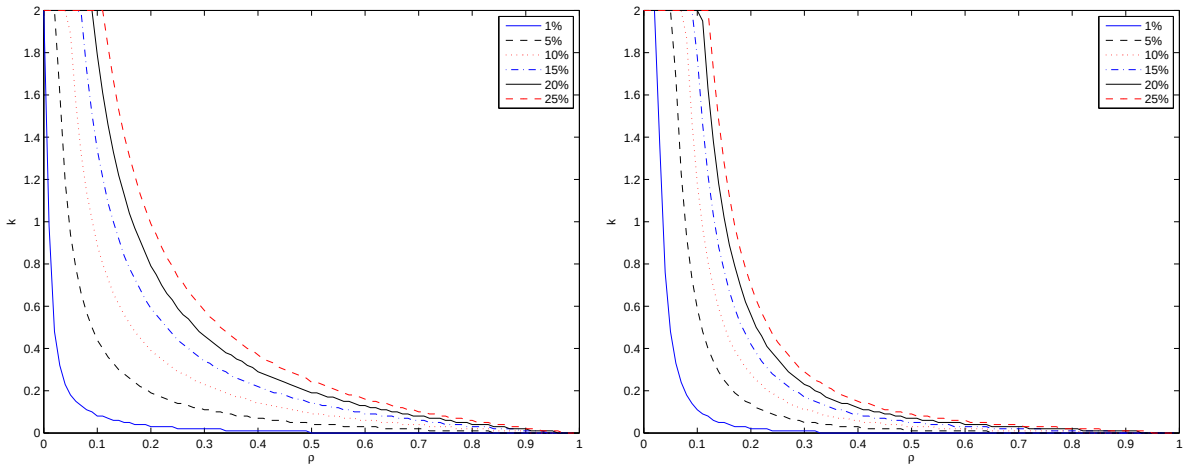
$$\begin{aligned}\frac{W^{UB}}{E[\underline{W}]} &= \frac{(1+k)^2(1-\rho^2)}{1-(1+k)^2\rho^2} - \frac{(2k+k^2)(1+2(1+k)\rho)(1-(1+k)\rho)(1-\rho^2)}{1+(1+k)\rho} \\ &\approx 1 + E[\underline{W}]\rho k + 6\rho^2 k + 4\rho^3 k\end{aligned}$$

around $k = 0$. Similar to the single server case, we see the delay effect introduces a quadratic term in $E[\underline{W}]$, the expected work of a traditional $M/M/2$ system. Thus, we again see a much more rapid growth according to $1/(1-\rho)^2$ rather than the traditional $1/(1-\rho)$ relationship.

Many traditional queueing models are used for performance evaluation and capacity management in healthcare settings (e.g. McManus et al. (2004)). However, the results here suggest that using such traditional tools could lead to substantial underestimates of the true delay. This is most pronounced when ρ and k are large; ignoring the delay effect can result in estimates of performance which are orders of magnitude too low. For instance, when $k = .3$ and $\rho = .95$, our upper bound estimates the expected work in the system to be a factor of 9 greater than the workload in a standard $M/M/2$ system. As such, it is important for hospital managers to be acutely aware that the delay effect can cause delays to be much worse than originally anticipated and take this into careful consideration when making capacity management decisions.

Of course, there are some instances where ignoring the delay effect will result in reasonably accurate estimates of the underlying system. At a high level, this will occur when the system has 1) low utilization and/or 2) a small delay effect. Suppose we are willing to accept underestimates of the expected work in system, there exist scenarios in which the estimates generated from the standard $M/M/s$ without delay-dependent service times are sufficiently close to the expected work in the underlying $M/M(f)/s$ system. Figure 3 demonstrates the k and ρ values below which the $M/M/s$ workload is sufficient for various percentage tolerances. One can then use this analysis to determine whether accounting for the delay effect when using queueing models to inform managerial decisions is necessary. For example, as long as $k < .2$ and $\rho < .3$, the estimates for $E[W]$ given by a standard $M/M/1$ model are likely to be within 10% of $E[W]$ in the delay-dependent system. Similarly, for a two server system, as long as $k < .4$ and $\rho < .15$, using an $M/M/2$ will be within 10% of the workload in the $M/M(f)/2$ system. Unfortunately, in most hospital settings–and especially the ICU–system loads tend to be between 65-90%. In such cases, the impact of delays cannot increase LOS by more than 4-10% in order for the resulting delays to be within 25% of the estimated delays from a traditional $M/M/s$ model. Thus, for some conditions, such as Skeletal, it might be reasonable to use traditional $M/M/s$ models. However, for other conditions, such as Fluid & Hematologic, using a traditional $M/M/s$ to model the system is likely to be quite inaccurate.

We can see that this bound is an important first step in understanding and characterizing the impact of the delay-dependent service times on the expected workload in the system. We find that the delay effect results in a dependency of the workload on the system load which grows much more rapidly than $1/(1-\rho)$. A direct takeaway from this observation is that reducing utilization–by reducing the arrival rate or by adding more servers–will have a highly non-linear impact on decreasing workload, and subsequently, delays. Note that in light of the delay effect, this reduction will be much more effective than the same reduction in a traditional $M/M/s$ system. As such, it is important for hospital managers to use caution when using traditional queueing models to inform capacity decisions, such as how many beds to staff in the ICU, as



(a) A single server system (b) A two server system

Figure 3 Numeric evaluation of when $W^{UB}/E[\underline{W}] < 1 + p\%$ where $p\%$ is the percentage tolerance a manager may be willing to accept.

delays are likely to be much worse than anticipated. Furthermore, because the $1/(1-\rho)^2$-type relationship becomes more pronounced with large delay-effects (as measured by k), it is crucial to understand the delay effect for the patient population of interest. The model we present in this work is an important first step at understanding how the effect of delays on service time can impact a queueing system. When making operational decisions, it is important for hospital managers to account for the delay effect, either via models such as the one presented here, simulation models, or other approaches. Certainly, ignoring the delay effect is likely to result in poor operational decisions.

6. Numerical Comparisons

We further examine the behavior of our delay-dependent queueing system along with the quality of the derived upper bound. In particular, we wish to examine how this delay effect may impact a real system. To do this we connect back to our empirical analysis in Section 2 to calibrate our model. We consider a setting with a fixed number of servers (beds). If a job (patient) arrives and there is an available server, it is immediately served. If there is no available server, he must wait. His expected service time is non-decreasing in the amount of time he must wait. We consider the expected workload in the systems.

Specifically, we simulate the behavior of these delay-dependent queueing systems for a small (6 beds) and moderately sized (15 beds) ICU. We compare the expected workload to three benchmarks:

1. **[$\mathbf{M/M/s}$ with $\rho = \lambda$]** This represents a traditional queueing system without delay effects. This is a (trivial) lower bound to the delay-dependent system.

2. **[$\mathbf{M/M/s}$ with $\rho = \lambda(\mathbf{1} + \mathbf{k})$]** This represents a queueing system where the amount of work each job brings is artificially inflated as if *all* jobs experienced delays. This is a (trivial) upper bound to the delay-dependent system.

3. **[Upper bound derived in Theorem 2]** This corrects for the miscalculation of work for jobs who are not delayed.

6.1. A Single Step Growth Function

To start, we consider a model where patients have a nominal ICU LOS. If a new patient is delayed admission, his LOS increases by a constant factor k (we consider alternative delay functions later). That is, our first experiments involve a delay function $f(m) = k$ if $m \geq s$ and 0 otherwise. We need to determine the value of k. To do so, we turn back to our empirical analysis in Section 2. Recall that we found evidence that patient delays (ED Boarding) are associated with longer ICU LOS. In order to capture this effect in our simulations, we account for an increase in service time whenever a job is delayed.

We focus on Vascular and Renal patients. We selected these condition categories because they have the lowest and highest statistically significant increase in ICU LOS when delayed. From Table 1, the mean ICU LOS of Vascular and Renal patients is 60.67 and 69.70 hours, respectively. From Table 2, each hour of boarding is associated with a 11.37% or 38.21% increase in ICU LOS.

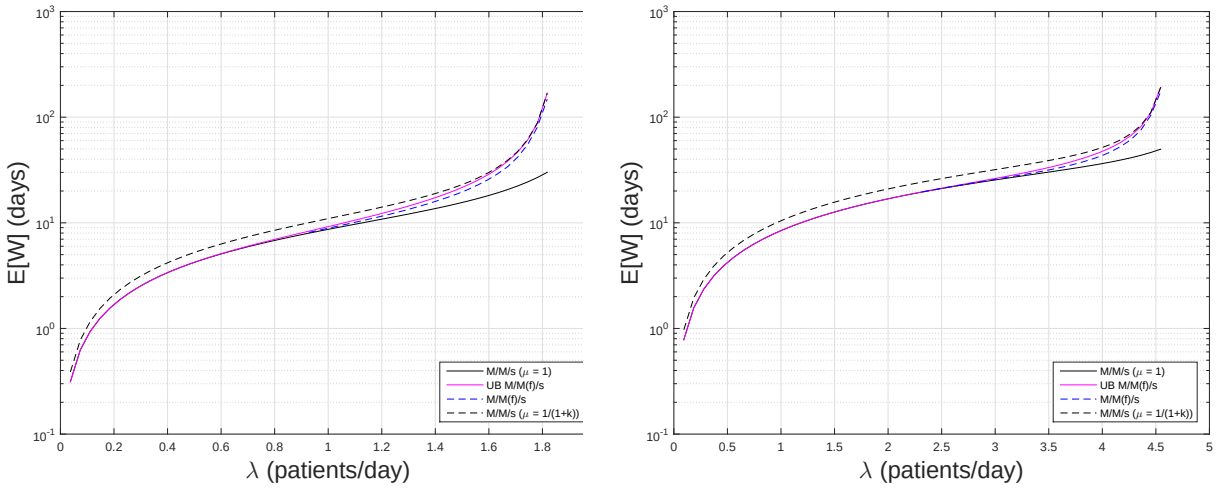
In our empirical analysis we estimated a log-linear growth function, which we will be approximating with a step function. While we consider multi-step functions later, simple simulation experiments suggest that the step function is quite accurate for low to moderate loads–as the system becomes more heavily loaded, more using additional steps will be more accurate. First, we consider Vascular patients and define the growth function $f = f_1$ as:

$$f_1(m) = \begin{cases} 0.1137, & m \geq s; \\ 0, & \text{otherwise.} \end{cases}$$

since the nominal LOS is normalized to 1, this corresponds exactly to a 11.37% increase in LOS when a patient has to wait.

Figure 4 plots the expected workload, $E[W]$, for different arrival rates. We make two observations about the delay-dependent system. First, as seen in Section 5, the upper bound is very accurate. Second, even with this very small delay effect, we can see the behavior of the system is quite different than that of an $M/M/s$ system. At low loads, the delay-dependent system looks like an $M/M/s$ system where no jobs are extended; this is because few jobs, if any are delayed. However, as the system load increases, more jobs are delayed and the delay-dependent system transitions between the $M/M/s$ system without any job growth to the $M/M/s$ system with constant job growth. It is clear that ignoring the delay effect can be misleading as to the actual work in the system.

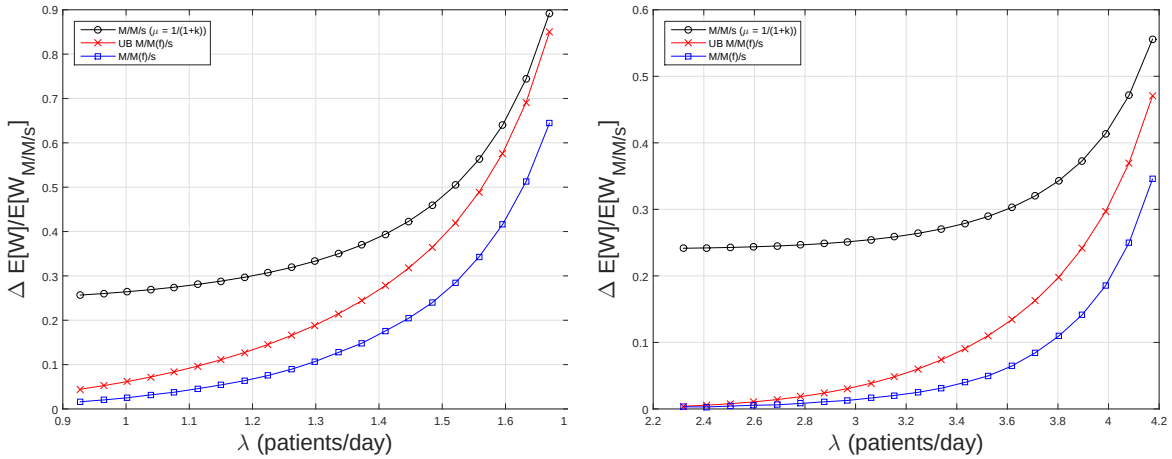
In order to get a better sense of the impact of the delay effect, in Figure 5, we examine the relative difference in the expected workload of different models compared to a traditional $M/M/s$ system where no jobs are extended, i.e. $\rho = \lambda/s$. Most ICUs are not operated in a regime where patients are rarely or always delayed, so we focus on arrival rates where at least a third of the beds turn over each day so there is some, but not excessive, congestion in our system. Again, we see that our bound is fairly accurate. Moreover, it



(a) 6 bed ICU (b) 15 bed ICU

Figure 4 **Vascular patients: Comparison of simulation of $M/M(f_1)/s$ system to the derived upper bound as well as traditional $M/M/s$ systems with *no* jobs or *all* jobs are inflated. Here the growth factor is 11.37%.**

provides more insight into the system workload than an $M/M/s$ system where all jobs are inflated. Note that an $M/M/s$ system with $\mu = 1/(1+k)$ precisely characterizes the stability condition for a delay-dependent queueing system (see Proposition 1). However, the dynamics of the workload are more nuanced.

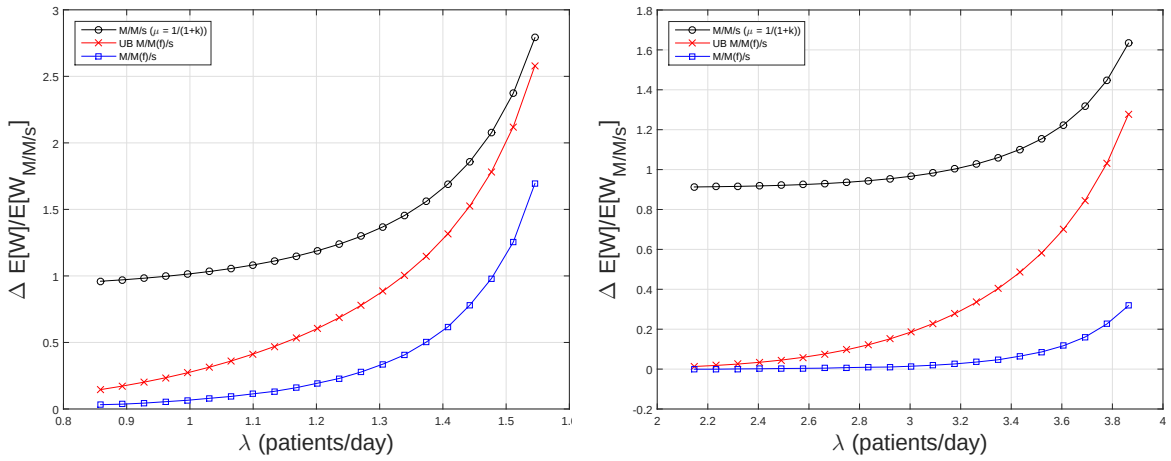


(a) 6 bed ICU (b) 15 bed ICU

Figure 5 **Vascular patients: Simulation of $M/M(f_1)/s$ system: Relative difference in workload compared to a standard $M/M/s$ system with $\rho = \frac{\lambda}{s}$. Here the growth factor is 11.37%.**

We now consider Renal patients, which have a larger delay effect. Our second growth function is: $f_2(m) = 0.3821$ if $m \geq s$ and 0 otherwise. In this case, being delayed increases a patient's ICU LOS by 38.21%. Figure 6 demonstrates the relative difference in workload in such a system. We notice that the

upper bound is looser. This is because the upper bound only corrects the work a single job brings in, but not the propagation effect it has on delaying/not delaying future jobs. This propagation is more substantial when the delay-effect is larger. Still, we can see the upper bound is a better measure of system load than the naive upper bound of an $M/M/s$ system with $\rho = \frac{\lambda(1+k)}{s}$, i.e. all jobs are extended.



(a) 6 bed ICU (b) 15 bed ICU

Figure 6 Renal patients: Simulation of $M/M(f_2)/s$ system. Relative difference in workload compared to a standard $M/M/s$ system with $\rho = \frac{\lambda}{s}$. Here the growth factor is 38.21%.

6.2. A Multi-step Growth Function

Next, we consider a multi-step growth function. To really stress the upper bound, we consider Renal patients. Thus, the maximum growth is 38.21% and we assume that the increase in expected service time is linearly increasing when there are between s and $2s$ jobs in the system. We have that

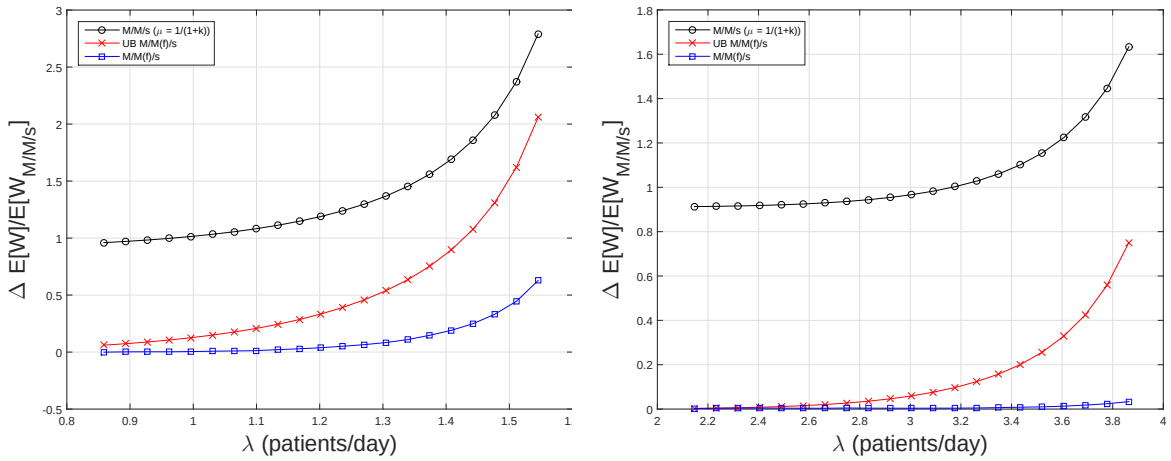
$$f_3(m) = \begin{cases} 0, & m < s; \\ \frac{.3821}{s+1}(m-s), & s \le m < 2s; \\ .3821, & m \ge 2s. \end{cases}$$

Figure 7 demonstrates the performance of the bound with such a growth function. As seen in Figure 2, the upper bound is more accurate in smaller systems. Still, even for large systems, there are regimes for λ where the upper bound is quite accurate. In all of these experiments, we see the very rapid growth of $E[W]$ versus ρ. This is particularly evident when considering Figures 5-7 are with respect to the traditional $M/M/s$ which already exhibits a $1/(1-\rho)$-type of relationship between $E[W]$ and ρ.

6.3. Realized waiting times

We now consider an alternative model to the initial $M/M(f)/s$ model introduced in Section 3. As a benchmark, we consider our original, $M/M(f)/s$ model with the following growth function:

$$f_4(m) = \begin{cases} .3821, & s \le m < 2s\,; \\ .7642 = .3821 \times 2, & 2s \le m; \\ 0, & \text{otherwise.} \end{cases}$$



(a) 6 bed ICU (b) 15 bed ICU

Figure 7 **Renal patients: Simulation of multistep $M/M(f)/s$ system: Difference in workload compared to a standard $M/M/s$ system with $\rho = \frac{\lambda}{s}$. Here the growth function is $f_3(m) = \min\{.3821, \frac{.3821}{s+1}(m-s)\} \times 1_{\{m \geq s\}}$.**

Service times increase with *realized* waiting times: Our empirical findings had that service times increased based on the wait each patient experienced, while our model uses the number of jobs in the system upon arrival. While this is certainly a proxy for the realized wait times, we also simulate a system which increases service times based on wait times. To translate f_4 into such a setting, we note that if the job is not delayed, there is no inflation of service time. If the job sees between $[s, 2s)$ jobs, its service requirement is inflated to 1.3821. In expectation, this corresponds to a wait time of less than 1[6]. If the number of jobs in the system is greater than $2s$, the expected wait (in an s-server system) will be at least 1. Hence, this system has expected service times given by:

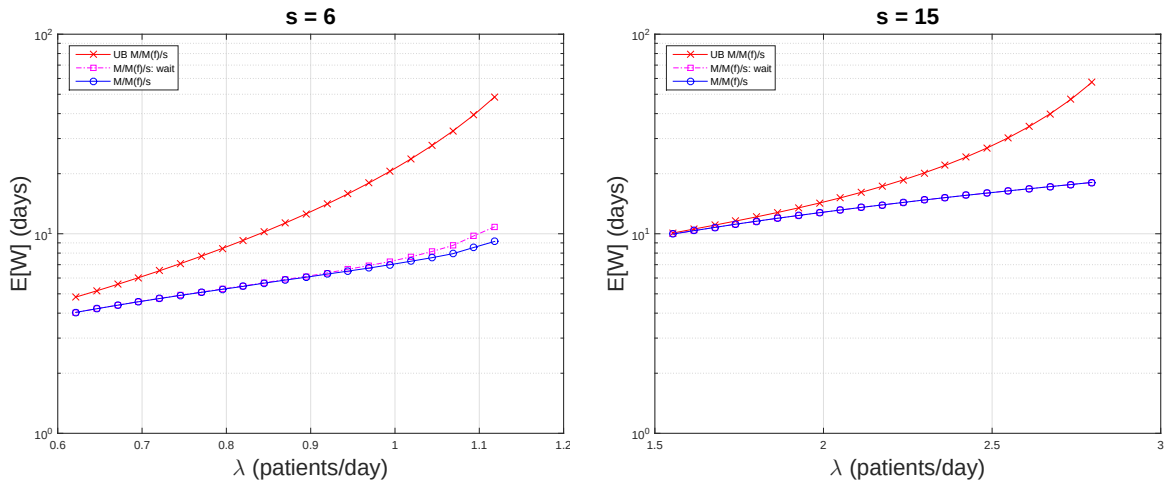
$$1 + \hat{f}(D) = \begin{cases} 1, & D = 0; \\ 1.3821, & 0 < D \leq 1; \\ 1.7642, & 1 < D. \end{cases}$$

where D is the realized delay of any job.

Figure 8 demonstrates that the model which depends on realized wait times is practically identical to our $M/M(f)/s$ system, thus our simplification still allows us to reasonably model our empirically estimated delay effect.

Through our simulations, we can see that our derived upper bound can be quite accurate. Moreover, we see that the expected workload for our $M/M(f)/s$ system is very different when comparing to a system without a delay effect. Ignoring the impact delays may have on service times may result in poor capacity management and substantial under provisioning when using traditional queueing models to guide such decisions. It is especially important to consider the delay effect when the system is heavily loaded and most

[6] Recall the service time of each job is normalized to 1



(a) 6 bed ICU (b) 15 bed ICU

Figure 8 **Simulation of $M/M(f)/s$ system with a delay effect that is wait (not number of jobs in system) dependent.**

jobs tend to experience some delay. Without accounting for the delay effect, a hospital ICU can become even more congested. In order to manage this increase in system load, hospitals may have to cancel surgeries and/or divert ambulances to reduce patient arrivals at a substantial loss in revenue. As the delay effect seems to be prevalent in a number of healthcare settings, reconsidering the management of these systems in light of delay sensitive service times may result in substantial operational and medical care improvements.

7. Conclusions and Future Directions

To summarize, this work quantifies a relatively unstudied queueing phenomenon in a critical care setting – the impact of delays on care requirements. We see that this natural phenomenon is substantially verified by data and attempt to incorporate the phenomenon into simple queueing models. The impact of this phenomenon is substantial and, as such, warrants careful attention.

In this work, we empirically estimated the impact of delays in ICU admission on ICU LOS for 10 different patient types. A number of estimation challenges arise due to the fact that patients are not randomly selected for ICU admission and that sicker patients are typically given priority, thereby lowering their waiting times, but also increasing the risk of longer LOS. Our empirical approach utilizes a Heckman selection model with endogenous regressors. Due to the large sample size of patients *not* admitted to the ICU, we utilize a bootstrapping approach to estimate our model and found that a number of patient types, but not all, demonstrate a significant effect of delay on LOS.

We then propose a stylized queueuing model which incorporates this effect. Analyzing queueing systems with delay-dependent service times exactly can be cumbersome and intractable. As such, we focus on the development of reasonable approximations for the system workload. We find that 1) our approximations are

quite accurate and 2) they provide expressions which allow for interpretations related to increases in system load. Our analysis reveals a relationship between system load and work which grows much more rapidly than the standard $1/(1-\rho)$ relationship seen in traditional queueing models. Ignoring the delay effect when using queueing models to guide operational decision making is likely to result in substantial underestimates of true delay and, simultaneously, shortages of resources such as beds, nurses, and physicians. In the ICU setting where access to timely care is crucial, it is essential that hospital managers are aware of such phenomena when considering staffing decisions. Moreover, because the delay effect can be quite substantial, especially under standard ICU loads, disregarding it may impede future attempts to make ICUs more efficient and effective. Accounting for a delay effect will result in more accurate estimates of system dynamics as well as targets for system improvement.

While we don't expect our models to directly translate into new capacity management criteria for hospital ICUs, this analysis demonstrates the impact of ignoring the delay effects when making such decisions. Ignoring the delay effects will result in ICUs continuing to be highly congested, which can lead to other reactive actions such as rerouting (Kim et al. 2015), patient speedup (Kc and Terwiesch 2012), and ambulance diversion (Allon et al. 2013), which can also be detrimental to patient outcomes. From both a patient as well as system level perspective, it is desirable to reduce delays. While reducing the average ED boarding time by an hour may be practically difficult, the adverse feedback of delays on increased service requirements suggests that, due to the faster than $1/(1-\rho)$ relationship between system load and work, even small reductions in boarding time on the order of 10 to 15 minutes may help reduce congestion. Hospital managers need to 1) take the time to characterize the extent of the delay effect within their own patient cohort and 2) to the extent that the delay effect exists, they must be careful when managing capacity of their units as delays will grow out of hand much faster than traditional queueing models suggest.

This work takes the first steps towards identifying and clarifying an important phenomenon: the impact of delays on service times. The foundation developed here suggests a number of directions for further research. For instance, our empirical models imposed a linear relationship between delay and $\log(LOS)$. It would be useful to further explore the nature of the growth function $f(\cdot)$. Doing so will likely require significant data collection for each different patient type. Our empirical results demonstrate that many patient types demonstrate a delay effect and our queueing models indicate that this phenomenon significantly alters and hinders insights that can be extracted from traditional queueing models. This suggests a need for better models to be used for capacity planning. It would be interesting to develop heuristics for capacity management which are easily understandable by hospital managers, yet also account for the added complexity introduced by the delay-dependent service times.

Acknowledgments We gratefully acknowledge the editors and reviewers for their many helpful suggestions and comments, which have greatly improved this paper. The first author would like to thank Sarang Deo, Linda Green, and Marcelo Olivares, for their time and valuable feedback. We thank Marla Gardner and John

Greene for their help in preparing the data, along with the staff in the Division of Research and hospitals in Kaiser Permanente Northern California for their time and invaluable contributions to this research. The work by Carri W. Chan was supported in part by NSF CAREER grant number CMMI-1350059. The work by Vivek F. Farias was supported in part by NSF CAREER grant number CMMI-1054034

References

Allon, G., S. Deo, W. Lin. 2013. The impact of hospital size and occupancy of hospital on the extent of ambulance diversion: Theory and evidence. *Operations Research* **61** 554–562.

Anand, K., M. F. Pac, S. Veeraraghavan. 2010. Quality-Speed Conundrum: Tradeoffs in Customer-Intensive Services. *Management Science* **57** 40–56.

Anderson, D., C. Price, B. Golden, W. Jank, E. Wasil. 2011. Examining the discharge practices of surgeons at a large medical center. *Health Care Management Science* 1–10.

Ata, B., S. Shnerson. 2006. Dynamic control of an M/M/1 service system with adjustable arrival and service rates. *Management Science* **52** 1778–1791.

Boxma, O.J., M. Vlasiou. 2007. On queues with service and interarrival times depending on waiting times. *Queueing Systems* **56** 121–132.

Buist, M.D., G.E. Moore, S.A. Bernard, B.P. Waxman, J.N. Anderson, T.V. Nguyen. 2002. Effects of a medical emergency team on reduction of incidence of and mortality from unexpected cardiac arrests in hospital: Preliminary study. *British Medical Journal* **324** 387–390.

Chalfin, D. B. 2005. Length of intensive care unit stay and patient outcome: The long and short of it all. *Critical Care Medicine* **33** 2119–2120.

Chalfin, D. B., S. Trzeciak, A. Likourezos, B. M. Baumann, R. P. Dellinger. 2007. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. *Critical Care Medicine* **35** 1477–1483.

Chan, C. W., V. F. Farias, N. Bambos, G. Escobar. 2012. Optimizing ICU discharge decisions with patient readmissions. *Operations Research* **60** 1323–1342.

Chan, P.S., H.M. Krumholz, G. Nichol, B.K. Nallamothu. 2008. Delayed time to defibrillation after in-hospital cardiac arrest. *New England Journal of Medicine* **358**(1) 9–17.

de Luca, G., H. Suryapranata, J.P. Ottervanger, E.M. Antman. 2004. Time delay to treatment and mortality in primary angioplasty for acute myocardial infarction: Every minute of delay counts. *Circulation* **109**(10) 1223–1225.

de Véricourt, F., O. B. Jennings. 2011. Nurse staffing in medical units: A queueing perspective. *Operations Research* **59** 1320–1331.

Dobson, G., H.H. Lee, E. Pinker. 2010. A model of ICU bumping. *Operations Research* **58** 1564–1576.

Dong, J., P. Feldman, G. Yom-Tov. 2015. Slowdown services: Staffing service systems with load-dependent service rate. *Operations Research, to appear* .

Durrett, R. 1996. *Probability: Theory and Examples*. Duxbury Press.

Escobar, G. J., J. D. Greene, P. Scheirer, M. N. Gardner, D. Draper, P. Kipnis. 2008. Risk-adjusting hospital inpatient mortality using automated inpatient, outpatient, and laboratory databases. *Medical Care* **46** 232–239.

George, J. M., J. M. Harrison. 2001. Dynamic control of a queue with adjustable service rate. *Operations Research* **49**(5) 720–731.

Ghahramani, S. 1986. Finiteness of moments of partial busy periods for M/G/C queues. *Journal of Applied Probability* **23**(1) 261–264.

Green, L. 2006. Queueing analysis in healthcare. Randolph W. Hall, ed., *Patient Flow: Reducing Delay in Healthcare Delivery, International Series in Operations Research & Management Science*, vol. 91. Springer US, 281–307.

Green, L. V., J. Soares, J. F. Giglio, R. A. Green. 2006. Using queuing theory to increase the effectiveness of emergency department provider staffing. *Academic Emergency Medicine* **13** 61–68.

Green, L.V. 2003. How many hospital beds? *Inquiry-The Journal Of Health Care Organization Provision And Financing* **39** 400–412.

Halpern, N. A., S. M. Pastores. 2010. Critical care medicine in the united states 2000-2005: An analysis of bed numbers, occupancy rates, payer mix, and costs. *Critical Care Medicine* **38** 65–71.

Heckman, J.J. 1979. Sample selection bias as a specification error. *Econometrica* **47** 153–161.

Kc, D., C. Terwiesch. 2009. Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* **55** 1486–1498.

Kc, D., C. Terwiesch. 2012. An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* **14** 50–65.

Kim, S-H, C. W. Chan, M. Olivares, G. Escobar. 2015. Icu admission control: An empirical study of capacity allocation and its implication on patient outcomes. *Management Science* **61** 19–38.

Kuntz, L., R. Mennicken, S. Scholtes. 2014. Stress on the Ward: Evidence of Safety Tipping Points in Hospitals. *Management Science, to appear* .

Litvak, E., M.C. Long, A.B. Cooper, M.L. McManus. 2001. Emergency department diversion: Causes and solutions. *Acad Emerg Med* **8** 1108–1110.

McClellan, M, BJ McNeil, JP Newhouse. 1994. Does more intensive treatment of acute myocardial infarction in the elderly reduce mortality?: Analysis using instrumental variables. *JAMA* **272**(11) 859–866.

McManus, M. L., M. C. Long, A. Cooper, E. Litvak. 2004. Queuing theory accurately models the need for critical care resources. *Anesthesiology* **100** 1271–1276.

Meijer, E., T. Wansbeek. 2007. The sample selection model from a method of moments perspective. *Econometric Reviews* **26**(1) 25–51.

Moreno, R.P., P. G. Metnitz, E. Almeida, B. Jordan, P. Bauer, R.A. Campos, G. Iapichino, D. Edbrooke, M. Capuzzo, J.R. Le Gall. 2005. SAPS 3–From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. *Intensive Care Med* **31** 1345–1355.

Norton, S.A., L.A. Hogan, R.G. Holloway, H. Temkin-Greener, M.J. Buckley, T.E. Quill. 2007. Proactive palliative care in the medical intensive care unit: Effects on length of stay for selected high-risk patients. *Crit Care Med* **35** 1530–1535.

Powell, S. G., K. L. Schultz. 2004. Throughput in serial lines with state-dependent behavior. *Management Science* **50** 1095–1105.

Rapoport, J., D. Teres, S. Lemeshow. 1996. Resource use implications of do not resuscitate orders for intensive care unit patients. *Am J Respir Crit Care Med* **153** 185–190.

Renaud, B., A. Santin, E. Coma, N. Camus, D. Van Pelt, J. Hayon, M. Gurgui, E. Roupie, J. Hervé, M.J. Fine, C. Brun-Buisson, J. Labarère. 2009. Association between timing of intensive care unit admission and outcomes for emergency department patients with community-acquired pneumonia. *Critical Care Medicine* **37**(11) 2867–2874.

Rivers, E., B. Nguyen, S. Havstad, J. Ressler, A. Muzzin, B. Knoblich, E. Peterson, M. Tomlanovich. 2001. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *New England Journal of Medicine* **345**(19) 1368–1377.

Schwiebert, J. 2012. Revisiting the Composition of the Female Workforce - A Heckman Selection Model with Endogeneity. Diskussionspapiere der Wirtschaftswissenschaftlichen Fakultt der Leibniz Universitt Hannover dp-502, Leibniz Universitt Hannover, Wirtschaftswissenschaftliche Fakultt. URL `http://ideas.repec.org/p/han/dpaper/dp-502.html`.

Sheridan, R., J Wber, K Prelack, L. Petras, M. Lydon, R. Tompkins. 1999. Early burn center transfer shortens the length of hospitilization and reduces complications in children with serious burn injuries. *J Burn Care Rehabil* **20** 347–50.

Shi, P., M. C. Chou, J. G. Dai, D. Ding, J. Sim. 2015. Models and insights for hospital inpatient operations: Time-dependent ed boarding time. *Management Science, to appear* .

State of California Office of Statewide Health Planning & Development. 2010-2011. Annual Financial Data. URL `http://www.oshpd.ca.gov/HID/Products/Hospitals/AnnFinanData/CmplteDataSet/index.asp`.

Thompson, S., M. Nunez, R. Garfinkel, M.D. Dean. 2009. Efficient short-term allocation and reallocation of patients to floors of a hospital during demand surges. *Operations Research* **57**(2) 261–273.

Wansbeek, T., E. Meijer. 2000. Measurement Error and latent Variables in Econometrics. C.J. Bliss, M.D. Intriligator, eds., *Advanced Textbooks in Economics*, vol. 37. Elsevier cience, Amsterdam, The Netherlands.

Whitt, W. 1990. Queues with service times and interarrival times depending linearly and randomly upon waiting times. *Queueing Systems* **6** 335–352.

Wooldridge, J.M. 2002. *Econometric Analysis of cross section and panel data*. The MIT Press.

Yankovic, N., S. Glied, L.V. Green, M. Grams. 2010. The impact of ambulance diversion on heart attack deaths. *Inquiry* **47** 81–91.

Yankovic, N., L. Green. 2011. Identifying good nursing levels: A queuing approach. *Operations Research* **59** 942–955.

Zimmerman, J. E., A. A. Kramer, D.S. McNair, F. M. Malila. 2006. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients. *Crit Care Med* **34** 1297–1310.

Appendix A: Supplementary Information for Empirical Analysis

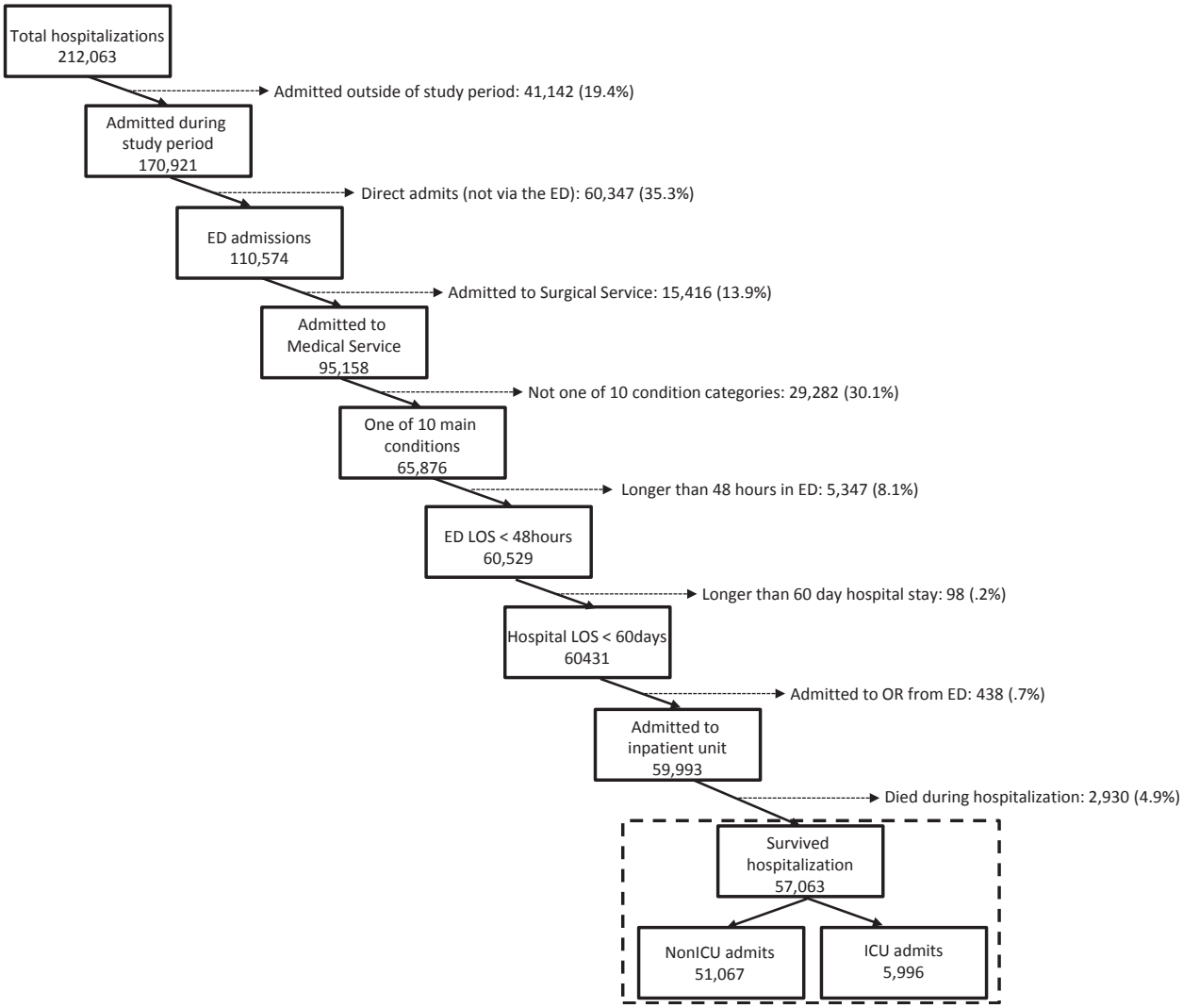


Figure 9 Selection of the patient sample. Final cohort in dotted-lined box: aggregated and split by patients who were admitted to the ICU and nonICU units.

Variable	Description
Age	Patient's age at hospital admission: coded as piecewise linear spline with knots at 40, 65, 75, and 85 years
Gender	Male or Female
LAPS	Laboratory-based Acute Physiology Score which uses information from 14 laboratory tests obtained 24 hours preceding hospital admission (see Escobar et al. (2008) for more information)
COPS	Comorbidity Point Score based on 40 different comorbidities recorded in inpatient and outpatient data 12 months preceding hospital admission (see Escobar et al. (2008) for more information)
Hospital	Indicator variables for each of the 19 hospitals a patient may be treated
Admission Day	Indicator variables for the day of week the patient was admitted to the hospital
Admission Month	Indicator variables for the month the patient was admitted to the hospital
Admission Shift	Indicator variables for the nursing shift the patient was admitted to the hospital. Nursing shifts are 8 hours: from 7am-3pm, 3pm-11pm, and 11pm-7am

Table 3 Control Variables.

Appendix B: Miscellaneous Proofs

PROOF OF PROPOSITION 1:

Weak Stability: First, we show that if $\frac{\lambda}{s} \leq \frac{1}{1+f_{\sup}}$, then the system is weakly stable. This follows by examining a traditional $M/M/s$ system with arrival rate λ and mean service requirement $1+f_{\sup} = 1+\sup_m f(m)$. By coupling the arrivals of this system and the service times so that if the mean service requirement in our delay-dependent system is $\sigma \leq 1+f_{\sup}$, its service requirement is σX and the service requirement in the $M/M/s$ system is $(1+f_{\sup})X$ where X is a mean 1, exponentially distributed random variable. It is easy to see that this $M/M/s$ system is an upper bounding system to our delay-dependent system. Hence, if the upper bounding system is stable, so is the $M/M(f)/s$ system. The stability condition for this upper bounding system is the desired criteria.

Instability: We now show that if $\frac{\lambda}{s} > \frac{1}{1+f_{\sup}}$, then the system is unstable. We do this in two steps: 1) we show that from any initial state, there is a non-zero probability that the time until the $M/M(f)/s$ system will reach the state where the number of jobs in the system is such that the service time of a new arrival would be maximally inflated and all the jobs in the system have been been delayed enough that their service rate is maximal is finite 2) we establish the transience of this state which will establish that our resulting system is transient and, hence, unstable.

We define the following notation: Let $N_{f_{\sup}} = \min\{N : f(N) = \sup_n f(n)\}$ be the minimum number of jobs in the system such that the service time for a new job is inflated maximally. Note that by assumption f takes values in a finite set, so $N_{f_{\sup}}$ exists. Our state at time t can be described by the $N_{f_{\sup}}$-dimensional vector, Z_t, where $(Z_t)_n$ is the number of jobs in the system which saw n jobs when it arrived ($Z_{N_{f_{\sup}}}$ is the number of jobs which see $N_{f_{\sup}}$ or more jobs in the system). Let $T_{xy} = \inf\{t > 0 : Z_t = y | Z_0 = x\}$ be the time to first passage to state y given we start in state x at time 0. Finally, we define the state with exactly $\hat{N} = \max(N_{f_{\sup}}, s)$ jobs in the system, all of whose service time is maximally inflated as $S^* = \{Z : Z_{N_{f_{\sup}}} = \sum_n Z_n = \hat{N}\}$.

We begin by showing that the time to reach state S^* is finite with non-zero probability from any initial state. Specifically, we will show that for any state x, $P(T_{xS^*} < \infty) > 0$. Consider a system which starts at state x, i.e. $Z_0 = x$. Let N_x be the number of jobs in the system in state x. We start with assuming $N_x < N_{f_{\sup}} + \hat{N}$. Our goal is to find the first time to state S^*. One way to get to S^* is to have $\hat{N} + N_{f_{\sup}} - N_x$ jobs arrive before any job departs the system and then have $N_{f_{\sup}} - N_x$ jobs depart from the system before another job arrives. Thus, the probability of this particular sample path occurring, which we denote as A, can be lower bounded by:

$$P(A) > \left(\frac{\lambda}{\lambda + s\mu_{\max}}\right)^{\hat{N}+N_{f_{\sup}}-N_x} \left(\frac{s\mu_{\min}}{\lambda + s\mu_{\max}}\right)^{N_{f_{\sup}}} > 0$$

where $\mu_{\max} = 1$ and $\mu_{\min} = 1/(1+f_{\sup})$ are the maximal and minimal service rates, respectively. Moreover, the time it takes for this cascade of events to occur is upper bounded by the sum of $\hat{N} + 2N_{f_{\sup}} - N_x$, mean $1/(\lambda + s\mu_{\min})$ exponentially distributed random variables. Specifically, the time has a gamma distribution $T_A \sim \Gamma(\hat{N} + 2N_{f_{\sup}} - N_x, 1/(\lambda + s\mu_{\min}))$, which is finite with non-zero probability. Hence, we have that:

$$P(T_{xS^*} < \infty) > P(A)P(T_A < \infty) > 0$$

Note that if $N_x > N_{f_{\sup}} + \hat{N}$, we simply need that $N_x - \hat{N}$ jobs must depart before the next arrival. Using the same argument as above, we can show that $P(T_{xS^*} < \infty) > 0$, for any x.

Next, we demonstrate that the recurrence time for state S^* is infinite with non-zero probability, i.e. $P(T_{S^*S^*} < \infty) < 1$. To do this, we will leverage the fact that a standard $M/M/s$ queueing system with $\rho = \frac{\lambda}{s\mu_{\min}} = \frac{\lambda(1+f_{\sup})}{s} > 1$ is unstable, and hence, transient. We consider two states in this $M/M/s$ system: state y, with $\hat{N}$ jobs in the system, and state y^+, with $\hat{N}+1$ jobs in the system. Because this $M/M/s$ system is transient, the time to first passage from y^+ to y satisfies the following: $P(T_{y^+y}^{M/M/s} < \infty) < 1$. Here we use the superscript $M/M/s$ to differentiate from the first passage time of our delay dependent $M/M(f)/s$ system, T_{xy}.

We leverage the the preceding observation and decompose the recurrence time $T_{S^*S^*}$ into whether the next event is an arrival or departure with the new state denoted by y^+ and y^-, respectively:

$$\begin{aligned} P(T_{S^*S^*} < \infty) &= \frac{s\mu_{\min}}{\lambda + s\mu_{\min}} P(T_{y^-S^*} < \infty) + \frac{\lambda}{\lambda + s\mu_{\min}} P(T_{y^+S^*} < \infty) \\ &\leq \frac{s\mu_{\min}}{\lambda + s\mu_{\min}} + \frac{\lambda}{\lambda + s\mu_{\min}} P(T_{y^+S^*} < \infty) < 1 \end{aligned}$$

The last inequality comes from the observation that we started at state S^*, an arrival occurred so we are now at start y^+ and we are considering the recurrence time to return to state S^*. Now there are $\hat{N}+1$ jobs in the system and any job that arrives to the system will see at least $\hat{N} \geq N_{f_{\sup}}$ jobs in the system before the system hits state S^*. If this were not the case, the system will have already returned to state S^*. Therefore, all new jobs will have service time exponentially distributed with mean $1 + f_{\sup} = 1/\mu_{\min}$. Hence, the dynamics of our $M/M(f)/s$ system are identical to the $M/M/s$ system with arrival rate λ and service rate $\mu_{\min}$ during the trajectory to the first visit to state S^* from state y^+. Because the $M/M/s$ system is transient, state S^* is also transient in our $M/M(f)/s$ system.

By Theorem 3.4 in Durrett (1996), all states in our $M/M(f)/s$ system are transient since the time to reach a transient state ($y \in S^*$) is finite with non-zero probability for all states. Hence, the $M/M(f)/s$ queue is unstable. □

Appendix C: A Markovian Model

For the sake of concreteness and simplicity of exposition we will consider a very simple $f(\cdot)$. In particular, we assume that the workload increase function, $f(\cdot)$, is defined as follows:

$$f(m) = \begin{cases} 0, & m < s; \\ k, & m \geq s. \end{cases}$$

Hence, the mean service time of each job is 1 if there are fewer than s jobs in the system upon arrival and $1+k$ otherwise. This means any job which is delayed will have an increased service requirement.

Let $X = (X_N, X_D)$ be the system state where X_N is the number of jobs in the system who arrived with less than s jobs currently in the system and X_D is the number of jobs in the system who arrives with s or more jobs in the system, and hence experiences an increase in service requirement. Note that due to the FCFS and non-preemptive service discipline, if $X_N > 0$, then necessarily there are $(X_N \wedge s)$ jobs currently in service at rate 1. The remaining servers, $(s - X_N)^+$, will be serving jobs at rate $\frac{1}{1+k}$ if any are available. Otherwise, they will idle. We can verify that the Markov Property holds for our state as defined.

Proposition 2 *An $M/M(f)/s$ system with $f(m) = k1_{\{m \geq s\}}$ can be represented as a Markovian system with state $X = (X_N, X_D)$.*

PROOF: We show that the Markov Property holds for our system. We let $X(i) = (X_N(i), X_D(i))$ be the state at the ith state transition. What's left to show is that

$$P(X(i+1) = (x_N, x_D)|X(0), X(1), \ldots, X(i-1), X(i)) = P(X(i+1) = (x_N, x_D)|X(i))$$

We demonstrate this by considering the precise transition probabilities:

$$\begin{aligned} P(X(i+1) =\ & (x_N, x_D)|X(0), X(1), \ldots, X(i-1), X(i) = (x'_N, x'_D)) \\ =\ & \begin{cases} \frac{\lambda}{\lambda + (x'_N \wedge s) + \frac{x'_D \wedge (s - x'_N)^+}{1+k}}, & \text{if } (x_N, x_D) = (x'_N + 1, x'_D) \text{ and } x'_N + x'_D < s; \\ \frac{\lambda}{\lambda + (x'_N \wedge s) + \frac{x'_D \wedge (s - x'_N)^+}{1+k}}, & \text{if } (x_N, x_D) = (x'_N, x'_D + 1) \text{ and } x'_N + x'_D \geq s; \\ \frac{x'_N \wedge s}{\lambda + (x'_N \wedge s) + \frac{x'_D \wedge (s - x'_N)^+}{1+k}}, & \text{if } (x_N, x_D) = (x'_N - 1, x'_D); \\ \frac{\frac{x'_D \wedge (s - x'_N)^+}{1+k}}{\lambda + (x'_N \wedge s) + \frac{x'_D \wedge (s - x'_N)^+}{1+k}}, & \text{if } (x_N, x_D) = (x'_N, x'_D - 1); \\ 0, & \text{otherwise.} \end{cases} \\ =\ & P(X(i+1) = (x_N, x_D)|X(i) = (x'_N, x'_D)) \end{aligned} \tag{9}$$

It is clear that the transition probabilities depend only on the current state and are independent of the past. □

For many other f functions, the system will still be Markovian with an appropriately defined state space; however, the size of the state space will grow rapidly with more complex f functions.

Appendix D: Proof of Theorem 2

We now proceed with the proof of our main result. The proof will examine the case of Theorem 1, which assumes that the growth function f is defined as:

$$f(m) = \begin{cases} 0, & m < N^*; \\ k, & m \geq N^*. \end{cases}$$

We note that the generalized result for Theorem 2 will follow similarly. The only changes required are additional notation and book keeping to keep track of each breakpoint in the growth function, f. The proof will proceed in several steps. Again we will refer to our $M/M(f)/s$ system as system 1 and an $M/M/s$ system with arrival rate λ and service rate $1/(1+k)$ as system 2.

Coupling: To begin we will construct a natural coupling between the $M/M(f)/s$ and $M/M/s$ systems above. In particular, we assume that both systems see a common arrival process. With an abuse of notation, let the workload introduced by the ith arriving job–equivalently, this is the service time as the service rates have been normalized to 1–in the latter system be $\overline{\sigma}_i$; the corresponding service time in the delay dependent system is then either $\sigma_i = \overline{\sigma}_i/(1+k)$ or $\sigma_i = \overline{\sigma}_i$ depending on whether the delay dependent system has low congestion ($N_{t_i^-} < N^*$) or is considered busy ($N_{t_i^-} \geq N^*$) upon the arrival of the ith job. Finally, we assume that both systems start empty. Now let τ_i ($\overline{\tau}_i$) denote the amount of time the ith arriving job waits in the former (latter) system respectively before beginning service. We have, as a consequence of our coupling, the following elementary result:

Proposition 3 *$\tau_i \leq \overline{\tau}_i$ for all i. Moreover, $N_t \leq \overline{N}_t$ for all t.*

PROOF: We prove the first statement. Proceeding by induction observe that the statement is true for $i = 1$: $\tau_1 = \overline{\tau}_1 = 0$. Assume the statement true for $i = l - 1$ and consider $i = l$. For the sake of contradiction assume that $\tau_l > \overline{\tau}_l$. Since the service discipline is FCFS in both systems, it follows that when job l starts service in system 2:

- There are at most $s-1$ jobs from among the first $l-1$ arriving jobs present in system 2.
- Simultaneously, at *least* s jobs from among the first $l-1$ arriving jobs are still present in system 1 since job l has not yet started service in system 1.

Consequently, given the induction hypothesis and the fact that by our coupling $\sigma_i \leq \overline{\sigma}_i$ for $i = 1, 2, \ldots, l-1$, there is a job among the first $l-1$ arrivals that finished service strictly earlier in system 2 than in system 1. This is a contradiction. We have consequently established that $\tau_i \leq \overline{\tau}_i$ for all i. The latter statement follows as a simple corollary. □

We next use the result above to construct a first upper bound. We have:

Proposition 4

$$\sigma_i\tau_i + \frac{1}{2}\sigma_i^2 \leq \overline{\sigma}_i\overline{\tau}_i + \frac{1}{2}\overline{\sigma}_i^2 - \mathbf{1}\left\{\overline{N}_{t_i^-} < N^*\right\}\frac{1}{2}\overline{\sigma}_i^2\left(\frac{2k+k^2}{(1+k)^2}\right)$$

PROOF: We begin with two elementary observations. First,

$$\sigma_i \leq \overline{\sigma}_i$$

always under our coupling and, in particular, if $\overline{N}_{t_i^-} \geq N^*$. Further

$$\sigma_i \leq \frac{\overline{\sigma}_i}{1+k}$$

if $\overline{N}_{t_i^-} < N^*$. This follows from the fact that $\overline{N}_t > N_t$ (Proposition 3), so that $\overline{N}_{t_i^-} < N^*$ implies $N_{t_i^-} < N^*$. It follows that

$$\begin{aligned}
\sigma_i\tau_i + \frac{1}{2}\sigma_i^2 &\leq \overline{\sigma}_i\overline{\tau}_i + \frac{1}{2}\sigma_i^2 \\
&\leq \overline{\sigma}_i\overline{\tau}_i + \mathbf{1}\left\{\overline{N}_{t_i^-} \geq N^*\right\}\frac{1}{2}\overline{\sigma}_i^2 + \mathbf{1}\left\{\overline{N}_{t_i^-} < N^*\right\}\frac{1}{2}\frac{\overline{\sigma}_i^2}{(1+k)^2} \\
&= \overline{\sigma}_i\overline{\tau}_i + \frac{1}{2}\overline{\sigma}_i^2 - \mathbf{1}\left\{\overline{N}_{t_i^-} < N^*\right\}\frac{1}{2}\overline{\sigma}_i^2\left(\frac{2k+k^2}{(1+k)^2}\right)
\end{aligned}$$

The first inequality follows from the fact that $\sigma_i \leq \overline{\sigma}_i$ (by our coupling) and $\tau_i \leq \overline{\tau}_i$ (Proposition 3). The second inequality follows from the two observations we made at the outset. □

We next connect this result to the average workload in both systems (over a finite interval). Let $N(T)$ be the number of jobs that have arrived during $t \in [0, T]$. We have:

Proposition 5

$$\frac{1}{T}\int_0^T W_t dt \leq \frac{1}{T}\int_0^T \overline{W}_t dt - \frac{1}{T}\sum_{i=1}^{N(T)} \mathbf{1}\left\{\overline{N}_{t_i^-} < N^*\right\}\frac{1}{2}\overline{\sigma}_i^2\left(\frac{2k+k^2}{(1+k)^2}\right) + \frac{\overline{W}_T}{T}$$

PROOF: Notice that the total workload contributed by job i over time in system 1 is given by the quantity $\sigma_i\tau_i + \frac{1}{2}\sigma_i^2$ where the first term in the sum corresponds to the workload contributed while job i waits, and the latter term corresponds to the workload contributed while job i is in service. We consequently have:

$$\begin{aligned}
\frac{1}{T}\int_0^T W_t dt &\leq \frac{1}{T}\sum_{i=1}^{N(T)}\left(\sigma_i\tau_i + \frac{1}{2}\sigma_i^2\right) \\
&\leq \frac{1}{T}\sum_{i=1}^{N(T)}\left(\overline{\sigma}_i\overline{\tau}_i + \frac{1}{2}\overline{\sigma}_i^2\right) - \frac{1}{T}\sum_{i=1}^{N(T)} \mathbf{1}\left\{\overline{N}_{t_i^-} < N^*\right\}\frac{1}{2}\overline{\sigma}_i^2\left(\frac{2k+k^2}{(1+k)^2}\right) \\
&= \frac{1}{T}\int_0^T \overline{W}_t dt + \frac{\overline{W}_T}{T} - \frac{1}{T}\sum_{i=1}^{N(T)} \mathbf{1}\left\{\overline{N}_{t_i^-} < N^*\right\}\frac{1}{2}\overline{\sigma}_i^2\left(\frac{2k+k^2}{(1+k)^2}\right)
\end{aligned}$$

□

Note that the last equality comes from the fact that not all of the work which arrives between $[0,T]$ is completed by time T; hence $\bar{W}_T$ remains. What remains is to take limits on both sides of the inequality established in the previous result. To that end we begin with a few intermediary results. First, we provide a few definitions. We let $E[W]$ and $E[\overline{W}]$ be the expected work in our $M/M(f)/s$ system and an $M/M/s$ system with $\rho = \frac{\lambda(1+k)}{s}$, respectively.

Lemma 1 $\lim_{T\to\infty} \frac{1}{T}\int_0^T W_t dt = E[W]$

PROOF: This result follows directly from the renewal reward theorem and the fact that the system is stable. The reward function is the cumulative work and is defined as: $R(t) = \int_0^t W_\tau d\tau$ □

Lemma 2 $\lim_{T\to\infty} \frac{1}{T}\int_0^T \overline{W}_t dt = E[\overline{W}]$

PROOF: Again, this result follows directly from the renewal reward theorem and the fact that the system is stable. The reward function is the cumulative work and is defined as: $R(t) = \int_0^t \overline{W}_\tau d\tau$ □

Lemma 3 $\lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}\{\overline{N}_t < N^*\} dt = P(\overline{N} < N^*)$

PROOF: Again, this result follows directly from the renewal reward theorem and the fact that the system is stable. The reward function is the total time the number of jobs in the system is less than N^* and is defined as: $R(t) = \int_0^t \mathbf{1}\{\overline{N}_\tau < N^*\} d\tau$ □

Lemma 4 $\lim_{T\to\infty} \frac{W_T}{T} = 0$

PROOF: This follows from the fact that the system is stable and thus recurrent. If we consider that W_T is upper bounded by the amount of work that arrives between $[T_0^*(T), T]$, where $T_0^*(T) = \sup\{t < T : W_t = 0\}$ is the last time before T, the system was empty, then the fact that the system is recurrent establishes that $P(T - T_0^*(T) < \infty) = 1$. Assuming a finite first moment for σ_i gives the desired result. □

We next establish a limit for the second term on the right hand side of the inequality in Proposition 4.

Proposition 6

$$\lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{N(T)} \mathbf{1}\left\{\overline{N}_{t_i^-} < N^*\right\} \frac{1}{2}\overline{\sigma}_i^2 \left(\frac{2k+k^2}{(1+k)^2}\right) = \lambda(2k+k^2) \lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}\left\{\overline{N}_t < N^*\right\} dt$$

PROOF: Let use denote, for notational convenience,

$$\frac{1}{2}E\overline{\sigma}_i^2 \left(\frac{2k+k^2}{(1+k)^2}\right) = 2k+k^2 \triangleq \alpha.$$

and

$$\frac{1}{2}\left(\frac{2k+k^2}{(1+k)^2}\right) \triangleq \beta.$$

We begin with observing that

$$\lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{N(T)} \mathbf{1}\left\{\overline{N}_{t_i^-} < N^*\right\} \alpha = \lambda\alpha \lim_{T\to\infty} \frac{1}{T}\int_0^T \mathbf{1}\left\{\overline{N}_t < N^*\right\} dt \tag{10}$$

by PASTA. Next, observe that

$$\begin{aligned} E\left[\lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\alpha\right] &= \lim_{T\to\infty}\frac{1}{T}E\left[\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\alpha\right] \\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N(T)}E\left[\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\alpha\right] \\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N(T)}E\left[\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\right]E\overline{\sigma}_i^2\beta \qquad (11) \\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N(T)}E\left[\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2\right]\beta \\ &= \lim_{T\to\infty}E\left[\frac{1}{T}\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2\right]\beta \end{aligned}$$

The first equality above follows by dominated convergence (using the dominating random variable $N(T)/T$). The fourth equality (which is crucial) follows since $\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}$ and $\overline{\sigma}_i^2$ are independent random variables. Recall these are defined for the standard $M/M/s$ system. Now, since $\lim_{T\to\infty}\frac{1}{T}\int_0^T\mathbf{1}\left\{\overline{N}_t<N^*\right\}dt$ is a constant (by Lemma 3), (10) and (11) together yield

$$\lim_{T\to\infty}E\left[\frac{1}{T}\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2\right]\beta=\lambda\alpha\lim_{T\to\infty}\frac{1}{T}\int_0^T\mathbf{1}\left\{\overline{N}_t<N^*\right\}dt.$$

But from Lemma 5, which will come in Appendix D.1

$$\begin{aligned} \lim_{T\to\infty}E\left[\frac{1}{T}\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2\right]\beta &= E\left[\lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2\right]\beta \\ &= \lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2\beta \end{aligned}$$

Using Lemmas 1 and 2 to replace the limit with expectations gives the desired result. This completes the proof. □

D.1. Existence of a Limit

The *partial busy period* of an $M/G/s$ queue is defined as the time between when an arriving customer sees an empty system and the first time after that at which a departing customer sees an empty system. We will use the following result:

Theorem 3 (Ghahramani (1986)) *The mth moments of the partial busy period of an $M/G/s$ queue are finite if and only if the service time distribution has finite mth moments.*

We denote by T_m the length mth partial busy period. We can now establish:

Lemma 5 *Assume the service time distribution has finite fourth moments. Then,*

$$\lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2$$

exists and equals a constant. Further,

$$\lim_{T\to\infty}E\left[\frac{1}{T}\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2\right]\beta=E\left[\lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2\right]\beta$$

PROOF: We first establish that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2\beta$$

exists and is constant. To see this denote by $1=j_1<j_2<j_3\ldots$ the arrivals i for which $\overline{N}_{t_i^-}=0$. Observe that the random variables

$$X_m\triangleq\sum_{i=j_m}^{j_{m+1}-1}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2$$

are independent random variables. Moreover, $\sum_{i=j_m}^{j_{m+1}-1}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2\le s^2T_m^2$. Note that since we have assumed the service time distribution has finite fourth moments, we have $ET_m^4<\infty$. Now let $M(T)=\sup\{l|A_{j_l}\le T\}$; $M(T)\to\infty$. The strong law of large numbers then implies that

$$\lim_{T\to\infty}\frac{\sum_{i=1}^{M(T)}X_m}{M(T)}$$

exists and is a constant a.s. Further, a simple argument using Chebyshev's inequality and the Borel Cantelli lemma implies that

$$\lim_{T\to\infty}\frac{X_{M(T)}}{T}=0\text{ a.s.}$$

Finally, the elementary renewal theorem implies that $\lim_{T\to\infty}\frac{M(T)}{T}=1/ET_1$. But,

$$\frac{\sum_{i=1}^{M(T)}X_m}{M(T)}\frac{M(T)}{T}-\frac{X_{M(T)}}{T}\le\frac{1}{T}\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2\le\frac{\sum_{i=1}^{M(T)}X_m}{M(T)}\frac{M(T)}{T}+\frac{X_{M(T)}}{T}$$

so that taking limits throughout and employing the above observations yields the first conclusion of the Lemma.

Now to establish the second conclusion, observe that

$$\sum_{i=1}^{N(T)}\mathbf{1}\left\{\overline{N}_{t_i^-}<N^*\right\}\overline{\sigma}_i^2\le\sum_{i=1}^{N(T)}\overline{\sigma}_i^2$$

and that

$$E\sum_{i=1}^{N(T)}\overline{\sigma}_i^2=EN(T)E\sigma_i^2=\lambda TE\sigma_i^2$$

where the first equality is Wald's identity. Consequently, we may apply the conclusion of the first part of the theorem along with the dominated convergence theorem to establish the second conclusion of the theorem.

□