Gatekeeping Under Uncertainty: An Empirical Study of Referral Errors in the Emergency Department

Michael Freeman
Judge Business School, University of Cambridge, Cambridge CB2 1AG, United Kingdom mef35@cam.ac.uk

Stefan Scholtes
Judge Business School, University of Cambridge, Cambridge CB2 1AG, United Kingdom s.scholtes@jbs.cam.ac.uk

We study admission and discharge errors made by physicians in a congested emergency department (ED) using a data set comprising more than 600,000 visits over a seven-year period. We find that when the ED becomes busier, physicians make increasingly more admission errors but also fewer discharge errors. This leads to a bullwhip-type effect: demand surges in the ED leads to relatively greater demand pressures in the hospital, as more patients are admitted unnecessarily. While this behavior can be rationalized at the level of the individual ED physician, who deploys a "safety first" principle and admits patients in case of doubt, the overall system effect is detrimental. In particular, unnecessary use of specialist hospital services leads to higher system costs, while a higher level of hospital occupancy is also known to have detrimental effects on patient outcomes e.g. longer lengths of stay and higher mortality rates. Having established the bullwhip phenomenon, we consider whether replacing the direct one-stage (ED physician to hospital) referral system with a two-stage process that also allows ED physicians to stream patients to an intermediate "semi-specialist" referral unit – in our context operationalized by a Clinical Decisions Unit – when diagnostic uncertainty is high can improve care coordination in gatekeeping systems. We find that such a unit can significantly reduce the negative demand-propagation effect, and the two-stage process to be significantly more effective than simply adding ED capacity. This beneficial effect is particularly pronounced for more error prone and risk averse gatekeepers, and when patients present with a higher degree of diagnostic uncertainty.

Key words: gatekeeper systems; routing; server behavior; uncertainty; health care: hospitals; service operations; econometrics
History: October 19, 2016

Preliminary version – Please do not cite or circulate.

1. Introduction

Many service settings (e.g., health care, call centers, maintenance) are characterized by the presence of multiple service tiers, with customers commencing service at a low-cost entry level (e.g., an emergency department (ED), a general enquiries help-desk, a local repair shop) from where they can be referred to a more specialized and hence more costly level of service (e.g., acute hospital

bed, complaints desk, engineering department) if necessitated by the complexity of their needs. The upstream server (e.g., ED physician, telephonist, technician) in such a setting assumes a dual role: They will service simple requests themselves while, at the same time, acting as a *gatekeeper* to downstream specialist units, thereby ensuring that customers receive the appropriate service intensity for their needs (Shumsky and Pinker 2003).

Empirical research has demonstrated that high utilization of specialist resources leads to a deterioration of system performance, resulting in delays (KC and Terwiesch 2009, Chan et al. 2016), reduced service quality (Needleman et al. 2011, Kuntz et al. 2014, Tan and Netessine 2014), and poorer financial performance (Powell et al. 2012). From a system perspective, it would, therefore, be desirable if gatekeepers smoothed demand variation by rationing access to specialist resources when demand surges. However, recent empirical evidence suggests that precisely the opposite may occur: As congestion in the system increases, gatekeepers *increase* the rate at which they refer customers to specialists, further increasing the busyness of the specialists (Freeman et al. 2016). Are gatekeepers "opening the floodgates" to specialist services precisely at times when they should ration access to these services? If so, then their behavior causes a bullwhip-type effect: Demand surges faced by upstream gatekeepers lead to even greater relative demand surges for the more expensive downstream specialist units, with a detrimental effect on the service received by customers for whom the specialist services are most valuable. This paper investigates this behavioral inefficiency in the context of admission and discharge decisions made by physicians in a busy ED and examines a mechanism that can be used to counteract this behavioral bullwhip effect.

An important assumption made in the gatekeeping literature (reviewed in Section 2) is that gatekeepers are able to diagnose and rank customers in order of increasing complexity. However, correctly diagnosing a customer's needs and identifying how best to meet them can be challenging, in particular for the type of knowledge work that characterizes many gatekeeping settings (and especially so in medicine). Moreover, servers are often time- and resource-constrained, and so must trade-off the benefit of investing to acquire additional information that improves diagnostic accuracy (e.g., through further testing) against the cost of reduced throughput and delayed service for waiting customers (Alizamir et al. 2013). As a consequence, gatekeepers will often make referral decisions with only partially complete information and can therefore not avoid referral errors altogether. This is a concern since an incorrect referral decision can be costly for the service provider. If a customer who could have been self-served effectively by the gatekeeper is instead referred to the specialist, then the specialists' valuable time is wasted. Moreover, more complex customers – who gain more value from specialist services – may experience worse service and

poorer outcomes because of the resulting increase in specialist congestion. On the other hand, if the gatekeeper attempts to resolve a customer's problem by herself but fails, then this can lead to expensive delays, rework or even harm.

Much of the analytical work in the operations management and economics literature on gatekeeping has focused on this trade-off and the problem of identifying and incentivizing the optimal rate of specialist referrals (see literature in Section 2). These papers assume that gatekeepers do not incur disutility from an incorrect referral or self-service decision, and instead maximize the time-average income from wages plus bonuses per customer diagnosed and per customer successfully treated (e.g. Shumsky and Pinker 2003, Hasija et al. 2005). If, however, gatekeepers experience disutility – whether monetary or otherwise – when an error occurs, then this may have implications for how they behave when faced with differing levels of diagnostic uncertainty. Specifically, they will refer at a rate above the system-optimal rate if their disutility from a "missed referral" is significantly higher than their disutility from an erroneous referral. Our data suggests this to be the case in EDs, where physicians weigh a failure to admit a patient to the hospital as a more severe error than an unnecessary hospital admission. While this may be the best decision for the patient at hand, it does not internalize the cumulative negative effect of false admissions on the patients already in the hospital. Such patients are exposed to higher levels of hospital occupancy, with negative implications for service quality (e.g. Kuntz et al. 2014). Our empirical research examines: (i) the role of diagnostic uncertainty on referral decisions in congested systems, and specifically the consequences of asymmetric gatekeeper disutilities for false positive and false negative referrals, and (ii) the effect that an intermediate semi-specialist unit has on mitigating the demonstrated behavioral inefficiencies.

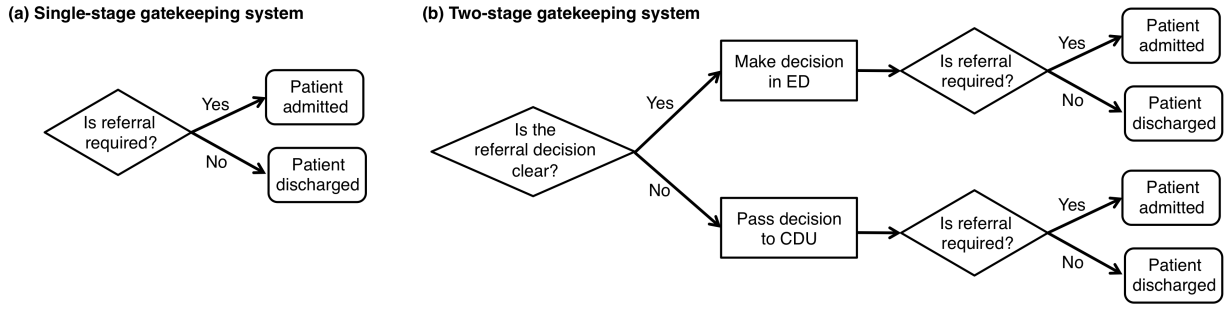
Our empirical study is based on over 650,000 patient attendances to the busy ED of a UK-based teaching hospital over a seven year period. ED physicians act as gatekeepers to expensive acute inpatient beds, responsible for restricting access to the main hospital to only those patients whose immediate treatment needs are too complex to be met by staff working within the ED itself. Despite the fact that one in 10 medical diagnoses are estimated to be wrong (Graber 2013), with errors in the diagnostic process the leading cause of internal investigation and malpractice claims in the ED (Kachalia et al. 2007, Cosby et al. 2008), ED physicians often experience high caseloads and must make decisions under considerable time pressure (see Section 3). As a consequence, unnecessary admission and inappropriate discharge decisions can occasionally occur: For the ED in our study, the error rate among admitted patients is estimated at 16.1% versus an error rate among discharged

patients of 1.3%.[1] The high rate of admission errors relative to discharge errors suggests that when faced with an uncertain decision ED physicians err on the side of caution and adopt a "safety-first" principle, preferring to minimize the risk that their patient leaves untreated over the risk of an incorrect admission that is costly for the hospital and may impede the service received by the other hospital patients, but is safe for the patient at hand (Roy 1952). To confirm this behavior, we study empirically how an increase in system congestion – which reduces the time available for diagnosis and so increases uncertainty – affects the rate of avoidable admissions to acute hospital beds and inappropriate discharges from the ED. We find that for every one standard deviation increase in ED busyness, ED physicians *increase* the rate at which they admit patients to the hospital by 7.7%. At the same time, there is also a *reduction* in the rate of errors in discharge by 3.3%. Thus, when faced with additional diagnostic uncertainty ED physicians adjust the rate of admissions in order to avoid a higher chance of an unlikely but potentially 'catastrophic' error in discharges. Moreover, they adjust beyond the rate that would be necessary to preserve prevailing rates of discharge error, which is suggestive of risk aversion. Importantly, this is precisely the opposite behavior to that which is desirable from a system perspective: Demand surges in the ED lead to more admission errors and therefore an amplification of the surge in the hospital (a bullwhip-type effect), with potentially negative implications for the other patients under their care.

Having established the undesirable behavioral effect of congestion on admission errors, the second part of this paper studies what can be done to mitigate the adverse impact of diagnostic uncertainty on performance (as measured by overuse of specialist resources) in gatekeeping systems such as this. We offer one potential solution: to allow the gatekeeper to classify a referral candidate as "unresolved" prior to making the referral decision, and offloading them instead to an intermediate second-stage gatekeeper who assumes responsibility for deciding whether a referral is necessary. Since these unresolved cases are more homogenous than the average patient arrival (with unambiguous referrals and non-referrals having already been filtered out by the first-stage gatekeeper), they can be looked after by a more specialized workforce and additional resources can be invested to acquire information to increase diagnostic accuracy and reduce referral errors. As it happens, this two-stage gatekeeping process already exists in the context of our study hospital by way of the presence of a clinical decisions unit (CDU). The CDU is a stand-alone unit attached to the ED into which a patient can be referred for further monitoring, diagnostic evaluation, and/or treatment. Beds in the CDU are of lower intensity and cost than acute beds in the main hospital,

[1] An admission error is defined to be any patient admitted to the hospital and subsequently discharged within 24 hours with no treatment provided, while a discharge error is defined to be a patient treated in the ED and sent home who returns to the ED within seven days and is at that point admitted. See Section 4.2 for more detail.

Figure 1 Flow charts of the traditional single-stage gatekeeping process (left) and the proposed two-stage gatekeeping process (right).



but patients are able to stay up to 24 hours (rather than 4 hours in the ED) and it is generally staffed by more experienced clinicians than the ED. The CDU thus provides ED physicians with an alternative to discharge or hospital inpatient admission that can be leveraged when it is unclear whether or not the patient should be admitted. A comparison of this two-stage process to the traditional gatekeeping set-up is shown in Figure 1.

After accounting for non-random assignment of patients to the CDU using appropriate sample selection methods, we show that patients routed through the CDU are 12.2% less likely to be admitted in error than patients admitted directly by ED physicians, while being no more likely to be discharged in error. Moreover, we find that patients admitted directly from the ED are 8.2% more likely to have a specialty transfer during their hospital stay than patients admitted via the CDU, indicating fewer routing errors within the hospital for CDU patients. We also identify conditions under which such a system adds the most value, specifically two such factors: (1) the degree of uncertainty associated with the customer's diagnosis, and (2) the risk attitude and error propensity of the assigned gatekeeper. We find that as diagnostic uncertainty increases, the CDU becomes increasingly beneficial as a means of reducing avoidable admissions, reducing the rate by 0.57% for patients estimated to be in the first quartile of the distribution of diagnostic uncertainty, which increases to a large 19.7% reduction for patients in the fourth quartile. We also find that while the rate of admission errors made on those patients referred to the CDU by less risk taking and more error prone physicians are reduced by 16.3% and 17.4%, respectively, this halves to approx. 10.0% and 10.7% for patients passed to the CDU by less error prone and more risk taking physicians. These findings suggest that a two-stage gatekeeping system is especially valuable in the presence of high levels of diagnostic uncertainty and less experienced gatekeepers.

Our study provides empirical support that intermediate "semi-specialist" gatekeeping units can help alleviate the trade-off between speed and quality in multi-tier service systems (see e.g. Anand et al. 2011, Kostami and Rajagopalan 2013, Alizamir et al. 2013): While it might appear desirable

to incentivise gatekeepers to make referral decisions faster when the system is congested, to increase throughput and ensure that customers receive prompt service, this can reduce the time available for accurate diagnosis and lead to increased referral errors which can erode the benefits of higher throughput and lead to worse outcomes and higher system costs. An example in point is the introduction of a waiting time target in the NHS in 2004, requiring that 98% (adjusted later to 95%) of patients be admitted or discharged within four hours of arrival to the ED. This target led to faster decision-making in the ED and reduced waiting times. However, it also coincided with a 30% increase in hospital admission rates, at a multi-billion pound cost to the UK healthcare system (NAO 2013). Our findings suggest that a more systematic use of a two-tier gatekeeping system, emphasizing the gatekeeping role of CDUs, could have moderated the unintended negative effects of the waiting time target in a significant way.

From a broader perspective, our study also offers evidence that may contribute to our understanding of the unnecessary care phenomenon – which is estimated to account for as much as a third of health care spending in the US (Smith et al. 2012). Variation in expensive specialist services is often attributed to financial incentives of specialists or hospitals. Our results suggest that a combination of three non-economic factors – (1) high levels of diagnostic uncertainty, (2) shorter decision times as a consequence of system congestion, and (3) gatekeeper preferences for risk avoidance ("safety-first principle") – may also play an important part in explaining the overuse of expensive specialist services.

2. Literature Review

The research in this paper relates primarily to three main streams of literature: (i) work on gatekeeping and referrals within multi-tier service contexts, (ii) analytical studies of diagnostic processes, and (iii) empirical research on factors that impact on service performance.

Most relevant to our study is extant literature on gatekeeping systems. Such systems are comprised of two service tiers, with the server in the first tier referred to as a 'gatekeeper' because of the dual nature of their role, either 'self-treating' the customer or else, if too complex, referring them to a more costly but higher-skilled second tier 'specialist' (Shumsky and Pinker 2003). This service system has been studied mainly in the health economics literature – due to parallels with systems of referrals between primary and secondary/tertiary care – with a focus on the conditions under which gatekeeping systems are preferable to direct access and the design of contracts to reduce information frictions (González 2010, Mariñoso and Jelovac 2003, Malcomson 2004, Brekke et al. 2007). In the operations management literature, early modeling work has looked at how the system optimal rate of referrals between gatekeeper and specialist can be incentivized in both

deterministic (Shumsky and Pinker 2003) and stochastic (Hasija et al. 2005) settings. This modeling framework has been extended to investigate e.g. outsourcing contract decisions (Lee et al. 2012) and the performance of security-check queues (Zhang et al. 2011).

The gatekeeping literature abstracts away from the problem of identifying *which* customers to refer, focusing instead on the average rate of referrals assuming customers present with varying but orderable levels of complexity. If service times and/or quality vary with demand, however, then this may affect the accuracy of these referral decisions. A second body of research investigates such a possibility in service systems in which the quality of service is affected by it's duration. In work on the so-called 'speed-quality trade-off', Anand et al. (2011) and Kostami and Rajagopalan (2013) study pricing strategies in static and dynamic settings, respectively, in which the value of a service is increasing in the time that the service provider spends with the customer, but where this is also a cost to waiting. Complementary work explores the relationship between service configuration decisions and congestion/waiting times. Hopp et al. (2007), for example, find that increasing capacity may, in contrast to standard queuing results, increase congestion as a result of discretionary service components being added when servers are under light load. For expert services, for which customers are unable to accurately ascertain their service needs, Debo et al. (2008) demonstrate that queuing dynamics create heterogeneity in the customer base that can be exploited to induce additional service when arrival rates are low, with Paç and Veeraraghavan (2015) showing that congestion also acts as a deterrent to expert overtreatment. In contrast, we study this problem in a two-tier system and investigate instead the impact of service times on the classification process. We show that congestion may in fact *increase* expensive specialist overuse because greater diagnostic uncertainty leads to misclassification errors and servers referring customers unnecessarily to the specialist.

Another stream of research focuses specifically on the classification problem. Both van der Zee and Theil (1961) and Argon and Ziya (2009) examine customer classification policies when there exists imperfect information about customer type (e.g. refer versus self-treat). While the classification threshold affects error rates in these papers, misclassification is not inherently affected by service times or effort. Alizamir et al. (2013), on the other hand, also examines the process of customer type identification, but with a server who can perform additional diagnostic testing to resolve type uncertainty. The more tests they perform, the better the accuracy of diagnosis at a cost of increasing levels of congestion and waiting times for other customers. Similarly, Wang et al. (2010) study diagnostic centers in which servers trade-off the dual concern of accuracy and congestion given that misclassification costs are incurred by both the service provider *and* customer. They

find that increases in capacity may increase congestion, extending the result from the centralized system in Hopp et al. (2007) to the decentralized system. We also expect classification thresholds and errors to depend on congestion levels in our ED setting and study this behavior empirically. We differ, however, in that we (i) are interested in the behavior of the server in response to varying levels of diagnostic uncertainty, rather than the system optimal response, and (ii) also study a mechanism that can be implemented to reduce rates of errors when faced with type uncertainty.

Our work is also similar to research on resource pooling and partitioning, with the two-stage gatekeeping process conceptually similar to a two-priority queuing system for patients with high and low levels of diagnostic uncertainty. Results from queueing systems research suggest that streaming customers into different (priority) classes may be beneficial when customers differ sufficiently in their service requirements (see e.g. Mandelbaum and Reiman 1998, Dijk and Sluis 2008). While these queuing studies consider the streaming of customers based on processing times (see also Hu and Benjaafar 2009), other prioritization schemes exist, such as triage. Triage is a process used in EDs and other medical settings that prioritizes customers mainly based on levels of urgency (see e.g. FitzGerald et al. 2010, for an excellent overview of the history and process of triage). Recent studies of the triage process in the operations management literature have explored ways in which the basic triage process might be augmented, by e.g. segmenting patients along other dimensions. Chan et al. (2013), for example, develop an effective triage algorithm to allocate burn victims to burn-beds based on their expected duration of stay and comorbidity profile. Most relevant to our work is two modeling papers that look at the ED triage process: Saghafian et al. (2012) and Saghafian et al. (2014a). These propose augmenting triage by streaming ED patients based not only on their severity but also using their (i) likelihood of being admitted and (ii) their complexity (i.e. the likely duration of the diagnostic process), respectively. Although we also consider separating patients into different streams, we propose doing so instead based on residual uncertainty at the end of service, rather than observables at the start of service. Moreover, our outcomes of interest also differ, focusing instead on admission/discharge misclassification errors, rather than costs associated with long ED waits. A combination of these two approaches may, though, have further benefits.

Finally, our work relates to recent empirical studies of health care and other service settings which have looked into the impact of organizational factors such as workload on service outcomes (see Freeman et al. (2016) for a recent overview), for example clinical safety (Kuntz et al. 2014), service times (KC and Terwiesch 2009), reimbursement (Powell et al. 2012) and sales performance (Tan and Netessine 2014). Also related is work on patient routing, with Kim et al. (2014) and KC and Terwiesch (2012) showing, respectively, that high occupancy levels in the intensive-care

unit (ICU) can reduce rates of ICU admission and increase early discharge propensity. While these behaviors preserve/free up capacity in the resource-constrained and expensive ICU (i.e. the 'specialist' resource) for higher priority patients, we find that when the ED (the first gatekeeping tier) is crowded this pattern may be reversed, with instead *more* patients being referred into acute inpatient beds (the second-tier 'specialist' resource in our context). Freeman et al. (2016) find a similar result in a maternity context. In the first empirical analysis of the two-tier gatekeeping system, they demonstrate that midwives (gatekeepers) refer high complexity patients to obstetricians (specialists) at higher rates in the presence of congestion. In contrast, we focus instead on the effect of diagnostic uncertainty on both referral (admission) *and* self-treatment (discharge) errors, rather than the one-sided case, as well as exploring a possible preventative measure.

3. Decision Making and Uncertainty in the Emergency Department

The ED at the study hospital operates in a manner similar to the majority of hospitals in the US, UK and worldwide. After a patient arrives, they are registered and then assessed by a triage nurse and assigned a triage level based on the acuteness and severity of their condition. The patient then joins a queue in a waiting room, and waits to be seen for further assessment, diagnostic testing (e.g., x-ray, blood test, cardiac echo) and, if appropriate, treatment by a nurse (for a more "simple" patient) or, in most cases, an ED physician. Patients can present with a variety of complaints and symptoms, some of which can be easily handled in the ED (e.g., wound suturing, casting, splinting), while others are more complex (e.g., hip fracture, heart attack, multiple trauma) and require more specialized, longer-term care than the ED is equipped to provide. If after assessment the physician determines that the patient requires a level of care beyond that which they can provide in the ED then they can admit the patient to an acute bed in the hospital. Else, after treating the patient for their symptoms, the patient will be discharged home. ED physicians thus act as gatekeepers to expensive hospital inpatient beds, rationing access to the hospital by admitting only those patients whose needs can not be met in the less resource-intensive ED setting (Blatchford and Capewell 1997). This study focuses on the pattern of hospital admission (referral) and discharge (self-treat/non-referral) decisions made by physicians (gatekeepers) working in the ED of a large UK-based teaching hospital.

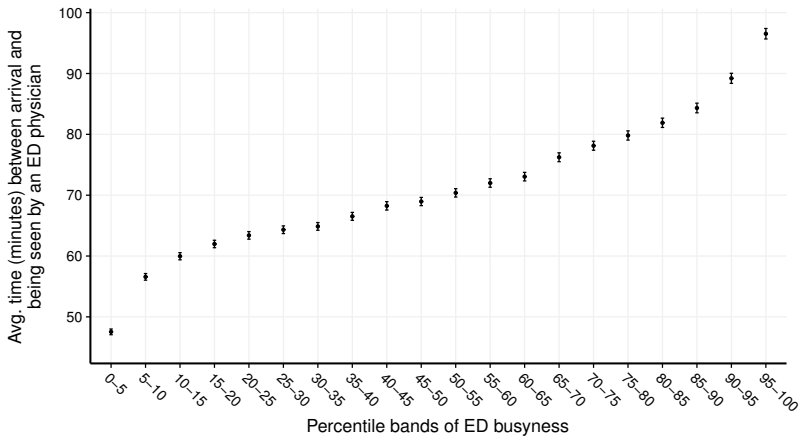
The ED is a highly time-pressured environment, with congestion and delays in care associated with e.g. higher complication rates and increased mortality (Bernstein et al. 2009, Huang et al. 2010, Sun et al. 2013). At the same time, there is an upwards global trend in ED attendances and ED crowding continues to worsen (Pines et al. 2011). In the US, for example, ED visits between 1997 and 2007 grew at almost twice the rate of population growth (Tang et al. 2010), while in

England between 1997 and 2012 ED admissions grew by 47% compared to population growth of 10% over this period (NAO 2013). In fact, the ED is now the primary point of entry to the hospital, admitting more than half of non-obstetric cases (Greenwald et al. 2016). Because of this, mitigating ED crowding is a significant policy concern and countries have adopted a wide range of interventions designed to manage this problem. Examples include telephone advice centers, implementation of fast tracks, increases in capacity and staffing, changes in boarding practices, and, most relevant to our study, the use of observation units and clinical decisions units (for an overview of the various approaches adopted in different countries see Pines et al. 2011). Yet these approaches have met with only limited success; February 2016 statistics from England, for example, revealed that only 87.8% of patients were admitted, transferred or discharged within four hours of their arrival at the ED – significantly lower than the target of 95% and the lowest rate since records began (NHE 2016).

As a consequence of growth in demand, ED physicians must increasingly make treatment and referral decisions under significant time and workload induced pressure. To demonstrate the impact of ED congestion on service times, in Figure 2 we have plotted for our study hospital the mean time between ED arrival and a patient being first seen by an ED physician. Each point in the plot corresponds to one of 20 percentile bands of ED busyness of width 5%. (Note that ED busyness is adjusted for differences across time of day and for various other time-related factors using a method described later in Section 4.4). As the ED becomes busier, the time between a patient's arrival and their first being seen by a physician increases also, and approximately (approx.) doubles from under 50 minutes to over 95 minutes when comparing the first and last percentile bands. Given that 95% of patients in our study hospital must be out of the ED within four hours of arrival (with failure to achieve this in any month attracting a fine of £200 per breach (NHS 2013)), this delay in the start of treatment has the effect of reducing time available to spend with each patient. The effect is also surprisingly large: average available service time is shortened by nearly 25% between the first and last percentile bands, falling from approx. 190 minutes to approx. 145 minutes. A natural question then is to ask what the consequence of this shortening of service times is on referral behavior in gatekeeping contexts such as this.

One characteristic of the ED context that might drive differences in outcomes as service times are compressed is the existence of high levels of clinical uncertainty and variation in diagnostic accuracy in emergency medicine (Sklar et al. 1991, Green et al. 2008). When service times are reduced, physicians have less time available to spend with each patient to perform diagnostic testing and to acquire the information necessary to make accurate and informed gatekeeping decisions

Figure 2 Mean time between patient arrival at the ED and being seen by an ED physician, by five-percentile bands of ED busyness, with 95% confidence bands.



(Smith et al. 2008, Alizamir et al. 2013). Decision density is also high, with ED physicians often caring simultaneously for multiple patients, which can lead to elevated cognitive loading. As a consequence they must regularly rely on heuristics and intuition, such as pattern recognition and rule-based decision-making, when assessing patients' needs (see Croskerry (2002) for an excellent overview of the types of heuristics employed by ED physicians). Significant time pressure and resource constraints prevalent in many EDs means that these cognitive shortcuts can result in higher than desired levels of costly but preventable errors (Leape 1994). For example, in a study of 100 cases of diagnostic error, Graber et al. (2005) found that cognitive factors contributed in 74% of cases. Two particularly costly errors in the gatekeeping context are referrals errors (type II errors) and self-treatment errors (type I errors). Inappropriate referrals use expensive inpatient beds that might otherwise be used for other types of activity (e.g. planned procedures); inappropriate discharges can lead to patients returning in a worse health state requiring more costly treatment and, potentially, the payment of compensation. As diagnostic uncertainty increases and servers have less information available to make gatekeeping decisions, therefore, we hypothesize higher rates of both of these types of error.

Hypothesis 1. *As service times decrease and the level of diagnostic uncertainty increases, gatekeepers make more errors in their referral decisions, i.e. they (i) refer more customers to the specialist and (ii) attempt to self-serve customers who they would have otherwise referred.*

As an alternative to Hypothesis 1, as diagnostic uncertainty increases the referral threshold may be adjusted to prevent an increase in one type of error at the expense of a higher rate of the other. Previous modeling literature suggests that such an effect may occur if the cost incurred by the customer and/or the service provider from the 'protected' error is significantly greater than that

of the 'sacrificed' error (Alizamir et al. 2013, Wang et al. 2010, Zhang et al. 2011). In gatekeeping settings, in addition to the provider and customer, the server may also associate costs with each of these types of error. Since the ultimate referral decision is made by the gatekeeper and not by the provider or the customer, the gatekeeper's chosen referral rate and the first-best referral rate for the service provider/customer may differ greatly. While various contracts have been proposed to align gatekeeper incentives with those of the provider (see Section 2), these neither adjust for the gatekeepers perceived or realized misclassification costs nor are such contracts typically used in practice. For example, ED physicians in the study hospital are salaried employees and their wages are not affected by the decisions that they take. There may, though, be other factors that these physicians consider when deciding whether to admit or discharge a patient. For example, medical errors have been shown to have a negative emotional impact on physicians (Christensen et al. 1992), can result in malpractice investigations and/or litigation (Studdert et al. 2006), and can also lead to reputation damage and peer disapproval (Leape 1994). The costs (whether financial or otherwise) that physicians associate with these concerns will affect how they respond to uncertainty.

In practice, asymmetry in error rates does exist; over-referrals occur more frequently that under-referrals (Bunik et al. 2007), with medical professionals having been shown to increasingly refer patients for higher intensity care when they perceive a risk (e.g. of litigation) from undertreatment (Shurtz 2013). Moreover, the 'overtreatment' phenomenon in health care suggests that medical professionals, when faced with uncertainty, will more often than not choose to do more rather than less (Gawande 2015). An extensive body of medical literature has also explored how physicians' attitudes toward risk and uncertainty affect resource use. In general, this finds that physicians act to reduce their feelings of uncertainty in clinical settings by e.g. ordering more diagnostic tests or prescribing multiple medications (McKibbon et al. 2007). More risk avoiding physicians have also been found to have e.g. lower primary care referrals (Franks et al. 2000), admit fewer patients from the ED to hospital (Pearson et al. 1995), and have overall lower costs of patient care (Allison et al. 1998, Fiscella et al. 2000). In the presence of risk avoiding gatekeepers and when there is a high cost of 'missed' diagnosis, therefore, we expect the gatekeeper's referral behavior to adjust to any increase in uncertainty in such a way so as to avoid additional under-referral errors.

HYPOTHESIS 2. *As service times decrease and the level of diagnostic uncertainty increases, if the cost to the gatekeeper of a non-referral error is perceived to be significantly higher than a referral error, then gatekeepers will (i) refer more customers to the specialist, resulting in (ii) no change or even a reduction in the number of self-service errors.*

A reduction in self-service errors would suggest an overreaction to the increase in diagnostic uncertainty: not only do they admit a large proportion of those additional patients with uncertain diagnosis, but they also admit more patients who they would have previously been willing to discharge. Such an overreaction would suggest that physicians in the ED have low risk-tolerance and weigh the cost of a non-referral error significantly higher than that of a referral error. This behavior would greatly increase the overuse of expensive specialist services at a high cost the provider.

Before moving on to describe our data and model set-up, one further point deserves attention. While we are interested in the effect of shortening service times on physician's referral decisions, capturing this using the time between a patient arriving and their being seen by a physician (e.g. as per Figure 2) is problematic. In particular, there will undoubtedly be many factors that we are unable to control for but that are correlated both with the time that it takes for a patient to be seen and with the decision of the physician (e.g., acuteness of their condition, medical history, the range of complications, etc.). This makes identification of a causal relationship challenging. Instead of this, therefore, we will use the busyness level of the ED as a proxy for the level of clinical uncertainty. This works because ED congestion and service times are (negatively) correlated (as shown in Figure 2), and so as ED congestion goes up we would expect ED physicians to be forced to make decisions with increased uncertainty (as they have less time available per patient for assessment, testing, and diagnosis). At the same time, since patients arrive for the most part at random, and there is no way for them to know in advance of arrival how busy the ED will be, there is little reason to suspect that patients will differ based on unobservable factors. One complication, however, is that evidence in the medical and operations literature has found (see Section 2) service quality and outcomes to deteriorate at higher workload levels. Thus, we need to be sure that any change in error rates is not simply a consequence of physicians becoming more error prone when making referral decisions under congestion. If this were the case, we would expect to see not only higher rates of admission and discharge errors but also higher rates of other types of admission error. Thus, we will also explore changes in specialty transfer error rates – which occur when patients are admitted to the incorrect medical area and must be subsequently transferred – though we make no apriori assumptions as to the direction of these effects, if at all significant.

Hypothesis 3. *As busyness levels for the gatekeeper increase and service times decrease, the rate of specialty transfer errors may increase, stay the same, or decrease.*

4. Data Description and Variable Definitions

The data for our study is comprised of detailed information relating to 651,044 ED attendances over a period spanning seven years from December 2006 through December 2013, as well as matching

inpatient records for all of those patients admitted from the ED into the hospital during this period. (All 8,527 observations (obs.) from the final month, December 2013, are dropped since data entry may not have been completed fully.) The ED we study is the largest in the region and has experienced increasing demand pressure over recent years, with attendances up by 4.2% year-on-year from 215 ED visits per day on average in the first year of our sample to 274 per day in the final year. On average 29.1% of patients who arrive at the hospital are admitted to an inpatient bed, with admissions and discharges increasing at approx. the same rate over the sample period (by 4.7% per annum (p.a.) for admissions versus 4.1% for discharges).

In order to prepare the data for analysis, we perform an initial cleaning round to ensure, as far as is possible, that our results are not affected by various data or time-related confounds. This includes dropping a small proportion ($< 2\%$) of obs. with missing data, excluding the first year of data so that it can be used generate a number of variables of interest, taking out dates close to public holidays when demand and staffing patterns vary significantly, dropping obs. for patients who left again medical advice, died in the ED or were transferred to another hospital, and excluding all patients treated by ED nurses rather than physicians. This process is described in full in Appendix A. After this, we are left with 429,313 observations to take forward for analysis. While we present findings using this cleaned data set, all results continue to hold when using the full sample.

We next describe the main variables used in the analysis. Summary statistics for these variables and correlations between them can be found in Table 1.

4.1. Referral to the CDU

Although in the first part of this study we are interested specifically in those referral decisions made directly by a physician in the ED, it is important that we account for the existence of the other option available to the ED physician: passing the patient on to the CDU. To see why, observe that to determine how physicians respond to increased uncertainty requires us to study only the top half of the two-stage gatekeeping process shown in Figure 1 (i.e. only those patients not passed to the CDU). However, as operating conditions in the ED change (e.g. busyness levels), so too might the rate at which ED physicians leverage the CDU option. Thus, despite only 8.2% of patients being passed to the CDU, we note that it will be necessary to ensure that our findings are not confounded by differences in patient case-mix arising from changes in CDU usage. (The method for doing so is described later in Section 5.1.) As the CDU itself is not at this stage of primary interest, we leave the discussion of how this unit operates to Section 6.1. For now, it is important to know only that at the end of assessment in the CDU the same two options exist: to either refer the patient into an acute inpatient bed or else discharge them.

Table 1 Descriptive statistics and correlation table.

		Mean			Correlation table			
	N	All	CDU = 0	CDU = 1	(1)	(2)	(3)	(4)
(1) Admission error (%)	429,313	4.71	4.68	5.02				
(2) Discharge error (%)	429,313	0.95	0.87	1.83	-0.02^{***}			
(3) Specialty change (%)	125,228	22.01	22.54	17.12	-0.09^{***}	N/A		
(4) CDU admission (%)	429,313	8.17	0.00	100.00	0.00^{**}	0.03^{***}	-0.04^{***}	
(5) ED busyness	429,313	0.00	0.00	−0.01	0.01^{***}	-0.01^{***}	0.01^{**}	-0.00^{*}

Notes: Columns 'All', 'CDU = 0' and 'CDU = 1' report mean values for the full sample, subsample of patients referred directly from the ED, and subsample referred from the CDU, respectively; Standard deviation of ED busyness equal to 1.00, 1.00 and 1.02 for 'All', 'CDU = 0' and 'CDU = 1', respectively; Correlation coefficients significant with $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

4.2. Admission and Discharge Errors

Turning next to the dependent variables of interest in our analysis, the first two described here capture errors made in referral (admission) and non-referral (discharge) decisions by ED physicians.

An admission error (or 'false admission') occurs when a patient is admitted to an acute hospital bed despite that admission being unnecessary or excessive to their needs. These patients block beds and use expensive specialist resources and time, with unnecessary hospital admissions estimated to have cost the NHS in England over £600 million in the 2012-13 financial year.[2] A patient is classed as an admission error (or 'false admission') if within 24 hours of being admitted to the hospital from the ED or CDU they are discharged with no treatment or procedure performed on them. The second of these conditions is met if a patient has no Classification of Interventions and Procedures OPCS-4.6 (HSCIC 2013) code – the UK equivalent of the American Medical Association's CPT coding system – associated with their post-admission inpatient record. The average rate of admission errors for the full sample of 429,313 visits is 4.7% and for the 125,228 visits which resulted in admission is 16.1%. There is evidence that at some of these types of admissions may be avoidable, e.g. Denman-Johnson et al. (1997) estimates that approx. 10% of ED admissions to hospital for short term care could be avoided, while it has also been suggested that increased imaging in EDs could prevent around 16% of admissions (Burgess 1998, Cooke et al. 2003).

A discharge error (or 'false discharge'), on the other hand, occurs when a patient who should have been admitted to the hospital is instead discharged from the ED. These patients often come back in a more serious state, requiring a higher intensity of care than would otherwise have been needed if correctly admitted. Pope et al. (2000), for example, found risk-adjusted mortality for patients with acute myocardial infarction who were inappropriately discharged from the ED to be 1.9 times higher than for hospitalized patients. A patient is recorded as a discharge error if after discharge

[2] Authors' calculations based on total hospital spend in 2012-13 of £12.5 billion on ED admissions (NAO 2013), with 49% of ED admissions staying less than 48 hours (NAO 2013), and estimated 10% of ED admissions for short-term care being avoidable (Denman-Johnson et al. 1997).

from the ED or CDU they re-attend the ED within 7 days and are at that point admitted to an inpatient bed in the hospital. The rate of discharge errors in the full sample is 0.9% and is 1.3% for the subset of 304,085 discharged patients. Note that the high rate of admission errors relative to discharge errors is already suggestive of physicians overweighing the low probability of a discharge error and taking the cautious approach of admitting patients when faced with uncertainty.

While not all patients we class as admission errors and discharge errors may be true errors (e.g., a patient may require admission for observation according to medical guidelines, or a patient may be discharged and re-visit the hospital for a problem unassociated with their initial visit), our study investigates how misclassification rates change under different organizational conditions.

4.3. Routing errors

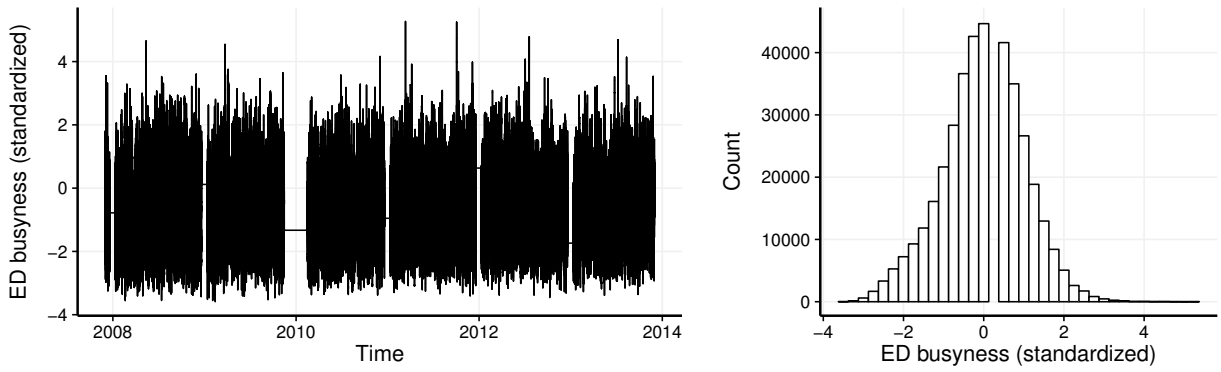
Our third dependent variable measures whether or not a patient is routed to the correct medical specialty when admitted to the hospital. This allows us to examine those factors that impact on the accuracy of specialist referral decisions (i.e., referral to the correct type of specialist). Patients who are transferred between medical units have been shown to experience delays in access to care, longer lengths of stay, and worse medical outcomes such as higher mortality (Beckett et al. 2013). We capture this using a binary variable that takes value one if the patient is transferred between medical specialties (e.g. between gastroenterology and endocrinology) within the first seven days after admission from the ED or CDU and zero otherwise. Note that this variable can only be calculated for the subsample of 125,228 patients who were admitted to the hospital.

4.4. ED Congestion

In order to measure how ED physicians respond when faced with increased diagnostic uncertainty arising due to congestion in the ED, we need a variable that captures ED busyness.

To generate this measure, we first determine which patients' ED visits overlapped with the period from arrival to one hour post-arrival of patient i, and calculate the sum of those overlapping periods $tOverlap_{-i}$. We then divide $tOverlap_{-i}$ by the total number of overlapping patients to get the time-weighted average busyness of the ED over the first hour after the arrival of patient i and denote this $QueueED_i$. It is well known that busyness levels in EDs vary across the day, on weekdays and weekends, in different seasons, and change over time. Since some of this is predictable and staffing can be partially set to meet demand, we will adjust $QueueED_i$ also to account for these differences. We achieve this by employing a variation on the approach used in Kuntz et al. (2014) and Berry Jaeker and Tucker (2016) which establishes an approx. upper bound on the available capacity. We estimate this upper bound using quantile regression to predict the 95th percentile level of occupancy at hour h. The dependent variable in this regression is the time-weighted average

Figure 3 Plot of standardized ED busyness over time (left) with frequency histogram (right).



occupancy level over every hour h starting midnight on 1st January 2007 and ending midnight on 31st December 2013. (Note that all dates dropped during the data cleaning process, as described in Appendix A, are also removed here.) We estimate this model with independent variables: (i) year, (ii) quarter of the year, (iii) time, split into six four-hour windows per day (e.g., midnight to 4a.m., etc.), (iv) a binary variable equal to one if a weekend and zero otherwise, (v) the interaction between (iii) and (iv), and (vi) the interaction between (v) and a binary variable equal to one if the date was between the years 2011 to 2013, and zero otherwise. The fitted values from this model then provide us with our estimate of capacity for each hour h, $QueueED_h^{95th}$. ED congestion, $OccED_i$, is then equal to $QueueED_i$ divided by $QueueED_{h_i}^{95th}$, where h_i is the hour of arrival of observation i. Finally, we normalize this by subtracting it's mean, $\mu(OccED_i)$, and dividing through by it's standard deviation, $\sigma(OccED_i)$, to form $zOccED_i$. Plots of $zOccED_i$ are provided in Figure 3.

4.5. Control Variables

In addition to the primary variables described above, we also have available and derive a large number of control variables that allow us to account for heterogeneity in the patient population and in the hospital that may be correlated with the dependent variables, and/or with the main independent variables of interest. These are reported in Table 10 in Appendix B, and capture temporal factors, differences in diagnosis and condition, contextual factors (e.g. arrival method), and attributes of the assigned physician. Any factors not reported in our data that might be correlated with the primary independent variables (and so through omission may bias the results) will be accounted for using appropriate empirical methods to be described in Section 5.1.

Two controls that will become important when discussing our empirical strategy that deserves mention here are variables for the historic admission and discharge error rate of the assigned physician. These account for the fact that particular physicians may be more or less predisposed to

making errors than others, and approx. speaking are calculated as the average case-mix adjusted rates of false admissions and discharge decisions made by each physician over the previous year (see Section 6.4 for a full description of the calculation of these variables).

5. Models and Results I: Response to Diagnostic Uncertainty

We are interested in how ED physicians respond when faced with additional uncertainty. As mentioned in Section 4.1, this requires us to initially study only those referral decisions made by ED physicians directly, i.e. not those cases referred into the CDU. Identification is complicated, however, by the fact that those patients who are passed to the CDU in our sample may be inherently different from those for whom the physician makes the referral decision themselves. While we account as far as possible for these differences with our set of controls (reported in Table 10), there may still exist factors unobservable to us, the researchers, but observable to the physician (e.g., fitness level, medical history) that influence whether or not the physician leverages the CDU option. Not accounting for this endogeneity could lead to biased coefficient estimates and invalidate our findings. In this section, we describe the empirical approach we adopt to resolve this.

5.1. Econometric Specification

Our empirical strategy separates the identification problem into two parts. The first looks to identify those factors that influence whether or not the patient is admitted into the CDU. The second determines whether or not a patient is admitted or discharged in error or referred to the wrong specialty while allowing this to depend on whether or not the patient was admitted to the CDU. More specifically, the first stage (selection) equation takes the form

$$CDU_i^* = \delta_0 + \mathbf{X}_i\boldsymbol{\delta}_1 + \mathbf{Z}_i\boldsymbol{\delta}_2 + zOccED_i\delta_3 + \epsilon_i^{\delta}\,, \tag{1}$$

$$CDU_i = \mathbb{1}[CDU_i^* > 0]\,, \tag{2}$$

where $\epsilon_i^{\delta} \sim \mathcal{N}(0,1)$, CDU_i^* is a latent variable, the vector $\mathbf{X}_i$ contain the set of all controls (reported in Table 10), the vector $\mathbf{Z}_i$ contains the set of instrumental variables (to be described in Section 5.2), CDU_i is the observed dichotomous variable that indicates whether the patient was sent to the CDU, and $\mathbb{1}[\cdot]$ is the indicator function. The second stage (outcome) equation takes the form

$$AdmErr_i^* = \beta_0 + \mathbf{X}_i\boldsymbol{\beta}_1 + CDU_i\beta_2 + zOccED_i\beta_3 + \epsilon_i^{\beta}\,, \tag{3}$$

$$AdmErr_i = \mathbb{1}[AdmErr_i^* > 0]\,, \tag{4}$$

where $\epsilon_i^{\beta} \sim \mathcal{N}(0,1)$, and where $AdmErr_i^*$ and $AdmErr_i$ are the latent and observed variables for admission errors, respectively. The latent variable equation for discharge errors is identical to that for admission errors, with coefficient vector $\boldsymbol{\beta}$ replaced with $\boldsymbol{\alpha}$.

When the dependent variable of interest is specialty transfer we use a different vector of controls. Specifically, we replace coefficient vector $\boldsymbol{\beta}$ with vector $\boldsymbol{\gamma}$ we also replace control vector $\mathbf{X}_i$ with $\mathbf{W}_i$, which includes all of the controls in $\mathbf{X}_i$ as well as: (i) a categorical (to allow for non-linearity) control equal to the number of days, up to a maximum of seven, that the patient stayed in the hospital after admission from the ED or CDU, (ii) a control for the age of the patient (using fifteen-year age bands), and (iii) a control for the specialty transfer rate of the assigned physician, similar to the admission and discharge error rates used as a control and described in Section 4.5. Note that the additional control for hospital length of stay up to seven days (which recall is the number of days we measure specialty changes over) accounts for the fact that the longer a patient stays in the hospital the more likely they are to change specialty. The estimations for specialty transfer can also only be run on the subsample of admitted patients (allowing us to introduce age as an additional control), since a transfer can only occur if a patient is admitted.

Rather than estimate the first and second stage models described above individually, instead, we estimate them simultaneously with a Heckman probit sample selection (heckprob) model using full information maximum likelihood (Maddala 1983). The heckprob model allows us to estimate the effect that ED congestion has on our outcomes for only those patients who were admitted or discharged directly by an ED physician (rather than by a physician in the CDU) – which is the effect we are interested in – while also allowing us to account for the fact that the rate of referrals into the CDU may also differ as the ED becomes congested. To achieve this we censor the outcome variable $AdmErr_i$, $DischErr_i$ or $SpecChg_i$ whenever $CDU_i = 1$, set $\alpha_2, \beta_2, \gamma_2 = 0$ in the outcome equation, remove ED length of stay (see Table 10) from the control vectors $\mathbf{X}_i$ and $\mathbf{W}_i$ (since ED busyness is very likely to affect length of stay in the ED), and then estimate the selection and outcome equations simultaneously under the assumption that their errors $(\epsilon_i^\delta, \epsilon_i^\alpha)$, $(\epsilon_i^\delta, \epsilon_i^\beta)$ or $(\epsilon_i^\delta, \epsilon_i^\gamma)$ are jointly distributed according to the standard bivariate normal distribution with unit variances and correlation coefficients ρ^α, ρ^β or ρ^γ which are estimated as parameters in the models.[3] We claim that ED physicians adjust the rate at which they admit patients to the hospital (rather than

[3] Traditionally, Heckman sample selection models are used when the outcome is not observed in the case of non-selection (for example, if we had no further information about those patients admitted to the CDU). In our case, however, we observe the outcome both when the ED physician makes the referral decision and when it is made in the CDU. It is possible, therefore, for us to estimate the coefficients under both regimes (i.e., when the referral decision is made by either the ED or a CDU physician). This estimation can be made jointly using an endogenous switching regression model, or instead by estimating both sides of the equation separately by "tricking" the Heckman selection model to do so, as described in Lee (1978). We employ this trick by censoring the dependent variable in the outcome equation ($AdmErr_i$ or $DischErr_i$) depending on whether CDU_i takes the value zero or one. Censoring when $CDU_i = 1$ allows us to estimate the effect of ED busyness on error rates made by ED physicians, while censoring when $CDU_i = 0$ allows us to estimate the effect on decisions made in the CDU instead. Joint estimation (not reported) results in nearly identical estimates of the coefficients and ρ.

simply making more mistakes in general) when faced with higher levels of diagnostic uncertainty due to shortening service times if as the system becomes more congested there is an increase in the rate of false admissions (i.e., $\beta_3 > 0$) without a similar increase in the rate of false discharges (i.e., $\alpha_3 \leq 0$) or referrals to the wrong specialty (i.e., $\gamma_3 \leq 0$).

5.2. Instrumental Variables

While the heckprob model can be estimated without instrumental variables (IVs), estimation is improved and coefficients more reliable when IVs are provided (Wilde 2000, Maddala 1983). These IVs should affect the CDU admission decision, and so appear in the selection equation (i.e., are relevant), but not affect the rate of admission errors, discharge errors or the likelihood of a patient transferring specialty, and so do not appear in the outcome equation (i.e., are valid). We use two IVs, included in the vector $\mathbf{Z}_i$. Summary statistics for these IVs are available in Table 2.

The first IV is the CDU admission propensity of the assigned physician. This is calculated in the same way as the physician's admission and discharge error propensity (as discussed in Section 4.5 and described later in more detail in Section 6.4), and approx. speaking is equal to the physician's average rate of CDU referrals over the previous twelve months relative to the rate expected given the case-mix of patients they treated. A patient assigned to a physician who is more predisposed to admit patients to the CDU will be more likely to be sent there themselves, satisfying the relevance condition. Furthermore, since we already control for the physician's admission, discharge and, where relevant, transfer propensity in the selection and outcome equations (see Table 10), the physician's predisposition to admit patients to the CDU should not affect the error rates other than through the CDU admission decision itself, satisfying the validity condition.

Our second IV is the busyness of the CDU. Congestion in the CDU, $zOccCDU_i$, is calculated in the same way as was ED busyness in Section 4.4, except that we time-weight instead over the one hour period leading up to the departure of patient i from the ED. If the CDU is congested then it becomes less available to ED physicians as an option, since beds and other resources are constrained. This is similar to findings in the literature relating to e.g. admission to the intensive care unit (Chan et al. 2016) and obstetric operating theaters (Freeman et al. 2016). Thus we expect when the CDU is busy there to be fewer CDU admissions, satisfying the relevance condition. For patients who are not admitted to the CDU, the busyness of the CDU should have no direct effect on their likelihood of being admitted or discharged in error or to experience a specialty transfer – and to the extent that CDU busyness is correlated with busyness in the main hospital, we control for this using the occupancy level of the hospital (calculated in the same way as CDU busyness). For patients who are admitted to the CDU, it is possible that admission and discharge decisions made

Table 2 Descriptive statistics and correlation table for the instrumental variables.

		Mean			Correlation table				
	N	All	CDU = 0	CDU = 1	(1)	(2)	(3)	(4)	(5)
(6) CDU busyness	429,313	12.84	12.97	11.31	0.07^{***}	0.01^{***}	-0.04^{***}	-0.01^{***}	0.03^{***}
(7) Phys. CDU use	429,313	0.00	0.01	−0.06	0.01^{***}	−0.00	0.01^{***}	-0.02^{***}	0.17^{***}

Notes: Columns 'All', 'CDU = 0' and 'CDU = 1' report mean values for the full sample, subsample where $CDU_i = 0$ and subsample where $CDU_i = 1$, respectively; (1) Admission error, (2) Discharge error, (3) Specialty transfer, (4) CDU admission, (5) ED busyness; Correlation coefficients significant with $^{***}p < 0.001$, else $p > 0.05$.

in the CDU are affected when the CDU becomes busier, which might impact on error rates. To account for this, we include in the selection and outcome equations a variable that takes value zero when the patient is not admitted to the CDU and is equal to $zOccCDU_i$ otherwise.

Hypothesis testing of the IVs to identify whether there are signs of over-, under- or weak identification provide strong evidence the IVs are not invalid (p-values > 0.10), are relevant (p-values < 0.001), and achieve significantly less than 10% maximal relative bias, as desired (see Section EC.3 of the e-companion). Our results are also robust to the omission of CDU busyness as a second IV.

5.3. Results

Before presenting the full set of results, we start by reporting in Table 3 coefficient (coef.) estimates with robust standard errors using a standard probit estimation for each of the four dependent variables in the selection and outcome equations. Examining the model coefficients, we find evidence that as ED physicians become more busy, and hence have less time to spend with each patient so increasing diagnostic uncertainty, they (1) increase the rate at which they refer patients to the CDU (coef. $= 0.067$, p-value < 0.001), (2) make more admission errors (coef. $= 0.025$, p-value < 0.001), and (3) make fewer discharge errors (coef. $= -0.016$, p-value $= 0.033$), with (4) no change in the probability that a patient is assigned to the wrong specialty (coef. $= 0.002$, p-value > 0.10). These responses to increasing levels of diagnostic uncertainty are consistent with Hypothesis 2, i.e. that physicians become more cautious and admit more patients to the hospital than need to be there, rather than Hypothesis 1, i.e. that they generally become more error-prone. In the rest of this section, we investigate our hypotheses using the empirical strategy outlined in Section 5.2.

Given that ED busyness is significant in the selection equation (model (1) of Table 3) we must correct with the heckprob models for potential endogeneity to ensure that the coefficient of ED busyness in the outcome equations are not biased by this. Heckprob model coefficients are reported in Table 4. In heckprob (1e), (2e) and (3e) we identify the effect of ED busyness for only the subset of patients for whom the referral decision is made directly by an ED physician, i.e. censoring when $CDU_i = 1$. For completeness, in heckprob (1c), (2c) and (3c) we report this instead for only those patients admitted to the CDU, i.e. censoring when $CDU_i = 0$. Heckprob (1e) shows evidence of

Table 3 Base coefficient estimates using probit model specification.

	(1) CDU	(2) AdmErr	(3) DischErr	(4) SpecChg
ED busyness	0.067^{***}	0.025^{***}	-0.016^{*}	0.002
	(0.004)	(0.005)	(0.008)	(0.006)
CDU referral	–	-0.150^{***}	0.244^{***}	-0.175^{***}
		(0.015)	(0.020)	(0.016)
CDU busyness	-0.035^{***}	–	–	–
	(0.002)			
Phys. CDU rate	0.827^{***}	–	–	–
	(0.022)			
N	429,313	429,313	429,313	125,228
Log-lik	−94,213	−65,704	−21,432	−53,054
Pseudo-R^2	0.224	0.193	0.068	0.196

Notes: All estimations made using a probit model specification; *Robust standard error* in parentheses; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

significant negative selection ($\rho = -0.420$, p-value < 0.001), meaning that patients who are not selected for admission to the CDU are less likely to be a false admission than a patient selected at random from the population, verifying the need to account for endogeneity.

After correcting for endogenous selection, we find evidence consistent with that of probits (2), (3) and (4) in Table 3. In particular, evidence from Table 4 suggests that when the ED is more busy ED physicians are significantly more likely (coef. $= 0.036$, p-value < 0.001 in heckprob (1e)) to admit patients to the hospital who do not require hospitalization. At the same time, ED physicians become less likely (coef. $= -0.016$, p-value $= 0.056$ in heckprob (2e)) to discharge patients in error when the ED becomes busy, and no more likely (coef. $= 0.010$, p-value > 0.10 in heckprob (3e)) to admit patients to the incorrect specialty. All of this evidence is consistent with ED physicians overcorrecting for the increased risk of a discharge error when clinical uncertainty rises by increasing the rate of at which they admit these uncertain cases. In particular, as fewer false discharge errors are made and there is no change in specialty routing errors, this is strongly indicative of the fact that ED physicians are not simply becoming more error prone as they become busy (since we would expect a similar increase in both of these types of errors). This supports Hypothesis 2.

Interestingly, the presence of the CDU appears to shelter the system from some of the effects of ED busyness: as the ED becomes busier, more patients are admitted to the CDU (coef. $= 0.067$, p-value < 0.001 in probit (1) from Table 3), and patients admitted to the CDU are unaffected by busyness in the ED (coef. $= 0.018$, p-value > 0.10 in heckprob (1c)). This indicates one potential benefit of decoupling the gatekeeper's referral decision: the additional service layer can act as a workload buffer for the gatekeeper. This finding is consistent with existing literature, with e.g. Freeman et al. (2016) showing that as midwives (the gatekeepers in their context) become busier, they increase the rate at which they refer patients to obstetricians (the specialists in their context).

Table 4 Coefficient estimates to establish ED physicians' response to increased uncertainty, using heckprob model specification.

	Decision made by ED physicians			Decision made in the CDU		
	(1e) AdmErr	(2e) DischErr	(3e) SpecChg	(1c) AdmErr	(2c) DischErr	(3c) SpecChg
ED busyness	0.036^{***}	$-0.016^{\dagger}$	0.010	0.019	−0.019	0.006
	(0.005)	(0.008)	(0.007)	(0.017)	(0.022)	(0.037)
ρ	-0.420^{***}	0.180	−0.148	0.092	0.047	0.272
	(0.035)	(0.133)	(0.194)	(0.127)	(0.171)	(0.520)
N	429,313	429,313	125,228	429,313	429,313	125,228
N uncensored	394,225	394,225	112,918	35,088	35,088	12,310
Log-lik	−152,824	−112,683	−83,293	−100,548	−103,185	−41,291

Notes: All estimations made using the heckprob model specification; *Robust standard error* in parentheses; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

To give an idea of the scale of the effects, we convert coefficient estimates into average partial (marginal) effects (APEs) with 95% confidence intervals (CI_{95}s). A one standard deviation increase in ED busyness increases the probability of admission to the CDU by 0.79%, $CI_{95} = (0.70\%, 0.88\%)$, of being admitted by an ED physician in error by 0.36%, $CI_{95} = (0.26\%, 0.46\%)$, and decreases the probability of being discharged in error by -0.03%, $CI_{95} = (\text{-}0.07\%, 0.00\%)$. Compared with the average rate of CDU use, false admissions and false discharges reported in Table 1, this represents a relative increase (decrease) of approx. 9.7%, 7.6% and -3.5%, respectively. Thus moving from a low to high busyness state in the ED, i.e. from -2σ to $+2\sigma$, will have a surprisingly large impact, especially on CDU use and false admission rates. For example, assuming a cost of £500 per false admission, if all 651,044 patients had been treated in the ED in a high busyness state rather than low then over-referral by ED physicians would have cost the hospital approx. £4.7 million more.

5.4. Robustness to Endogeneity Concerns

While in the above we have argued that the increase in false hospital admissions and decrease in false discharges is an indication of ED physicians becoming more cautious and over-admitting patients when faced with increasing levels of diagnostic uncertainty, an alternative explanation could be that as the ED becomes busier the risk profile of the patients increases, e.g. if more complex cases arrive, or if more simple cases are instead seen by ED nurses. This then might necessitate an increase in hospital admissions by ED physicians, and hence cause the higher false admission rate. First, we note that this is unlikely since if patients were becoming riskier then we would also expect an increase in the rate of false discharges also, which we do not find. However, to address this concern more robustly, in Section EC.2 of the e-companion we (i) demonstrate that, based on observables, patients do not appear to differ in their false error propensity as the ED becomes more congested, and (ii) use an instrumental variable approach, with ED busyness from the previous week as an IV, to demonstrate that our results hold up even after accounting for potential correlation between ED busyness and the error terms.

6. The Two-Stage Gatekeeping System

Having established that physicians in the ED make overly cautious decisions and over-admit patients to expensive acute inpatient beds when faced with diagnostic uncertainty – at a significant cost to the provider – we next look at approaches that might be taken to mitigate this effect. Given that we have identified a cause of this to be high levels of uncertainty in diagnosis, interventions that act to reduce this uncertainty should improve performance. Two proposed suggestions for achieving this are: (i) to replace existing gatekeepers with those who are more experienced/skilled, and (ii) to increase the time available for diagnosis by increasing capacity. While both of these approaches would improve the accuracy of diagnosis, they each come with a cost: more experienced servers demand higher wages, while increasing capacity requires e.g. the hiring of more staff. In addition, it is not immediately clear that these changes would have as much of an impact as desired: the more experienced gatekeepers would spend a high proportion of their time with customers for whom the referral/non-referral decision was already unambiguous and could have been made just as effectively by less experienced and less costly servers; similarly, there is no guarantee that any increase in capacity would be used only to attend to those customers whose diagnosis is unresolved, as e.g. servers may add discretionary components to the service of the unambiguous customers (Hopp et al. 2007, Debo et al. 2008). A better approach, therefore, would be one that targets those more skilled gatekeepers and that additional capacity at those customers who would stand to benefit the most, i.e. those customers for whom there exist higher levels of diagnostic uncertainty. This is the idea behind the two-stage gatekeeping system.

The two-stage gatekeeping system enables gatekeepers to before making a referral decision judge whether sufficient information is available for accurate diagnosis and, if not, to pass the customer downstream to another gatekeeper who assumes responsibility for the referral decision (see e.g. Figure 1). The gatekeepers in this second service stage are more experienced than those in the first stage and are allocated more resources and time in order to resolve uncertainty in diagnosis. Since only those customers for whom the original referral decision is ambiguous should be passed to this second gatekeeping stage, the more experienced (and costly) servers will be expected to spend little of their time with unambiguous cases. Moreover, since capacity in the first service stage is left unchanged, the increase in time available for resolving uncertainty will be allocated specifically to those customers who stand to benefit the most (i.e. those passed to the second stage). So long as gatekeepers in the first stage refer to the second stage only those customers for whom there exists high enough classification uncertainty, then fewer errors should be made in referral decisions than if those patients had instead been referred directly by the first-stage gatekeeper. This is similar

in concept to that of complexity-augmented triage proposed in Saghafian et al. (2014b), which recommends first triaging patients who arrive at the ED on the relative complexity of diagnosis and then on their degree of urgency. Similarly, we expect that streaming patients based on residual uncertain should increase the overall accuracy of referral decisions.

HYPOTHESIS 4. *In the two-stage gatekeeping system, those customers referred through the second stage are significantly less likely to be referred in error than in the single-stage gatekeeping system.*

HYPOTHESIS 5. *In the two-stage gatekeeping system, those customers referred through the second stage are more likely to be referred to the correct specialist.*

Although one might expect dedicating additional capacity to reduce uncertainty should improve the quality of decision-making, a question of practical relevance is under what conditions the two-stage gatekeeping system is preferable to the single stage alternative. In this study, we investigate two such characteristics of multi-tier service systems that might affect the potential scale of the benefits of the two-stage system: (1) the prevailing levels of uncertainty surrounding the appropriate level of service in the setting being studied, and (2) the skill and accuracy of the first-stage gatekeepers when making classification decisions. We posit that the higher the underlying level of uncertainty in diagnosis, the more potential there is for the second-stage gatekeeper to add value. This is intuitive: if decisions are clear-cut and carry little ambiguity then the single-stage gatekeeping system is perfectly sufficient and capable of referring customers appropriately. As levels of diagnostic uncertainty increase, gatekeepers in the first-stage are increasingly likely to make incorrect decisions when routing customers, and so the value of a second opinion from a more experienced decision-maker increases. In our particular context, since we find that ED physicians over-refer patients to specialist acute inpatient beds in the presence of diagnostic uncertainty, we should expect to find a significant reduction in admission errors, thus we hypothesize:

HYPOTHESIS 6. *When first-stage gatekeepers are prone to over-referring to the specialist, the two-stage gatekeeping system is especially beneficial in reducing false referral errors when levels of diagnostic uncertainty are higher.*

With respect to the first-stage gatekeepers, we note that they are not homogenous in their skill or experience and that this made lead to discrepancies in service delivery, particularly for the type of knowledge work that characterizes many gatekeeping systems. EDs are typically staffed primarily by junior physicians, and are known to vary greatly in ability and confidence (Pinkney et al. 2016) and may "practise defensively and lack confidence to resist an admission." (Blatchford and Capewell 1997). In addition, these decision makers will vary in their attitudes towards risk

taking and uncertainty. Studies in the medical literature, for example, have shown that physicians' with higher risk taking scores admit more patients with chest pain from the ED (Pearson et al. 1995) and have higher patient care costs (Allison et al. 1998, Fiscella et al. 2000). We can thus think of approx. dividing the gatekeepers into one of four types: less error prone versus more error prone, and less risk taking versus more risk taking. The more error prone the first-stage gatekeepers, the more likely they are to make both referral and non-referral errors than their less error prone counterparts. The more risk taking the first-stage gatekeepers, the more likely they are to attempt to self-treat customers resulting in higher rates of self-treatment errors and lower specialist referral error rates. The less risk taking they are, the more likely they are to refer customers to the specialist who they might otherwise have successfully treated themselves, resulting in higher rates of referral errors and lower self-treatment errors. Thus, we would expect the second-stage gatekeeper to reduce variation in service delivery (i.e. variation across both types of error) caused by customers being seen by heterogeneous servers.

Hypothesis 7. *The two-stage gatekeeping system reduces heterogeneity in service quality (i.e. rates of referral and self-treatment error) that arises from customers being treated by first-stage gatekeepers who are different in their error propensity and tolerance to risk.*

In our specific context, this should translate to reduced physician-induced heterogeneity in error rates for those patients referred through the two-stage system. In particular, we should expect a significant reduction in admission error rates for less risk taking and more error prone physicians, and potentially a reduction in discharge errors for more risk taking and more error prone physicians (though identifying such an effect is likely to be empirically more challenging given the low rate of discharge errors relative to admission errors).

In the rest of this section, we first describe how the ED–CDU interaction operates like the two-stage gatekeeping system, before introducing the variables that are used to capture diagnostic uncertainty and the physician's profile. Summary statistics for these variables and their correlations with the main outcome variables from Section 4 are reported in Table 5.

6.1. The Clinical Decisions Unit

The clinical decisions unit (also known as an observation unit) is a dedicated area for emergency patients of low to moderate risk that exists separate to the main ED and general hospital units. The unit is designed to provide services such as further diagnostic evaluation, additional testing, and continuation of therapy for patients who require care beyond the initial level that can be provided in the ED (Ross et al. 2012). Patients admitted to the CDU are expected to have symptom complexes

Table 5 Descriptive statistics and correlation table.

		Mean			Correlation table				
	N	All	CDU = 0	CDU = 1	(1)	(2)	(3)	(4)	(5)
(8) Low diag. uncertainty (%)	429,313	25.00	26.91	3.55	-0.12^{***}	-0.04^{***}	-0.06^{***}	-0.15^{***}	-0.02^{***}
(9) Low-med. diag. uncertainty (%)	429,313	25.00	25.53	19.08	-0.10^{***}	-0.01^{***}	-0.06^{***}	-0.04^{***}	0.01^{***}
(10) Med.-high diag. uncertainty (%)	429,313	25.00	23.58	40.91	0.01^{***}	0.03^{***}	0.02^{***}	0.11^{***}	-0.01^{***}
(11) High diag. uncertainty (%)	429,313	25.00	23.98	36.46	0.21^{***}	0.02^{***}	0.04^{***}	0.08^{***}	0.01^{***}
(12) Less error prone phys. (%)	429,313	33.61	34.40	24.68	-0.08^{***}	-0.02^{***}	-0.01^{***}	-0.06^{***}	0.00^{*}
(13) More risk taking phys. (%)	429,313	25.80	25.15	33.10	-0.03^{***}	0.01^{***}	-0.03^{***}	0.05^{***}	-0.02^{***}
(14) Less risk taking phys. (%)	429,313	17.85	17.83	18.10	0.08^{***}	-0.01^{***}	0.08^{***}	0.00	0.03^{***}
(15) More error prone phys. (%)	429,313	12.84	12.97	11.31	0.07^{***}	0.01^{***}	-0.04^{***}	-0.01^{***}	0.03^{***}

Notes: Columns 'All', 'CDU = 0' and 'CDU = 1' report mean values for the full sample, subsample of patients referred directly from the ED, and subsample referred from the CDU, respectively; Standard deviation of ED busyness equal to 1.00, 1.00 and 1.02 for 'All', 'CDU = 0' and 'CDU = 1', respectively; Correlation coefficients significant with $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

that can be resolved within a six-to-24 hour period, with further assessment determining whether inpatient admission is required at the end of their CDU stay (Hassan 2003). These units also typically benefit from the presence of specialist trained and more senior staff, as well as advanced diagnostic capacity. As a consequence, various advantages of such units have been identified in the literature, such as improved patient satisfaction and safety and shorter stays (see Cooke et al. 2003, for an excellent survey). It is also thought that making greater use of decisions units can result in considerable cost savings, estimated in one study at $3.1 billion per year (Baugh et al. 2012). Thus, while generally it is believed that CDUs are an effective alternative to inpatient admission, to our awareness no studies have looked at the impact on transfer or discharge error rates, nor at those characteristics of patients and physician's that influence the CDU's effectiveness.

Of the 35,097 ED patient that end up in the CDU, 35.1% are subsequently admitted with the rest discharged home. Once a patient is in the CDU, decisions are made quickly, with a median CDU length of stay (LOS) of 4.5 hours for those who are subsequently admitted, and 4.0 hours for those who are subsequently discharged. This compares with a median LOS in an inpatient hospital bed of 14.8 hours for a patient classed as an admission error, suggesting that the CDU is able to more quickly process patients than can be achieved in a standard inpatient setting. Moreover, of those patients admitted only 14.3% are then identified to be admission errors, compared with 16.2% for those admitted directly from the ED. This is despite the fact that patients admitted from the CDU are those for who we anticipate there exists considerably more diagnostic uncertainty and hence should be inherently more likely to be admitted in error. Further analysis (documented in Section EC.1 of the e-companion) indicates that the CDU is conservatively around 42% faster in processing those patients routed through it than if instead they had been admitted to a hospital inpatient unit. Thus, while referral through the CDU does extend the service episode, this is by an amount less than if all patients were instead referred directly into the hospital. This is consistent with findings in the medical literature (e.g. Baugh et al. 2012).

Table 6 Table of conditions with total error and admission rates.

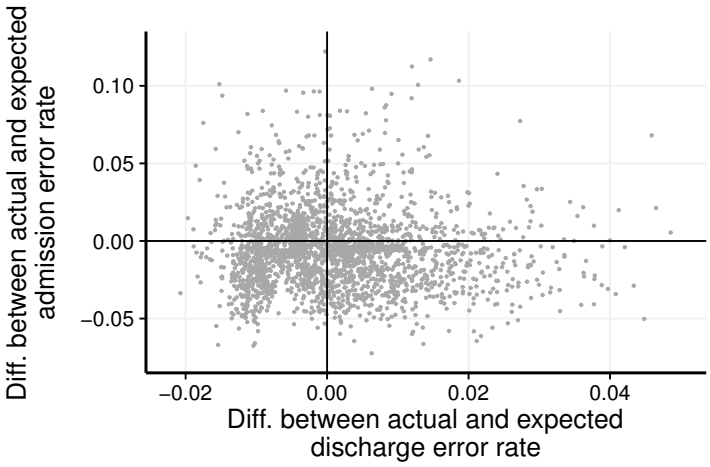
Condition category	N	% Error	% Admitted
Cardiac	37,892	17.0	59.6
Endocrinological (incl. diabetes)	2,619	13.3	73.2
Obstetrics and gynecological	7,108	13.2	28.0
Hematological	2,747	10.8	80.9
Gastrointestinal	44,878	10.4	54.6
Infectious disease	8,129	9.1	50.6
Respiratory	28,934	8.5	54.9
ENT	7,954	8.2	27.5
Genitourinary	17,796	7.0	53.0
Poisoning (incl. overdose)	5,948	6.8	13.4
Dermatological	9,230	5.8	27.3
Central nervous system (incl. stroke)	33,264	5.1	38.8
Rheumatological	4,538	5.0	25.6
Diagnosis not classifiable	29,742	5.0	25.9
Psychiatric	6,260	4.7	13.1
Minor wounds and injuries	173,709	1.0	8.3
Ophthalmological	8,565	0.7	1.6

6.2. Diagnostic Uncertainty

The degree of diagnostic uncertainty associated with a particular patient attendance is not observed but will be highly correlated with the patient's likelihood of being either admitted or discharged in error. In particular, patients for whom a referral is clearly necessary will have a low likelihood of being admitted as a false admission, and patients for whom the non-referral (i.e., self-treat and discharge home) decision is clear will have a low likelihood of being discharged as a false discharge. Therefore, we distinguish between the levels of diagnostic uncertainty associated with patients by using as a proxy their probability of being either admitted or discharged in error. One way to determine this probability it to classify patients based on the condition that they present with: for some conditions the appropriate course of action can be unambiguous (e.g., minor wounds can often be treated in the ED and the patient safely discharged home), while for others there may exist considerable uncertainty when making the referral decision (e.g., undiagnosed chest pain). In Table 6 we report for each condition group the total error rate along with the percentage of patients admitted to an acute inpatient bed. This shows there to exist significant variability in error propensity across conditions that could be exploited to segment patients.

While separating patients based on their diagnosis is possible, within each of these diagnosis categories there can still exist considerable heterogeneity in the level of diagnostic uncertainty associated with individual patients. For example, a patient who arrives at the ED experiencing a heart attack will almost certainly be admitted (appropriately). On the other hand, if a patient arrives with unexplainable chest pain then whether or not a referral to the hospital is necessary is less clear: this could be something relatively routine, such as acid reflux, or a sign of something more serious. We thus use an approach that also accounts for other factors that can explain differences

Figure 4 Scatterplot of the difference between actual and expected admission against discharge error rates, per physician per month (restricted to physician-months with more than 50 obs.).



in levels of diagnostic uncertainty. Examples of such factors include the actual diagnosis (e.g., 'myocardial infarction' or 'chest pain with unknown cause'), arrival method (e.g., by helicopter versus walk-in), and history of ED attendances (e.g., zero visits in the last year versus multiple visits). This is achieved by estimation of a latent variable (probit) model of the form

$$TotErr_i^* = \tau_0 + \mathbf{T}_i\boldsymbol{\tau}_1 + \mathbf{D}_i\boldsymbol{\tau}_2 + \mathbf{C}_i\boldsymbol{\tau}_3 + \epsilon_i^\tau\,, \tag{5}$$

$$TotErr_i = \mathbb{1}[TotErr_i^* > 0]\,, \tag{6}$$

where $\epsilon_i^\tau \sim \mathcal{N}(0,1)$, $TotErr_i^*$ is a latent variable, $TotErr_i$ is the observed dichotomous variable indicating false admission or discharge, and the vectors $\mathbf{T}_i$, $\mathbf{D}_i$ and $\mathbf{C}_i$ contain the set of all temporal, diagnosis related and contextual factors described in Table 10. We then use this estimated model to calculated fitted values, $\widehat{TotErr}_i$, which specifies the degree of diagnostic uncertainty associated with a particular patient i. Note that $\widehat{TotErr}_i$ thus adjusts not only for the assigned diagnosis, but also all of those factors in vectors $\mathbf{T}_i$, $\mathbf{D}_i$ and $\mathbf{C}_i$.

Since the relationship between the level of diagnostic uncertainty and the dependent variables may not necessarily be linear, we use the quartiles of $\widehat{TotErr}_i$ to separate our observations into four categories which form the variable $DiagType_i$. This variable takes values Low, LowMed, MedHigh and High, which capture increasing levels of diagnostic uncertainty, respectively.

6.3. Physician Type

We will classify physicians based on whether they make fewer or more admission and discharge errors than the average physician after controlling for observable differences in the patient assignment. The idea is to perform a classification similar to that demonstrated in Figure 4, in which

we plot the difference between observed and expected error rates for each physician in each month and could then e.g. class physicians based on the quadrant in which they lie. To do this, we first estimate a probit model that takes the same form as that specified in Equations (5) and (6), but with latent and observed dichotomous variables instead to indicate admission error, i.e. $AdmErr_i^*$ and $AdmErr_i$. This model gives the baseline risk of a patient being admitted in error if treated by an 'average' physician. We then take the fitted values from the auxiliary equation, $\widehat{AdmErr}_i^*$, and estimate a fixed effects probit model, unique for each physician, of the form

$$AdmErr_{i_p}^* = \mathbf{M}_{i_p}\boldsymbol{\beta}_p + \widehat{AdmErr}_{i_p}^* + \epsilon_{i_p}^{\beta}\,, \tag{7}$$

$$AdmErr_{i_p} = \mathbb{1}[AdmErr_{i_p}^* > 0]\,, \tag{8}$$

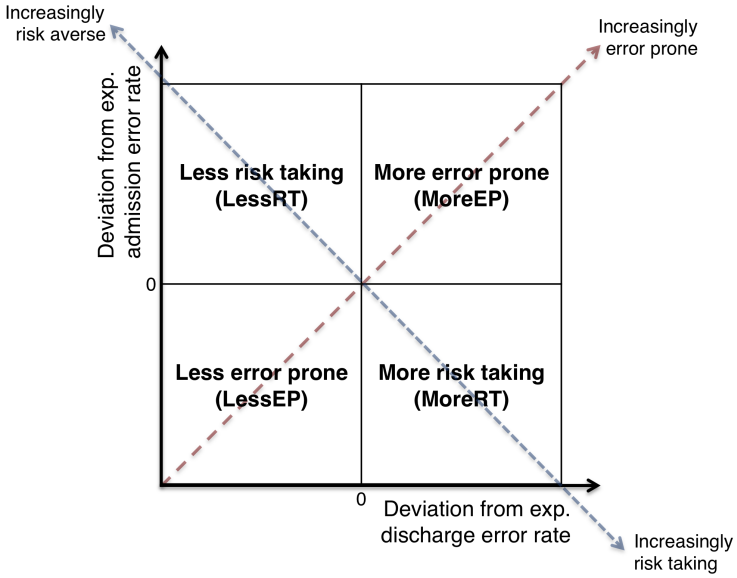
where $\epsilon_{i_p}^{\beta}$, $AdmErr_{i_p}^*$ and $AdmErr_{i_p}$ are as defined before but for the subset of observations i assigned to physician p, indexed i_p, and with $\mathbf{M}_{i_p}$ a vector indicating the month in which patient i_p visited the ED. Coefficient vector $\boldsymbol{\beta}_p$ then allows the intercept (i.e., the admission error rate) to vary in each month m for a physician p, with $(\hat{\boldsymbol{\beta}}_p)_m$ the estimated intercept for physician p in month m. Note that in Equation (7) we constrain the slope of $\widehat{AdmErr}_{i_p}^*$ equal to one so that for each physician in each month we are estimating their deviation from the expected error rate.

We would now like to classify physicians based on their historic error rates, but we cannot simply take a weighted average of values of $(\hat{\boldsymbol{\beta}}_p)_m$ due to non-linearity of the probit link function Φ^{-1}, where Φ is the cumulative distribution function of the Gaussian distribution. Instead, to estimate the error rate for physician p in month m we start by calculating their expected error rate in that month, $ExpErrRate_{pm} = \mu\left(\Phi(\widehat{AdmErr}_{i_{pm}}^*)\right)$, where i_{pm} is the subset of patients seen by physician i in month m and $\mu(\cdot)$ returns the arithmetic mean of the vector within parentheses. We then subtract this from the error rate we estimate for physician p in month m, $EstErrRate_{pm} = \mu\left(\Phi(\widehat{AdmErr}_{i_{pm}}^* + (\hat{\boldsymbol{\beta}}_p)_m)\right)$, to get $\delta_{pm} = EstErrRate_{pm} - ExpErrRate_{pm}$, the difference between physician p's estimated and the expected admission error rates in month m. This δ_{pm} is approx. equivalent to the y-axis in Figure 4. For a patient i_{pm} who is treated by physician p in month m the historic admission error propensity of the assigned physician is then calculated as a weighted average of δ_{pm} over the previous 12 months, i.e.

$$PhysAdmDev_{i_{pm}} = \frac{\sum_{t=m-12}^{m-1} n_{pt}\cdot\delta_{pt}}{\sum_{t=m-12}^{m-1} n_{pt}}\,, \tag{9}$$

where n_{pm} is the number of patients seen by physician p in month m. A higher value of $PhysAdmDev_{i_{pm}}$ indicates that the physician makes more admission errors that we would expect based on the risk profile of patients who they treat. We also introduce a binary value,

Figure 5 Classification of physicians based on error propensity and risk attitude.



$LowVolPhys_{i_{pm}}$, that is equal to one when $\sum_{t=m-12}^{m-1} n_{pt} < 118$ (the tenth percentile) and zero otherwise. This variable accounts for the fact that when the number of observations is low the value of $PhysAdmDev_{i_{pm}}$ is likely to be estimated with significant error, which makes any classification of physicians based off of this quantity unreliable. Finally, we map each of the i_{pm} indices back to the original i indices to form variable $PhysAdmDev_i$. Using the same method we create the variable $PhysDischDev_i$, equal to the weighted average of the assigned physician's deviation from the expected rate of discharge errors over the past 12 months.

If a physician has a higher than expected rate of admission errors, but a lower than expected rate of discharge errors, then this suggests that they may be less risk taking than the average physician, while a physician with a lower than expected rate of admission errors and higher than expected rate of discharge errors may be more risk taking. High rates of both false admissions and discharges suggests that the physician is more error prone, while lower rates of both suggest they are more experienced and less error prone. Thus we define the variable

$$PhysType_i = \begin{cases} \text{LessEP} & \text{if } PhysAdmDev_i < 0 \text{ and } PhysDischDev_i < 0 \\ \text{MoreRT} & \text{if } PhysAdmDev_i < 0 \text{ and } PhysDischDev_i \geq 0 \\ \text{LessRT} & \text{if } PhysAdmDev_i \geq 0 \text{ and } PhysDischDev_i < 0 \\ \text{MoreEP} & \text{otherwise} \end{cases} \tag{10}$$

where LessEP, MoreRT, LessRT, and MoreEP capture whether the physician is less error prone, less risk taking, more risk taking or more error prone, respectively. When $LowVolPhys_i = 1$ we set $PhysType_i = \text{LowVol}$. This variable is summarized visually in Figure 5.

6.4. Physician-level Controls

Given our interest in the influence of physician type, we elaborate on the physician related controls listed in Table 10. In particular, we would like to account for non-random assignment of patients to physicians. One way to achieve this would be through the inclusion of physician-level fixed effects. This is problematic, however, since we observe over 500 physicians in our data set, with 258, 229, 173 and 101 physicians seeing more than 100, 250, 500 and 1000 patients, respectively. Controlling with physician-specific fixed effects would thus introduce a large number of parameters, leading to problems in estimation. Specifically, ill-conditioned model data caused by near-linear dependence amongst the columns of the model matrix can result in inflated standard errors or even rank-deficiency (Belsley et al. 2005). Here we describe an alternative control method which follows intuitively from the formulation of the historic physician-level error rates in Section 6.3. In particular, similar to $PhysAdmDev_{i_{pm}}$ we can formulate the variable $ExpAdmErr_{i_{pm}}$, equal to physician p's expected admission error rate over the 12 months leading up to month m. This is calculated as in Equation (9), except with δ_{pm} replaced with $ExpErrRate_{pm}$, the expected error rate in month m. We can then convert back with the probit link function to the original scale by taking the difference between the observed and expected error rates to arrive at control variable $PhysAdm_{i_{pm}} = \Phi^{-1}\left(ExpAdmErr_{i_{pm}} + PhysAdmDev_{i_{pm}}\right) - \Phi^{-1}\left(ExpAdmErr_{i_{pm}}\right)$. This can then be mapped back to the original i indices. If deviation from expectation in the past is an indicator of future deviation then when we should expect $AdmErr_i$ and $PhysAdm_i$ to be correlated. To check this, we estimate a probit model with dependent variable $AdmErr_i$ and independent variable $PhysAdm_i$. The coefficient of $PhysAdm_i$ is estimated to be 1.039, very close to one with p-value $<$ 0.001, which suggests that a physician's past deviation from expectation is a very good indicator of their future deviation, with the past and future performance almost identical.

7. Models and Results II: Evaluating the Two-Stage Gatekeeping System

We would like to know whether the two-stage gatekeeping process, which decouples the gatekeeping decision by introducing a refer-out option, reduces the high rate of errors in referrals of patients from the ED into acute inpatient beds, as well the conditions under which the two-stage system is especially preferable over the single-stage alternative. In this section, we describe the method of estimation and present results.

7.1. Empirical Specification

The empirical approach that we adopt is similar to that described in Section 5.1, except that rather than use a heckprob model we estimate the models instead with a recursive bivariate probit

(biprobit) model, again with full information maximum likelihood (Maddala 1983). These models have the same error structure as the heckprob model but differ in that censoring is not performed and $\alpha_2, \beta_2, \gamma_2$ are left as free parameters to be estimated in the models. We first ask whether there is evidence that (and, if so, the extent to which) decoupling the gatekeeping decision and allowing ED physicians to, when they are uncertain, pass on the referral decision to a second gatekeeping stage can help to reduce the overuse of specialists and referral of patients to the wrong specialists. This would be confirmed by coefficients $\beta_2 < 0$ and $\gamma_2 < 0$ in the respective outcome equations. We are also interested in if there is any evidence of a change in discharge errors, estimated by α_2, when patients are routed through the CDU.

After establishing the benefits of the two-stage gatekeeping process, we then look to determine the characteristics that define those multi-tiered service contexts which stand to benefit from it most. We hypothesized the largest reduction in unnecessary referrals when (a) there are high levels of diagnostic uncertainty, and (b) in the presence of less risk taking and more error prone gatekeepers. To identify if these hypotheses are supported, we re-run the biprobit models described above but re-specify the selection and outcome equations to include as additional variables $DiagType_i$ (or $PhysType_i$, depending on the hypothesis under investigation). We also add to the outcome equation the interaction between these variables and the CDU admission decision, i.e. $(CDU \times DiagType)_i$ or $(CDU \times PhysType)_i$, so that the updated outcomes equation takes (e.g. for admission errors) one of the following two forms

$$AdmErr_i^* = \phi_0 + \mathbf{X}_i\boldsymbol{\phi}_1 + CDU_i\phi_2 + zOccED_i\phi_3 + DiagType_i\boldsymbol{\phi}_4 + (CDU \times DiagType)_i\,\boldsymbol{\phi}_5 + \epsilon_i\,,$$
$$AdmErr_i^* = \varphi_0 + \mathbf{X}_i\boldsymbol{\varphi}_1 + CDU_i\varphi_2 + zOccED_i\varphi_3 + PhysType_i\boldsymbol{\varphi}_4 + (CDU \times PhysType)_i\,\boldsymbol{\varphi}_5 + \epsilon_i\,.$$

Note that the base category in the first (second) equation is a patient with a low level of diagnostic uncertainty (who is treated by a less error prone physician) who does not go to the CDU, while ϕ_2 (φ_2) captures the impact on admission errors if that patient had instead been referred to the CDU. The vector of coefficients $\boldsymbol{\phi}_4$ ($\boldsymbol{\varphi}_4$) estimates the change in error propensity if the base patient had instead had a low-medium, medium-high or high level of diagnostic uncertainty (treated by a more risk taking, less risk taking, or more error prone physician), and $\boldsymbol{\phi}_5$ ($\boldsymbol{\varphi}_5$) captures the additional impact (on top of that contributed by ϕ_2 (φ_2) and $\boldsymbol{\phi}_4$ ($\boldsymbol{\varphi}_4$)) if that patient is also treated in the CDU. If as hypothesized the CDU referral option is more beneficial the higher the level of diagnostic uncertainty then we should expect that $(\boldsymbol{\phi}_5)_d$ is negative and significant for all d, and becomes increasingly negative as the level of diagnostic uncertainty increases from the first to the fourth quartile. If also the CDU referral option reduces variation in error propensity across different physician types then we should expect the coefficients of $(\boldsymbol{\phi}_5)_p$ to reflect this.

Table 7 Coefficient estimates for CDU impact.

	(1) AdmErr		(2) DischErr		(3) SpecChg	
	(1s) CDU	(1o) AdmErr	(2s) CDU	(2o) DischErr	(3s) CDU	(3o) SpecChg
CDU referral	–	-0.740^{***}	–	0.075	–	-0.351^{***}
		(0.032)		(0.066)		(0.083)
CDU busyness	-0.046^{***}	–	-0.047^{***}	–	-0.052^{***}	–
	(0.002)		(0.002)		(0.003)	
Phys. CDU rate	0.536^{***}	–	0.362^{***}	–	0.212^{***}	–
	(0.024)		(0.028)		(0.037)	
ρ	0.292^{***}		0.074^{*}		0.095^{*}	
	(0.018)		(0.036)		(0.044)	
N	429,313		429,313		125,228	
Log-lik	−151,007		−112,884		−87,296	

Notes: All estimations made using a biprobit model specification; *Robust standard error* in parentheses; Columns (1s), (2s) and (3s) report coefficient estimates for the first-stage (selection) equation, while columns (1o), (2o) and (3o) report coefficients for the second-stage (outcome) equation; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

7.2. Results: The CDU Effect

Looking first at the question of whether patients admitted to the CDU have lower admission and discharge error rates and less chance of being referred to the wrong specialist, we find in Table 7 evidence of positive correlation in each of the three respective biprobit models, with estimated correlation coefficients $\rho = 0.292$ (p-value < 0.001), $\rho = 0.074$ (p-value $= 0.038$), and $\rho = 0.095$ (p-value $= 0.033$), respectively. This suggests that patients selected for admission to the CDU are more likely to be a false admission, false discharge, and to require specialty transfer than an 'average' patient who visits the ED. This is consistent with expectation: patients admitted to the CDU should be more complicated than the average ED arrival, else this more expensive service would be being used inappropriately. These biprobit model estimates provide strong evidence that patients admitted to the CDU are significantly less likely to (i) result in false admission (coef. $= -0.740$, p-value < 0.001 in column (2o)) and to (ii) require a transfer of specialty after admission (coef. $= -0.351$, p-value < 0.001 in column (4o)), with (iii) no corresponding increase in discharge errors reported (coef. $= 0.075$, p-value > 0.10 in column (3o)). This confirms our hypothesis that routing customers with unresolved diagnostic uncertainty through a two-stage gatekeeping system can help to significantly reduce the number of referral errors made in systems staffed by gatekeepers with a low tolerance for risk of non-referral errors, as well as helping as a secondary benefit to ensure that customers are referred to the correct specialist.

To see how much better admission decisions are when made in the CDU rather than by an ED physician, in Table 9 we convert coefficient estimates to average treatment effects (ATEs) and average treatment effects on the treated (ATTs). The ATEs and ATTs for biprobits (1o), (2o) and (3o) are given in columns (3) and (4) of Table 9. These results show that if no patients were referred through the CDU the rate of admission, discharge and specialty routing errors would have been

5.72%, 0.92%, and 22.8%, respectively. These change to 1.45%, 1.12% and 15.2%, respectively, if all patients are instead routed through the CDU. Thus the CDU acts to significantly reduce admission and specialty transfer errors with little change in discharge errors. Moreover, the especially large and negative ATT for admission errors, −12.2%, suggests that ED physicians are especially good at routing into the CDU patients who they would have otherwise been admitted in error, and also that the CDU significantly reduces the rate of false admissions for these patients.

7.3. Results: The Role of Diagnostic Uncertainty and Physician Type

Given the significant reduction in hospital admission errors and specialty transfers when patients are referred through the CDU, we next attempt to determine the characteristics of those service systems that are most likely to benefit from allowing the gatekeepers the option of deferring their referral decision to another more specialized decision maker.

First looking at the role of diagnostic uncertainty, the relevant coefficients are reported in biprobits (1d), (2d) and (3d) of Table 8. When the level of diagnostic uncertainty is low (i.e. in the base case) there is no change in false admission rates (coef. $= -0.100$, p-value > 0.10) while there are more false discharges (coef. $= 0.590$, p-value < 0.001) and fewer specialty routing errors (coef. $= -0.238$, p-value $= 0.067$). As the level of diagnostic uncertainty increases, routing patients through the CDU becomes increasingly valuable. First, the CDU acts to significantly reduce the rate of false admissions, with coefficients −0.337, −0.669 and −0.719 (p-values < 0.001 in biprobit (1d)) when diagnostic uncertainty increases to low-medium, medium-high and high levels. At the same time, although discharge errors become more likely as diagnostic uncertainty increases, routing through the CDU acts to counteract this effect, with coefficients taking values −0.429, −0.555 and −0.440 (p-values < 0.001 in biprobit (2d)) for low-medium, medium-high and high levels of diagnostic uncertainty. Coefficients for specialty transfer errors, while negative, are not individually significant at the 10% level in biprobit (3d). These results confirm our hypothesis that the two-stage system offers greater benefits in the presence of higher levels of diagnostic uncertainty.

Finally, we look in biprobits (1p) and (2p) to identify whether there is evidence that characteristics of the gatekeeper impact on the benefits achievable through the two-stage system. We see in biprobit (1p) that even when ED physicians are less error prone, fewer admissions errors are made (coef. $= -0.685$, p-value < 0.001) when referral decisions are made in the CDU rather than by an ED physician, with no difference in benefit for less error prone and less risk averse physicians (coefs. $= -0.027$, p-value > 0.10). As ED physicians become more risk averse (coef. $= -0.076$, p-value $= 0.037$) and more error prone (coef. $= -0.239$, p-value < 0.001) there are also further reductions in admission errors when patients are admitted through the CDU. For discharge errors in biprobit

Table 8 Coefficient estimates for CDU impact based on level of diagnostic uncertainty and physician type.

	Diag. type interactions			Phys. type interactions		
	(1d) AdmErr	(2d) DischErr	(3d) SpecChg	(1p) AdmErr	(2p) DischErr	(3p) SpecChg
CDU referral	-0.100	0.590^{***}	$-0.238^{\dagger}$	-0.685^{***}	0.166^{*}	-0.310^{***}
	(0.094)	(0.127)	(0.136)	(0.040)	(0.074)	(0.091)
Diag. type/Phys. type						
Low/LessEP	0	0	0	0	0	0
LowMed/MoreRT	$-0.072^{\dagger}$	0.206^{***}	-0.052	0.005	$0.045^{\dagger}$	0.021
	(0.039)	(0.027)	(0.058)	(0.017)	(0.025)	(0.021)
MedHigh/LessRT	-0.037	0.387^{***}	-0.016	0.025	-0.017	$-0.035^{\dagger}$
	(0.049)	(0.040)	(0.069)	(0.018)	(0.032)	(0.021)
High/MoreEP	$-0.092^{\dagger}$	0.296^{***}	-0.036	0.059^{**}	0.030	0.036
	(0.054)	(0.045)	(0.074)	(0.021)	(0.034)	(0.026)
CDU x Diag. type/Phys. type						
Low/LessEP	0	0	0	0	0	0
LowMed/MoreRT	-0.337^{***}	-0.429^{***}	-0.113	-0.027	-0.124^{**}	0.057
	(0.094)	(0.096)	(0.119)	(0.036)	(0.047)	(0.045)
MedHigh/LessRT	-0.669^{***}	-0.555^{***}	-0.141	-0.076^{*}	0.001	0.127^{**}
	(0.091)	(0.092)	(0.114)	(0.038)	(0.056)	(0.046)
High/MoreEP	-0.719^{***}	-0.440^{***}	-0.102	-0.239^{***}	$-0.107^{\dagger}$	0.069
	(0.089)	(0.092)	(0.114)	(0.046)	(0.063)	(0.056)
ρ	0.299^{***}	0.046	0.097^{*}	0.298^{***}	$0.061^{\dagger}$	0.044
	(0.019)	(0.041)	(0.045)	(0.018)	(0.037)	(0.047)
N	429,313	429,313	125,228	429,313	429,313	125,228
Log-lik	$-150{,}827$	$-112{,}014$	$-87{,}287$	$-150{,}982$	$-112{,}864$	$-87{,}279$

Notes: All estimations made using the biprobit model specification; Rows Low/LessEP corresponds to the base category, which is constrained to 0 by design; *Robust standard error* in parentheses; Likelihood ratio ($\Pr > \chi^2$) < 0.0001 in all models.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

(2p), we find that the patients treated by more risk taking (coef. $= -0.124$, p-value < 0.01) and more error prone (coef. $= -0.107$, p-value < 0.10) physicians experience the biggest reduction in false discharge error rates when referred through the two-stage gatekeeping system.

The ATEs and ATTs for the models in Table 8 are again given in Table 9. For the subset of patients who we classify as having low levels of diagnostic uncertainty, the ATE indicates that only a 0.04% reduction in admission errors is achieved by treating all patients in the CDU rather than in the ED, which increases to an impressive 11.3% reduction for patients classed as having high levels of diagnostic uncertainty. Comparing the $CDU(0)$ and $CDU(1)$ columns for the physician-type models, we can see clearly that, as hypothesized, there is a significant reduction in the variability of error rates when patients are admitted or discharged through the CDU rather than directly by the first-stage gatekeepers.

8. Managerial Implications and Conclusions

Our results suggest that, especially in systems with high levels of diagnostic uncertainty and less risk taking and more error prone decision makers, decoupling the gatekeeping decision and allowing gatekeepers the possibility of referring patients on to a third party who assumes responsibility for

Table 9 Table of average treatment effects (ATEs) and average treatment effects on the treated (ATTs).

	Biprobits (1o), (2o) and (3o)				Biprobits (1d), (2d) and (3d)				Biprobits (1p), (2p) and (3p)			
	CDU(0) (1)	*CDU(1)* (2)	ATE (3)	ATT (4)	*CDU(0)* (5)	*CDU(1)* (6)	ATE (7)	ATT (8)	*CDU(0)* (9)	*CDU(1)* (10)	ATE (11)	ATT (12)
Admission errors												
Aggregate (%)	5.72	1.45	−4.27	−12.2	5.75	1.38	−4.37	−12.5	5.74	1.38	−4.36	−12.4
Low / LessEP (%)					0.17	0.12	−0.04	−0.57	2.84	0.76	−2.08	−10.0
LowMed / MoreRT (%)					1.31	0.45	−0.87	−4.01	4.82	1.23	−3.59	−10.7
MedHigh / LessRT (%)					6.50	1.28	−5.23	−11.2	9.97	2.55	−7.43	−16.3
High / MoreEP (%)					15.0	3.66	−11.3	−19.7	9.90	1.84	−8.07	−17.4
Discharge errors												
Aggregate (%)	0.92	1.12	0.20	0.31	0.91	1.45	0.54	0.44	0.92	1.18	0.27	0.38
Low / LessEP (%)					0.27	1.40	1.13	1.64	0.63	0.98	0.35	0.64
LowMed / MoreRT (%)					0.68	1.04	0.36	0.47	1.08	1.20	0.12	0.18
MedHigh / LessRT (%)					1.49	1.62	0.13	0.15	0.78	1.20	0.42	0.56
High / MoreEP (%)					1.21	1.75	0.54	0.62	1.32	1.53	0.21	0.24
Specialty routing errors												
Aggregate (%)	22.8	15.2	−7.65	−8.62	22.8	15.1	−7.70	−8.70	22.6	17.2	−5.40	−6.16
Low / LessEP (%)					7.60	5.33	−2.27	−3.56	21.9	15.1	−6.86	−7.30
LowMed / MoreRT (%)					16.3	10.6	−5.75	−7.03	20.5	15.1	−5.46	−5.80
MedHigh / LessRT (%)					24.1	15.5	−8.56	−9.66	26.5	21.9	−4.62	−4.55
High / MoreEP (%)					24.3	16.5	−7.87	−8.89	18.6	14.1	−4.46	−5.85

Notes: Column $CDU(0)$ and $CDU(1)$ report expected rates of errors and specialty transfers in the cases where no patients receive treatment (i.e., if no patients were referred to the CDU) and where all patients receive treatment (i.e., if all patients are referred to the CDU), respectively; Columns ATE and ATT report average treatment effects and average treatment effects on the treated, respectively.

the referral decision may help to improve system performance by reducing the rate of over-referrals. One possibility that we should consider, however, is that if the CDU were not present then resources used to operate the CDU could instead be redeployed to the ED. Since we have found that fewer false admissions occur when the ED is less busy, this increase in ED resourcing might perhaps be more beneficial than the presence of the CDU. To assess whether this is the case, we now perform a counterfactual analysis based around merging the CDU with the ED and pooling capacity.

Over our six year sample period the total number of hours spent by patients in the ED was 1.5m, with 326k hours spent by patients in the CDU. If the ED and CDU were merged, therefore, capacity in the ED would increase by approx. 21.7%.[4] To estimate the effect that the increase in capacity would have on the rate or admission errors, we must adjust observed ED busyness downwards to account for the fact that more resources (e.g., physicians, nurses, treatment rooms) would have been available. To do this we multiply $QueueED_h^{95th}$, our measure of capacity based off of the 95th percentile of ED busyness in Section 4.4, by 1.217 and re-estimate $OccED_i$ for all i. To ensure that the original and updated measures of ED busyness are on the same scale, we standardize using the original mean, $\mu(OccED_i)$, and standard deviation, $\sigma(OccED_i)$. On average this has the effect of reducing workload by $\sim 0.64\sigma$. Substituting the original values of $zOccED_i$

[4] Note that we take a very conservative view and assume that all of those patients who were treated in the CDU could have instead been relocated elsewhere in the hospital without any additional capacity needing to be installed, meaning that all resources from the CDU can be redeployed to the ED. We thus estimate an upper bound on the gains that could be achieved from pooling ED and CDU capacity.

for the updated values achieved through pooling ED and CDU capacity suggests there would be an approx. 0.23% reduction in false admissions. On the other hand, using the model estimated by biprobit (1o) in Table 7, had the CDU not been available then the average rate of admission errors would have been 5.72% (see Table 9), as compared to an average rate in the sample of 4.71%, a reduction of approx. 1.01%. This suggests that the pre-screening of patients that takes place in the ED prior to offload to the CDU allows resources in the CDU to be targeted at those patients who benefit the most from more specialized and higher intensity service, which confers advantages above and beyond those that would occur is those resources were allocated at random in the ED.

Gatekeepers play an important role in ensuring that patients receive care of the appropriate intensity and that they are seen by the right specialist for their specific needs. We find that their behavioral response to higher levels of diagnostic uncertainty as busyness increase and service times are compressed may increase the rate of overuse of expensive specialist services. This observation provides further insight into the trade-off between speed and quality: while it might be desirable to encourage workers to act faster and make quicker decisions – reducing waiting times for other customers and increasing throughput – this might not only reduce service quality but may also come at the expense of additional errors in routing. In particular, in service contexts in which the server not only provides the service but also must diagnose customer's needs, shortening service times can reduce diagnostic accuracy resulting in both unnecessary and/or inaccurate referrals. When the gatekeeper associates a significantly greater cost to under-referral than to over-referral this problem is amplified. From a theoretical perspective, these empirical observations suggest that the modeling literature on gatekeeping systems, which ignores the endogenous response to congestion on service times and routing accuracy and does not account for the cost of error incurred by the gatekeeper, may need updating. From a practical perspective, these results suggest that caution should be taken if pursuing policies to reduce waiting time at the potential expense of shorter service times. Waiting time targets – which are pervasive in health care (Viberg et al. 2013) as well as in other service settings e.g. call centers (Gans et al. 2003) – are an example of this.

Our findings also provide an alternative strategy to adding capacity that can be implemented in multi-tiered service contexts to improve the accuracy at which customers are routed to the appropriate service provider. By allowing gatekeepers the option of passing the referral decision to a higher intensity gatekeeping tier (rather than directly to the even more costly specialist) when diagnostic uncertainty is high, the first-stage gatekeeper can focus on processing those more unambiguous cases while the second-stage gatekeeper searches for the appropriate treatment option for the more complex cases. In medical contexts and others in which diagnostic uncertainty is

high and where referral accuracy is highly dependent on the skill and behavior of the first-stage gatekeeper, we demonstrate that such a system is especially valuable. An interesting possible direction for future research might be to translate our empirical findings to a modeling approach and to investigate the two-stage gatekeeping system analytically.

While we focus in this paper on emergency care, such benefits are likely to extend beyond this to other industries and health contexts. For example, accurate detection and diagnosis of rare diseases in primary care takes on average over seven years in the US and five years in the UK. These patients are costly, visiting their primary care physician (PCP) multiple times, being subject to multiple tests, and seeing multiple specialists. Our results suggest that one potential solution to this would be to designate a subset of more experienced PCPs, e.g. who have a track-record of identifying more complex diseases, as second-stage gatekeepers and allowing PCPs to refer to them their patients (Shire 2013). Our findings suggest that such a two-stage gatekeeping system can help to reduce overuse of inappropriate specialist services while also improving the accuracy of referral, a win-win for both the payer and the patient.

Appendix A: Data Preparation

We now describe the method that we use to prepare the data for analysis. First, for the subset of admitted patients we merge their ED records with their inpatient records using a unique time-invariant patient identifier. To ensure an accurate matching is made we require that the ED departure and inpatient admission timestamps are within two hours of each other. After doing this we are left with 7,771 unmatched records corresponding to 7,496 patients. For the 7,249 patients for who we are unable to match their records only once, we drop only those obs. corresponding to the unmatched ED attendances. For the 247 patients with multiple missing records we drop all 3,208 visits that they make to the ED. We also drop 299 obs. corresponding to nine patients for whom our matching algorithm assigns two or more ED records to the same inpatient record. Finally, we drop 311 ED attendances where the patient was supposed to have been admitted to the CDU but where no corresponding record exists, 351 attendances where the patient was not meant to be admitted to the CDU but where we find a record of them being in the CDU, and 17 obs. with timestamps indicating ED discharge prior to arrival. This leaves an initial sample of 631,082 obs. (98.2% of the data).

Next, we use the first year (December 1st 2006 through November 31st 2007) as a warm-up period to generate various measures of patient risk (e.g., ED visits in the last year) and physician behavior (e.g., propensity to admit – see Section 6.3). This reduces the sample to 552,512 obs. over six full years. As staffing levels and patient behavior may differ greatly over the Christmas period and around public holidays, we also drop in each year all obs. corresponding to the dates December 20th through January 10th, as well as the three day period from one day before until one day after each public holiday. Finally, due to a temporary change in coding convention in December 2009 and January 2010 that makes identification of patient admissions to the CDU challenging, we drop all observations from November 15th 2009 to February 15th 2010. After applying all such temporal restrictions we are left with 479,678 obs.

Table 10 Table of controls.

	Type	Description
Temporal ($\mathbf{T}_i$)		
Year	Categorical (6)	Observation year (offset by one month so e.g. December '07 falls in '08), 2008 through 2013
Daily time trend	Continuous	A variable that takes value one on the first observation date and increases in value by one per day
Month	Categorical (12)	Month of the year in which the visit falls, January through December
School break	Categorical (7)	If visit occurs during a school break, equals the break type (e.g., Easter, Fall), else set to None
Day of week	Categorical (7)	Specifies the day of the week on which the visit occurred, Monday through Sunday
Window of arrival x weekend	Categorical (24)	A two-hourly arrival window (e.g., 2am to 4am) for weekdays, and a separate one for weekends
Diagnosis related factors ($\mathbf{D}_i$)		
Diagnostic category	Categorical (17)	The general medical specialty assigned to the patient's condition (e.g., gastrointenstinal, cardiac)
Diagnosis	Categorical (46)	The specific diagnosis assigned to the patient (e.g., cardiac arrhythmia, urinary tract infection)
Affected region of body	Categorical (7)	The area of the body affected (e.g., head, trunk, multiple)
Procedure performed	Categorical (22)	The procedure performed, if any (e.g., splint, chest drain, provision of oral medicine)
Contexual factors ($\mathbf{C}_i$)		
Mode of arrival	Categorical (8)	The mode of transport used to get to the hospital (e.g., helicopter, private, ambulance)
ED visits, last year	Continuous	The number of times the patient visited the ED in the previous 12 months
ED visits, last month	Continuous	The number of times the patient visited the ED in the previous one month
Admissions per ED visit, last year	Continuous	The rate of hospital admissions to ED visits in the previous 12 months
Admissions per ED visit, last month	Continuous	The rate of hospital admissions to ED visits in the previous month
Zero ED visits, last year	Binary	A variable equal to one if the patient did not attend the ED in the previous 12 months, else zero
Zero ED visits, last month	Binary	A variable equal to one if the patient did not attend the ED in the previous month, else zero
Physician related factors ($\mathbf{P}_i$)		
Admission errors	Continuous	The admission error propensity of the assigned physician, calculated as in Section 4.5
Discharge errors	Continuous	The discharge error propensity of the assigned physician, calculated as in Section 4.5
Admission errors x Discharge errors	Continuous	The interactions between the two variables defined above
Physician category	Categorical (14)	Specifies the type of physician (e.g., orthopedic, plastics) for 33% of the visits where the physician name is not specified due to treatment being provided by a junior (non-consultant grade) physician
Operational/other factors ($\mathbf{O}_i$)		
Length of stay in ED	Categorical(13)	30 minute time windows capturing how long the patient spent in the ED (e.g., 60-90 mins). Stays beyond 6 hours are merged into an 'over 360 minutes' category.
Hospital congestion	Continuous	The overall busyness of the main hospital inpatient units in to which ED patients are admitted, calculated using the same method as for ED congestion in Section 4.4

Notes: All diagnostic related factors for which in the raw data there were fewer than 3,000 obs. (approx. 0.5% of the data) are combined in to an "Other" category, prior to reporting and analysis; If a patient did not visit the ED in the previous 12 months (or month) then the "Admission per ED visit, last year" ("last month") variable is set equal to zero.

Lastly, since we are interested in referral errors made by ED physicians we perform a final round of cleaning by removing all visits by patients who died in or before arrival to the ED (478 obs.), who left without being seeing, against medical advice, or who refused treatment (11,535 obs.), or who were transferred to another hospital (724 obs.). We also exclude 37,628 ED visits during which the patient was seen by an ED nurse rather than a physician. This leaves 429,313 ED attendances that we take forward for analysis.

Appendix B: Control variables

In Table 10 we describe the variables used as controls in the models.

References

Alizamir S, de Véricourt F, Sun P (2013) Diagnostic accuracy under congestion. *Management Science* 59(1):157–171.

Allison JJ, Kiefe CI, Cook EF, Gerrity MS, Orav EJ, Centor R (1998) The association of physician attitudes about uncertainty and risk taking with resource use in a Medicare HMO. *Medical Decision Making* 18(3):320–329.

Anand K, Paç M, Veeraraghavan S (2011) Quality-speed conundrum: Trade-offs in customer-intensive services. *Management Science* 57(1):40–56.

Argon N, Ziya S (2009) Priority assignment under imperfect information on customer type identities. *Manufacturing & Service Operations Management* 11(4):674–693.

Baugh CW, Venkatesh AK, Hilton JA, Samuel PA, Schuur JD, Bohan JS (2012) Making greater use of dedicated hospital observation units for many short-stay patients could save $3.1 billion a year. *Health Affairs* 31(10):2314–2323.

Beckett DJ, Inglis M, Oswald S, Thomson E, Harley W, Wilson J, Lloyd RC, Rooney KD (2013) Reducing cardiac arrests in the acute admissions unit: a quality improvement journey. *BMJ Quality & Safety* bmjqs–2012.

Belsley DA, Kuh E, Welsch RE (2005) Detecting and assessing collinearity. *Regression Diagnostics*, 85–191 (John Wiley & Sons, Inc.).

Bernstein S, Aronsky D, Duseja R, Epstein S, Handel D, Hwang U, McCarthy M, McConnell K, Pines J, Rathlev N, Schafermeyer R, Zwemer F, Schull M, Asplin B (2009) The effect of emergency department crowding on clinically oriented outcomes. *Academic Emergency Medicine* 16(1):1–10.

Berry Jaeker J, Tucker A (2016) Past the point of speeding up: The negative effects of workload saturation on efficiency and quality. *Management Science (forthcoming)* .

Blatchford O, Capewell S (1997) Emergency medical admissions: taking stock and planning for winter. *British Medical Journal* 315(7119):1322–1323.

Brekke K, Nuscheler R, Straume O (2007) Gatekeeping in health care. *Journal of Health Economics* 26(1):149–170.

Bunik M, Glazner JE, Chandramouli V, Emsermann CB, Hegarty T, Kempe A (2007) Pediatric telephone call centers: how do they affect health care use and costs? *Pediatrics* 119(2):e305–e313.

Burgess C (1998) Are short-stay admissions to an acute general medical unit appropriate? Wellington Hospital experience. *The New Zealand Medical Journal* 111(1072):314–315.

Chan C, Farias V, Escobar G (2016) The impact of delays on service times in the intensive care unit. *Management Science (Forthcoming)* .

Chan CW, Green LV, Lu Y, Leahy N, Yurt R (2013) Prioritizing burn-injured patients during a disaster. *Manufacturing & Service Operations Management* 15(2):170–190.

Christensen J, Levinson W, Dunn P (1992) The heart of darkness: The impact of perceived mistakes on physicians. *Journal of General Internal Medicine* 7(4):424–431.

Cooke M, Higgins J, Kidd P (2003) Use of emergency observation and assessment wards: A systematic literature review. *Emergency Medicine Journal* 20(2):138–142.

Cosby KS, Roberts R, Palivos L, Ross C, Schaider J, Sherman S, Nasr I, Couture E, Lee M, Schabowski S, Ahmad I, Scott RD (2008) Characteristics of patient care management problems identified in emergency department morbidity and mortality investigations during 15 years. *Annals of Emergency Medicine* 51(3):251–261.

Croskerry P (2002) Achieving quality in clinical decision making: Cognitive strategies and detection of bias. *Academic Emergency Medicine* 9(11):1184–1204.

Debo L, Toktay L, Van Wassenhove L (2008) Queuing for expert services. *Management Science* 54(8):1497–1512.

Denman-Johnson M, Bingham P, George S (1997) A confidential enquiry into emergency hospital admissions on the Isle of Wight, UK. *Journal of Epidemiology and Community Health* 51(4):386–390.

Dijk NM, Sluis E (2008) To pool or not to pool in call centers. *Production and Operations Management* 17(3):296–305.

Fiscella K, Franks P, Zwanziger J, Mooney C, Sorbero M, Williams GC (2000) Risk aversion and costs: a comparison of family physicians and general internists. *Journal of Family Practice* 49(1):12–12.

FitzGerald G, Jelinek GA, Scott D, Gerdtz MF (2010) Emergency department triage revisited. *Emergency Medicine Journal* 27(2):86–92, URL `http://dx.doi.org/10.1136/emj.2009.077081`.

Franks P, Williams GC, Zwanziger J, Mooney C, Sorbero M (2000) Why do physicians vary so widely in their referral rates? *Journal of General Internal Medicine* 15(3):163–168.

Freeman M, Savva N, Scholtes S (2016) Gatekeepers at work: An empirical analysis of a maternity unit. *Management Science (Forthcoming)* .

Gans N, Koole G, Mandelbaum A (2003) Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management* 5(2):79–141.

Gawande A (2015) Overkill. *New Yorker* URL `http://www.newyorker.com/magazine/2015/05/11/overkill-atul-gawande`.

González P (2010) Gatekeeping versus direct-access when patient information matters. *Health economics* 19(6):730–754.

Graber ML (2013) The incidence of diagnostic error in medicine. *BMJ Quality & Safety* 22(Suppl 2):ii21–ii27.

Graber ML, Franklin N, Gordon R (2005) Diagnostic error in internal medicine. *Archives of Internal Medicine* 165(13):1493–1499.

Green SM, Martinez-Rumayor A, Gregory SA, Baggish AL, O'Donoghue ML, Green JA, Lewandrowski KB, Januzzi JL (2008) Clinical uncertainty, diagnostic accuracy, and outcomes in emergency department patients presenting with dyspnea. *Archives of Internal Medicine* 168(7):741–748.

Greenwald PW, Estevez RM, Clark S, Stern ME, Rosen T, Flomenbaum N (2016) The ED as the primary source of hospital admission for older (but not younger) adults. *The American Journal of Emergency Medicine* 34(6):943–947.

Hasija S, Pinker E, Shumsky R (2005) Staffing and routing in a two-tier call centre. *International Journal of Operational Research* 1(1/2):8–29.

Hassan T (2003) Clinical decision units in the emergency department: Old concepts, new paradigms, and refined gate keeping. *Emergency Medicine Journal* 20(2):123–125.

Hopp W, Iravani S, Yuen G (2007) Operations systems with discretionary task completion. *Management Science* 53(1):61–77.

HSCIC (2013) OPCS-4 classification. Technical report, Health & Social Care Information Centre, URL `http://systems.hscic.gov.uk/data/clinicalcoding/codingstandards/opcs4`.

Hu B, Benjaafar S (2009) Partitioning of servers in queueing systems during rush hour. *Manufacturing & Service Operations Management* 11(3):416–428.

Huang Q, Thind A, Dreyer J, Zaric G (2010) The impact of delays to admission from the emergency department on inpatient outcomes. *BMC Emergency Medicine* 10(1):1–6.

Kachalia A, Gandhi TK, Puopolo AL, Yoon C, Thomas EJ, Griffey R, Brennan TA, Studdert DM (2007) Missed and delayed diagnoses in the emergency department: A study of closed malpractice claims from 4 liability insurers. *Annals of Emergency Medicine* 49(2):196–205.

KC D, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* 55(9):1486–1498.

KC D, Terwiesch C (2012) An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* 14(1):50–65.

Kim S, Chan C, Olivares M, Escobar G (2014) ICU admission control: An empirical study of capacity allocation and its implication for patient outcomes. *Management Science* 61(1):19–38.

Kostami V, Rajagopalan S (2013) Speed-quality trade-offs in a dynamic model. *Manufacturing & Service Operations Management* 16(1):104–118.

Kuntz L, Mennicken R, Scholtes S (2014) Stress on the ward: Evidence of safety tipping points in hospitals. *Management Science* 61(4):754–771.

Leape LL (1994) Error in medicine. *JAMA* 272(23):1851–1857.

Lee H, Pinker E, Shumsky R (2012) Outsourcing a two-level service process. *Management Science* 58(8):1569–1584.

Lee LF (1978) Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International Economic Review* 19(2):415–433.

Maddala G (1983) *Limited Dependent and Qualitative Variables in Econometrics* (New York: Cambridge University Press).

Malcomson J (2004) Health service gatekeepers. *The RAND Journal of Economics* 35(2):401–421.

Mandelbaum A, Reiman MI (1998) On pooling in queueing networks. *Management Science* 44(7):971–981.

Mariñoso BG, Jelovac I (2003) GPs' payment contracts and their referral practice. *Journal of Health Economics* 22(4):617–635.

McKibbon KA, Fridsma DB, , Crowley RS (2007) How primary care physicians' attitudes toward risk and uncertainty affect their use of electronic information resources. *Journal of the Medical Library Association* 95(2):138–?146.

NAO (2013) Emergency admissions to hospital: managing the demand. Technical report, National Audit Office, URL `https://www.nao.org.uk/report/emergency-admissions-hospitals-managing-demand/`, Last accessed: 2016-09-21.

Needleman J, Buerhaus P, Pankratz V (2011) Nurse staffing and inpatient hospital mortality. *N. Engl. J. Med.* 364(11):1037–1045.

NHE (2016) Worst NHS figures ever triggered by 'unprecedented funding slowdown'. Technical report, National Health Executive, URL `http://www.nationalhealthexecutive.com/Health-Care-News/worst-nhs-performance-figures-ever-triggered-by-unprecedented-funding-slowdown`, Last updated: 2016-04-16, Last accessed: 2016-09-15.

NHS (2013) 2014/15 NHS standard contract. Technical report, NHS England, URL `https://www.england.nhs.uk/nhs-standard-contract/14-15/`, Last accessed: 2016-09-15.

Paç M, Veeraraghavan S (2015) False diagnosis and overtreatment in services, the Wharton School, Working paper.

Pearson SD, Goldman L, Orav EJ, Guadagnoli E, Garcia TB, Johnson PA, Lee TH (1995) Triage decisions for emergency department patients with chest pain. *Journal of General Internal Medicine* 10(10):557–564.

Pines JM, Hilton JA, Weber EJ, Alkemade AJ, Al Shabanah H, Anderson PD, Bernhard M, Bertini A, Gries A, Ferrandiz S, Kumar VA, Harjola VP, Hogan B, Madsen B, Mason S, Öhlén G, Rainer T, Rathlev N, Revue E, Richardson D, Sattarian M, Schull MJ (2011) International perspectives on emergency department crowding. *Academic Emergency Medicine* 18(12):1358–1370.

Pinkney J, Rance S, Benger J, Brant H, Joel-Edgar S, Swancutt D, Westlake D, Pearson M, Thomas D, Holme I, Endacott R, Anderson R, Allen M, Purdy S, Campbell J, Sheaff R, Byng R (2016) How can frontline expertise and new models of care best contribute to safely reducing avoidable acute admissions? a mixed-methods study of four acute hospitals. *Health Services and Delivery Research* 4(3):1–202.

Pope JH, Aufderheide TP, Ruthazer R, Woolard RH, Feldman JA, Beshansky JR, Griffith JL, Selker HP (2000) Missed diagnoses of acute cardiac ischemia in the emergency department. *New England Journal of Medicine* 342(16):1163–1170, URL `http://dx.doi.org/10.1056/NEJM200004203421603`.

Powell A, Savin S, Savva N (2012) Physician workload and hospital reimbursement: Overworked physicians generate less revenue per patient. *Manufacturing & Service Operations Management* 14(4):512–528.

Ross MA, Aurora T, Graff L, Suri P, O?Malley R, Ojo A, Bohan S, Clark C (2012) State of the art: Emergency department observation units. *Critical Pathways in Cardiology* 11(3):128–138.

Roy AD (1952) Safety first and the holding of assets. *Econometrica* 20(3):431–449.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2012) Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research* 60(5):1080–1097.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2014a) Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management* 16(3):329–345.

Saghafian S, Hopp WJ, Van Oyen MP, Desmond JS, Kronick SL (2014b) Complexity-augmented triage: A tool for improving patient safety and operational efficiency. *Manufacturing & Service Operations Management* 16(3):329–345.

Shire (2013) Rare disease impact report: Insights from patients and the medical community. Technical report, Shire, URL `https://globalgenes.org/wp-content/uploads/2013/04/ShireReport-1.pdf`, Last accessed: 2016-10-02.

Shumsky R, Pinker E (2003) Gatekeepers and referrals in services. *Management Science* 49(7):839–856.

Shurtz I (2013) The impact of medical errors on physician behavior: Evidence from malpractice litigation. *Journal of Health Economics* 32(2):331–340.

Sklar DP, Hauswald M, Johnson DR (1991) Medical problem solving and uncertainty in the emergency department. *Annals of Emergency Medicine* 20(9):987–991.

Smith M, Higgs J, Ellis E (2008) Factors influencing clinical decision making. *Clinical Reasoning in the Health Professions*, 89–100 (Butterworth Heinemann Elsevier), 3rd edition.

Smith M, Saunders R, Stuckhardt L, McGinnis J, eds. (2012) *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America* (Washington, D.C.: National Academies Press).

Studdert DM, Mello MM, Gawande AA, Gandhi TK, Kachalia A, Yoon C, Puopolo AL, Brennan TA (2006) Claims, errors, and compensation payments in medical malpractice litigation. *New England Journal of Medicine* 354(19):2024–2033.

Sun BC, Hsia RY, Weiss RE, Zingmond D, Liang LJ, Han W, McCreath H, Asch SM (2013) Effect of emergency department crowding on outcomes of admitted patients. *Annals of Emergency Medicine* 61(6):605–611.

Tan T, Netessine S (2014) When does the devil make work? An empirical study of the impact of workload on worker productivity. *Management Science* 60(6):1574–1593.

Tang N, Stein J, Hsia RY, Maselli JH, Gonzales R (2010) Trends and characteristics of US emergency department visits, 1997-2007. *JAMA* 304(6):664–670.

van der Zee SP, Theil H (1961) Priority assignment in waiting-line problems under conditions of misclassification. *Operations Research* 9(6):875–885.

Viberg N, Forsberg BC, Borowitz M, Molin R (2013) International comparisons of waiting times in health care – limitations and prospects. *Health Policy* 112(1):53–61.

Wang X, Debo L, Scheller-Wolf A, Smith S (2010) Design and analysis of diagnostic service centers. *Management Science* 56(11):1873–1890.

Wilde J (2000) Identification of multiple equation probit models with endogenous dummy regressors. *Economics Letters* 69(3):309–312.

Zhang Z, Luh H, Wang C (2011) Modeling security-check queues. *Management Science* 57(11):1979–1995.

e-companion to
"Gatekeeping Under Uncertainty: An Empirical Study of Referral Errors in the Emergency Department"

Appendix EC.1: Comparison of Inpatient (Specialist) and CDU Efficiency

The results in our paper suggest that implementing an intermediate unit that exists between the ED and hospital inpatient units (the CDU in our case) where patients for whom there exists considerable diagnostic uncertainty can be admitted can help to reduce the number of unnecessary hospital admissions. However, we must also show that this intermediate unit can operate more efficiently than a standard inpatient unit else it offers little benefit (instead all patients who are currently referred in to the CDU could simply be admitted to the hospital). Here we compare these two alternatives.

Ignoring false discharges for which our analysis shows there to be no additional advantage of the CDU, in our sample of admitted patients there exist five classes of patient: those admitted from the ED to an inpatient bed who are (1) not false admissions or (2) are false admissions, and in addition those instead admitted to the CDU who are (3) discharged or are (4) admitted and subsequently not deemed to be a false admission or are (5) admitted and then classed as a false admission. Assume, conservatively, that for every patient who was admitted from the CDU (i.e., those of class (4) or (5)) all of the time they spent in the CDU was wasted, i.e., their LOS is not reduced at all despite the additional tests, better routing, etc. of patients after assessment in the CDU. For all 12,313 patients in our sample who enter the hospital via the CDU this thus adds up to 87,092 'wasted' hours. For the CDU to break-even, therefore, each of the 22,784 patients who are discharged from the CDU (i.e., those in class (3)) must have an average stay that is more than 3.8 hours shorter than it would have been if they had instead been admitted to the hospital.

To determine whether the condition above is satisfied, again we take a conservative approach and assume that if those patients who were discharged from the CDU had been admitted to the hospital instead then *all* of them would have been identified and discharged within 24h (with no treatment performed), i.e., they would instead have been false hospital admissions (i.e., of class (2)). Thus we need to compare the length of stay associated with patients of classes (2) and (3). In doing so we should account for differences in the characteristics of those patients admitted and subsequently discharged from the hospital directly rather than through the CDU, since e.g. the

former may be inherently more risky and hence more likely to stay longer. To do this, we construct an ordinary least squares (OLS) model that takes the form

$$LOS_i = \lambda_0 + \mathbf{W}_i\boldsymbol{\lambda}_1 + CDU_i\lambda_2 + \epsilon_i^{\lambda}\,, \tag{EC.1}$$

where $\epsilon_i^{\lambda} \sim \mathcal{N}(0, \sigma_{\lambda}^2)$ and $\mathbf{W}_i$ is a control vector that contains all of the temporal, diagnosis related and contextual controls from Table 10, as well as controlling for the age (using five-year age bands) and gender of the patient. This model indicates that a patient treated in the CDU would have spent 8.7 hours more in the hospital if they had instead been admitted directly, meaning that the hospital 'saves' 182,515 hours of time as a consequence of ED physicians referring these patients to the CDU rather than admitting them directly to the hospital. The longer processing time of patients in the hospital than in the CDU is not surprising, since once admitted to a general inpatient ward heterogeneity of the patient pool increases, while the CDU is specifically set up to route patients in to the hospital who require hospitalization and discharge those who do not.

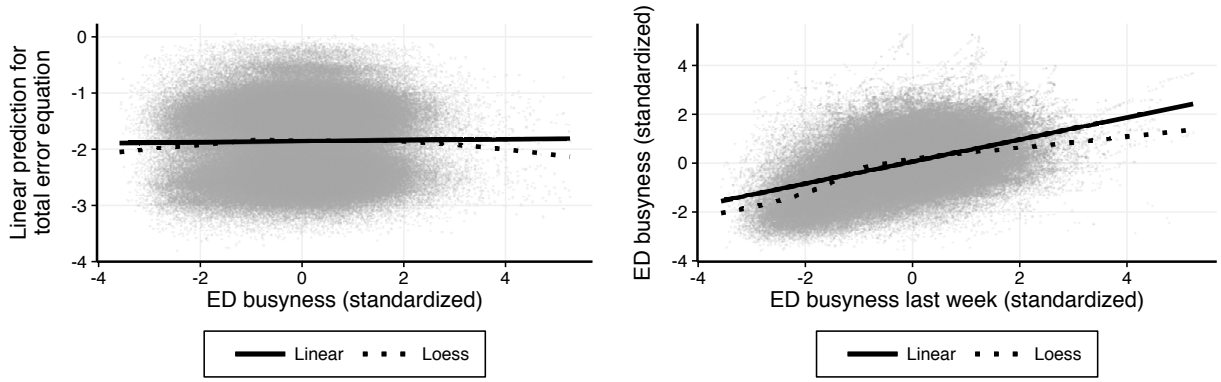
Combining the 'wasted' and 'saved' hours, we find the CDU saves, relative to hospital use, 95,423 hours over 1,840 days, reducing required capacity at our study hospital by approx. 2.2 beds (assuming 100% bed utilization). Put another way, over the sample period 251,581 hours (and the equivalent resources) were consumed by the CDU, however, a conservative estimate of the number of hours that would have been required had the CDU not been in place is 434,096 $(= 251,581 + 182,515)$. This implies an efficiency saving of approx. 42.0% $(= 1 - \frac{251,581}{434,096})$.

Appendix EC.2: Endogeneity of ED Busyness

In the paper we claim that the increase in the rate of false admissions (without a similar increase in false discharges) at higher levels of ED busyness can be explained as a consequence of physician risk aversion (i.e., a preference for 'safety-first') in the face of increased diagnostic uncertainty brought about by shorter service times. An alternative explanation that we must consider is that when the ED is more congested the reason why it is more busy is because of a surge in arrivals of patients who have a higher chance of being admitted (lower chance of being discharged) in error. We demonstrate here that this is not the case.

First, it is important to note that in our models we adjust for temporal variation with a range of controls (e.g., hourly dummies, day of the week, month, school holidays, trend, year) and also remove time periods where the patient population may significantly differ (e.g., public holidays, the Christmas period). Thus, the effects we observe is highly unlikely be a consequence of systematic time-related correlation between patient error propensity and ED busyness. Next, to see whether there is any evidence that patients differ in their error propensity at different levels of ED busyness

Figure EC.1 Plots with lines of best fit (linear and using locally weighted least squared (loess)) of predicted errors against ED busyness (left) and ED busyness against ED busyness from the previous week (right).



we plot in Figure EC.1 (left) the linear prediction from the total error equation estimated in Section 6.2, $\widehat{TotErr}_i^*$, against ED busyness, $zOccED_i$. The fitted values, $\widehat{TotErr}_i^*$, are calculated as a function observable factors associated with the patient that affect their risk of false admission or discharge. Since we would expect observable and unobservable risk factors to be correlated, any relationship between observed risk of admission or discharge error and ED busyness would be concerning. As can be seen in Figure EC.1 (left), however, there appears to be no association between them. Predicting $\widehat{TotErr}_i^*$ as a function of $zOccED_i$ in a standard ordinary least squares (OLS) model results in an R^2 of only 0.01%, suggesting that ED busyness essentially explains none of the variation in patients inherent error propensity. Thus, based on observables we have no reason to suspect that the profile of patients in the ED differs significantly when it is more or less busy.

To add further credibility to our findings, we perform a further robustness check and use an instrumental variable approach to correct for any potential endogeneity that results from unobservable factors being correlated with both ED busyness and error propensity. To do this we need an instrumental variable (IV) that is correlated with ED busyness but uncorrelated with a patient's likelihood of being admitted or discharged in error. We choose for this IV the busyness of the ED the week prior to the patients arrival in the ED. Clearly this should have no direct effect on the patient's likelihood of being admitted or discharged in error. However, to the extent that busy periods in hospitals tend to cluster (due to e.g., seasonal flu, heat waves, etc.), how busy the ED was in the previous week is expected to be correlated with ED busyness one week later. To check this, in Figure EC.1 (right) we plot ED busyness against ED busyness from one week prior (denote this $zOccEDPrevWk_i$). As can be seen there appears to be a positive association between the two variables, with correlation equal to $\rho = 0.455$, p-value < 0.0001. Thus, we take a two-step approach and first regress ED busyness against all of the exogenous covariates plus our IV using OLS, i.e.

$$zOccED_i = \omega_0 + \mathbf{X}_i\boldsymbol{\omega}_1 + \mathbf{Z}_i\boldsymbol{\omega}_2 + zOccEDPrevWk_i\omega_3 + \epsilon_i^{\omega}\,, \tag{EC.2}$$

where $\epsilon_i^{\omega} \sim \mathcal{N}(0, \sigma_{\omega}^2)$, and then substitute $zOccED_i$ in the selection and outcome equations (i.e. in Equations (1) and (3)) with the fitted values from the regression specified in Equation (EC.2), $\widehat{zOccED}_i$.[5] Doing so we estimate a coefficient (coef.) of 0.057 and APE of 0.56%, p-value $= 0.005$, (versus coef. of 0.035 and APE of 0.34%, originally) in the heckprob model for false admissions with censoring when a patient is admitted to the CDU, and an insignificant coef. of 0.011, p-value > 0.10, (versus coef. of -0.011 in the equivalent model for false discharges. Thus, if anything, based on unobservables as the ED becomes busy these findings suggest that patients become less, rather than more, likely to be admitted in error. This is not too surprising, because if we assume all patients with 'serious' conditions attend the ED anyway, then a surge in ED admissions is likely a result of an increase in the worried well, who all else being equal we would expect to be less likely to be admitted or discharged in error.

Appendix EC.3: Relevance and Validity of the Instruments

In this section formal testing is performed to assess the relevance and validity of the two instrumental variables (IVs) employed in the paper.

EC.3.1. Tests of Under- and Weak Identification

The underidentification test is a Lagrange multiplier (LM) test to determine whether the equation is identified. Specifically, the test determines whether the excluded instruments are correlated with the potential endogenous regressor, i.e. that the excluded instruments are "relevant" in the selection (first-stage) equation. "Weak identification", on the other hand, arises when the excluded instruments are correlated with the endogenous regressors, but only weakly. Estimators can perform poorly when instruments are weak: estimates may be inconsistent, tests for the significance of coefficients may lead to the wrong conclusions, and confidence intervals are likely to be be incorrect. Here we describe how we test for both of these properties.

First it is important to note that the majority of tests are based on a linear IV regression model where the dependent variable in the outcome equation and the endogeneous variable are continuous. In order to perform formal testing we therefore follow convention and treat the binary admission (discharge) error and CDU admission variables as continuous. While this means that the true critical values of the tests and significance levels may differ from those that are reported here, we note that differences in estimated parameters that arise from using a continuous rather than binary model specification are often small, and that the estimated coefficients using these models (not shown) are consistent with those reported in the main paper.

[5] The estimated coefficient for ω_3 is highly significant in Equation (EC.2) with p-value < 0.0001, suggesting that $zOccEDPrevWk_i$ is a good IV.

In testing for both underidentification and weak identification we use the method of Sanderson and Windmeijer (2016), implemented in and reported by the `ivreg2` command in Stata 12.1 (Baum et. al. 2010). The Sanderson-Windmeijer (SW) first-stage chi-squared Wald statistic is distributed as chi-squared with $(I_E - N_{EN} + 1)$ degrees of freedom under the null that the particular endogenous regressor of interest is underidentified, where I_E is the number of excluded instruments ($= 2$ here) and N_{EN} is the number of endogenous regressors ($= 1$ here). For the false admission model, the SW Chi-sq statistic is calculated to take a value of 825.8 with 2 d.f., which has corresponding p-value < 0.0001. For the false discharge model, the SW Chi-sq statistic takes value 885.6 with 2 d.f. and corresponding p-value < 0.0001. This means that there is strong evidence to reject the null hypothesis of underidentification in both cases at e.g. the 0.1% significance level, and so it is possible to conclude that the excluded instruments are "relevant".

Turning next to the issue of weak identification, the SW first-stage F-statistic is the F form of the SW chi-squared test statistic and can be used as a diagnostic for whether a particular endogenous regressor is "weakly identified". In particular, the F-statistic can be compared against the critical values for the Cragg-Donald F-statistic reported in Stock and Yogo (2005) to determine whether or not the instruments perform poorly. The relevant test has null hypothesis that the maximum bias of the IV estimator relative to the bias of ordinary least squares, i.e. $\left|\frac{\mathbb{E}[\hat{\beta}_{IV}]-\beta}{\mathbb{E}[\hat{\beta}_{OLS}]-\beta}\right|$, is b, where b is some specified value such as 10%. For a single endogenous regressor, assuming the model to be estimated under limited information maximum likelihood, the critical F-values are 8.68, 5.33 and 4.42 for maximum biases of $b = 10\%$, 15%, and 20%, respectively. If the estimated F-statistic is less than a particular critical value then the conclusion is that the instruments are weak for that level of bias. Here, the estimated SW F-statistic is equal to 412.7 for the false admission model, and equal to 442.6 for the false discharge model, indicating that the maximal bias is likely to be tiny. Thus we are not concerned that our models are affected by the problem of weak instruments.

EC.3.2. Testing for Overidentification

In addition to the excluded instruments being "relevant", it is also important to check that they are "valid", i.e. (1) uncorrelated with the error term (i.e., orthogonal to epsilon) and (2) correctly excluded from the outcome equation (i.e., only indirectly influence dependent variable y). The test for overidentification for the biprobit model uses the χ^2 statistic in a test of the joint significance of the instruments in the outcome equation. In particular, we include the instruments in both the selection and outcome equations and rely on identification based on the nonlinear functional form alone. The null hypothesis is that the instruments are not jointly significant in the outcome equation (Guilkey and Lance 2014, footnote 8, p. 31). For the false admissions biprobit model

$\chi^2 = 1.17$, p-value $= 0.557 > 0.10$, for the false discharge model $\chi^2 = 3.18$, p-value $= 0.204 > 0.10$. Together these results indicate no evidence of joint significance of the instruments, and hence we have no reason to suspect that they are not valid.

References

Baum CF, Schaffer ME and Stillman S (2010) ivreg2: Stata module for extended instrumental variables/2SLS, GMM and AC/HAC, LIML and k-class regression. URL `http://ideas.repec.org/c/boc/bocode/s425401.html`, Accessed: 2016-01-06.

Guilkey DK and Lance PM (2014) Program impact estimation with binary outcome variables: Monte Carlo results for alternative estimators and empirical examples. Sickles R, Horrace W, eds., *In Festschrift in Honor of Peter Schmidt*, 5–46 (New York: Springer).

Hansen L (1982) Large sample properties of generalized method of moments estimators. *Econometrica* 50:1029–1054.

Sanderson E and Windmeijer F (2016) A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics* 190(2):212–221.

Sargan J (1958) The estimation of economic relationships using instrumental variables. *Econometrica* 26:393–415.

Stock J, Yogo M (2005) Testing for weak instruments in linear IV regression. Andrews D, Stock J, eds., *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, 80–108 (Cambridge University Press).