Gatekeeping under Congestion: An Empirical Study of Referral Errors in the Emergency Department

Michael Freeman
INSEAD, 138676 Singapore, Republic of Singapore michael.freeman@insead.edu

Susan Robinson
Cambridge University Hospitals NHS Foundation Trust, Cambridge CB2 0QQ, United Kingdom
susan.robinson@addenbrookes.nhs.uk

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

Using data from over 350,000 visits to an emergency department (ED), we study the effect of congestion on the accuracy of gatekeeping decisions (hospital admission or discharge home) and the effectiveness of a second gatekeeping stage (a clinical decision unit (CDU)) in reducing errors. While ED physicians make more gatekeeping errors when congestion increases, the change in the rates of false positives (avoidable hospitalization) and false negatives (wrongful discharge) differ substantially. We find that when congestion increases, physicians prevent an increase in wrongful discharges – a more safety-critical concern – by lowering the threshold for hospital admission. This leads to a surge in avoidable hospitalizations and creates "false demand" for hospital beds at precisely the time when ED physicians should protect this constrained resource. We show that introducing a second gatekeeping stage – to which front-line gatekeepers can pass customers if they are unable to make an accurate referral decision – can mitigate this effect. When used as a second gatekeeping stage, we find evidence that the CDU reduces both avoidable admissions and wrongful discharges, by 16.5% and 13.8%, respectively. We also demonstrate that the two-stage gatekeeping system performs better than a combined system with pooled capacity.

Key words: gatekeeping; congestion; referral error; health care: hospitals; service operations; econometrics
History: September 5, 2017

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., a hospital emergency department (ED), general enquiries help-desk, local computer repair shop) from where they can be referred to a more specialized, and hence more costly, level of service (e.g., an acute hospital bed, complaints desk, central engineering department) according to the complexity of their needs. The upstream servers (e.g., ED physicians, telephonists, technicians) in

such a setting assume a dual role: They will service simple requests themselves, while at the same time acting as *gatekeepers* to downstream specialist units, thereby ensuring that customers receive the appropriate service intensity for their needs (Shumsky and Pinker 2003).

Empirical research in operations management has demonstrated that high utilization of specialist resources leads to a deterioration of system performance, resulting in delays (KC and Terwiesch 2009, Berry Jaeker and Tucker 2016, Chan et al. 2016), reduced service quality (Needleman et al. 2011, KC and Terwiesch 2012, Tan and Netessine 2014, Kim et al. 2014, Kuntz et al. 2015), and poorer financial performance (Powell et al. 2012). From a system perspective, it would therefore be desirable for gatekeepers to smooth demand variation by "rationing" access to specialists when demand surges. However, recent empirical evidence suggests that the opposite may occur: As congestion in the system increases, gatekeepers may *increase* the rate at which they refer customers to specialists (Freeman et al. 2016, Gorski et al. 2017). Are gatekeepers "opening the floodgates" to specialist services, i.e. referring greater numbers of customers to specialists who could instead be served directly by the gatekeepers themselves, at the very time when they should ration access to these services? If so, demand surges faced by upstream gatekeepers turn into relatively greater demand surges for more expensive downstream specialists, a demand amplification effect. This excess "false demand" on the specialist unit not only increases system costs but, by elevating utilization of specialist services, is likely to negatively affect the quality of service received by the "true demand" in the specialist unit.

We argue that this demand amplification effect is a natural behavioral response to congestion by gatekeepers who regard a "missed" referral (i.e. when a customer should have been sent to the specialist but is not) as a more serious error than an "unnecessary" referral. This is, to the best of our knowledge, the first empirical study to explore the trade-off between these two types of gatekeeping error. If the undesirable demand amplification effect is a consequence of differential error weights and worker behavior, which will be hard to affect directly, then system changes that reduce both types of error simultaneously are especially helpful. We study one such mechanism – a two-stage gatekeeping system – that counteracts the undesirable demand amplification effect in gatekeeping systems with asymmetric error weights.

Our empirical study uses a database of over 650,000 patient attendances at the ED of a large UK-based teaching hospital over a seven-year period. We use this data to study how variations in system congestion affects the rate of short-stay observational admissions (patients discharged within 24 hours of hospital admission with no treatment or procedure recorded in the electronic discharge record) and wrongful discharges from the ED (patients discharged from the ED who

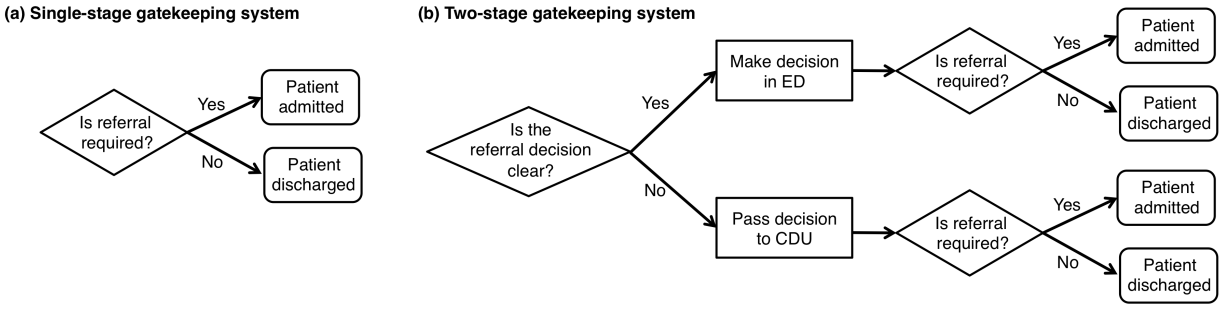
return to the ED within seven days and are then admitted to the hospital). We use *relative changes* in the rate of short-stay observational admissions as a conservative estimate of the *relative changes* in avoidable hospitalizations (see Section 5.2 for details).

We find that for every one standard deviation increase in ED congestion, the rate of avoidable hospitalization increases by 7.7%, while the rate of wrongful discharge reduces by 3.3% (the latter significant at the 10% level). Thus, when faced with higher congestion levels, ED physicians allow more patients into the hospital than necessary in order to mitigate the increased chance of a potentially catastrophic error in discharge. Surprisingly, in doing so, they may even over-compensate and adjust admission beyond the rate that would preserve the prevailing wrongful discharge rate. Importantly, the data provide evidence that the differential severity weight of the two types of error causes the aforementioned demand amplification effect: Rather than taming a demand surge in the ED by rationing access to scarce hospital resources more stringently, ED physicians lower the bar for admission and thereby increase the rate of avoidable hospitalization. The demand surge in the ED is thus amplified into a relatively larger demand surge in the hospital.

Having established the demand amplification phenomenon, the second part of this paper studies what can be done to mitigate it. We offer one potential solution: if the gatekeeper is unable to make a clear referral decision under the time and resource constraints, she can classify the customer as "unresolved". These unresolved customers are then passed to a second-stage gatekeeper who assumes responsibility for deciding whether a specialist referral is necessary. Since unresolved cases are more homogenous than the overall customer population (unambiguous referrals and non-referrals having been filtered out by the first-stage gatekeeper), they can be served by a more specialized workforce and targeted with customized resources that increase diagnostic accuracy and reduce referral errors.

As it happens, such a two-stage gatekeeping process already exists within our study hospital in the form of a clinical decision unit (CDU). This is a bedded unit attached to the ED to 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, but patients may stay up to 24 hours (rather than four hours in the ED) in the care of an experienced senior ED physician, who will eventually decide whether to discharge the patient or admit her into an acute hospital bed. The CDU thus provides front-line ED physicians with an alternative to the binary gatekeeping decision (hospital admission or discharge home) when the uncertainty of the decision is high but could be reduced substantially by the longer stay and the expertise in the CDU. A comparison of this two-stage process with the traditional gatekeeping set-up is shown in Figure 1.

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



After accounting for non-random assignment of patients to the CDU using appropriate sample selection methods, we show that the presence of the CDU reduces both avoidable hospitalization rates and wrongful discharge rates. We estimate that in our study hospital the presence of the CDU leads to the prevention of an avoidable hospitalization for 93 of every 1,000 patients routed through the unit (average treatment effect on the treated) and to the avoidance of a wrongful discharge for 12 of every 1,000 patients routed through the unit.

While it is perhaps unsurprising that making additional resources (the CDU) available for patients for whom the referral decision is uncertain will improve the accuracy of the decision for these patients, we note that the aggregate effect of the second gatekeeping stage in the combined system is not obvious as it comes at a cost. In particular, the second gatekeeping stage consumes resources (beds, staff time) that might instead be redeployed to the first stage to reduce congestion and thereby increase the accuracy of gatekeeping decisions. We demonstrate in Section 8 that this is not the case in our study hospital and that the second gatekeeping stage has a beneficial effect beyond that which could be achieved by redeploying resources back to the first gatekeeping stage. In fact, our data suggest that the CDU in the study hospital is under-capacitated and that referral errors could be reduced further if more resources were redeployed from the ED to the CDU.

2. Contribution to the Literature

This paper contributes to the operations management literature in two main ways. First, as an empirical study, it contributes to and complements the predominantly analytical work on (i) gatekeeping and referrals within multi-tier service contexts, (ii) the speed-quality trade-off in queuing systems with discretionary service completion, and (iii) resource pooling and partitioning. Second, the paper contributes to research on empirical healthcare operations as a first study of how organizational factors, specifically congestion, can affect errors in medical decision making, as well as introducing a process change that can help to mitigate these effects.

2.1. Gatekeeping

Gatekeeping systems are comprised of two service tiers, with the server in the first tier referred to as the 'gatekeeper'. Gatekeepers are typically generalists who can service a range of relatively simple customer needs. If those needs are too complex, they have the option to refer the customer to a more highly skilled and more costly second-tier specialist (Shumsky and Pinker 2003). In the operations management literature, early modeling work has addressed the question of how a system-optimal rate of referrals between the gatekeeper and specialist can be incentivized (Shumsky and Pinker 2003, Hasija et al. 2005). More recently, the framework has been extended and adapted to specific applications such as security-check queues (Zhang et al. 2011) and outsourcing decisions (Lee et al. 2012). This literature models gatekeepers as economic agents who maximize their time-average income from wages plus bonuses per-customer-diagnosed and per-customer-successfully-treated. Insights from this research are not easily transferable to contexts in which gatekeeping decisions are not economically motivated but may instead, for example, follow professional and social norms, as is likely the case for salaried ED physicians. In such a context, empirical or experimental studies may provide new insights into gatekeeping behavior. To date, such studies are rare (see e.g. Freeman et al. 2016, Gorski et al. 2017) and the effects of environmental conditions on the accuracy of gatekeepers' referral decisions have not been addressed. These effects are important, however, as gatekeeper referral errors are both costly (e.g. resulting in unnecessary specialist referrals) and may lead to worse outcomes (especially when a necessary referral is missed). We contribute to the emerging gatekeeping literature by studying the effect of congestion on gatekeeping errors in a context where gatekeepers are professionally rather than financially incentivized.

2.2. Speed-quality tradeoff

Our paper also contributes to the literature on the speed-quality trade-off in queuing systems when servers have discretion over task completion (e.g. Hopp et al. 2007, Anand et al. 2011, Kostami and Rajagopalan 2013). As with the gatekeeping literature, most of this work is analytical and empirical studies (e.g. Tan and Netessine 2014) are rare. Discretionary service completion in queuing systems can lead to surprising results, in particular in relation to overtreatment, which is where our study interacts most closely with this stream of literature. For example, Hopp et al. (2007) find that, in contrast to standard queuing systems, increasing capacity when service completion is discretionary may, in fact, increase congestion as a result of additional service components being added when servers are under light load. Wang et al. (2010) extend these results to a decentralized context where servers in diagnostic centers trade off diagnostic accuracy and congestion, and explore the effects of asymmetric error costs, and Alizamir et al. (2013) characterizes the optimal policy for

the diagnosis of customer types (e.g. patients requiring hospitalization or not) when servers can decide to perform additional diagnostic testing to resolve type uncertainty. Focusing on services for which customers cannot themselves ascertain their needs (as is often the case in healthcare), Debo et al. (2008) demonstrate analytically that queuing dynamics can create heterogeneity in the customer base that can be exploited to induce additional service when arrival rates are low. More recently, however, Paç and Veeraraghavan (2015) show that congestion may act as a deterrent to such overtreatment. In contrast to these studies, which are all concerned with single-tier queuing systems and provide analytical insights, we study a two-tier gatekeeping system and offer empirical observations. Specifically, we show that in contrast to the analytical results for the single-tier case, upstream congestion in the two-tier case *induces* overtreatment when gatekeepers weigh a missed referral as a more serious error than an unnecessary referral.

2.3. Pooling and partitioning

Our study of the effect of adding a second gatekeeping stage on the propensity for gatekeeping errors is naturally related to the literature on resource pooling and partitioning. In fact, the two-stage gatekeeping system is conceptually similar to a priority queue with two classes of patients, with either high or low levels of diagnostic uncertainty. Queueing theory suggests that dividing customers into different classes may be beneficial when customers differ sufficiently in their service requirements (see e.g. Mandelbaum and Reiman 1998, Dijk and Sluis 2008). However, in contrast to our setting, where customers are streamed by residual diagnostic uncertainty, these queuing studies stream customers based on processing times (see also Hu and Benjaafar 2009).

More closely related to our work is a series of recent papers on triage. While triaging has traditionally prioritized customers based on levels of urgency (see e.g. FitzGerald et al. (2010) for an excellent overview), recent analytical studies in the operations management literature have explored ways in which the basic triage process might be improved by 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 are two modeling papers that study the ED triage process (Saghafian et al. 2012, 2014). These propose augmenting triage by segmenting ED patients based not only on severity but also using their (i) likelihood of being admitted, and (ii) complexity (i.e. the likely duration of the diagnostic process), respectively. Saghafian et al. (2017) also use a modeling approach to identify the impact of allowing nurses to offload triage decisions to more experienced telemedical physicians, extending the standard single-stage triage process to a two-stage process. While our paper complements these studies with an empirical examination, our context differs in two important ways. First, a two-stage gatekeeping process streams patients into the second stage during

service itself, while triaging puts patients into a specific queue before the start of service. We therefore study the effect of congestion on ED physicians' admission decisions rather than on the typically much faster triage decision made by triage nurses (Saghafian et al. 2017). Second, our outcomes of interest differ from the prevailing average cost and waiting time concerns in that our study is a first empirical investigation of congestion effects on admission and discharge errors.

2.4. Empirical healthcare operations

In addition to the predominantly analytical operations management literature reviewed above, our study fits naturally into a stream of empirical papers on healthcare operations. Much of this literature is concerned with the impact of organizational factors, such as congestion, on clinical, operational and financial outcomes, specifically on clinical safety (e.g. KC and Terwiesch 2012, Kim et al. 2014, Kuntz et al. 2015), service times (e.g. KC and Terwiesch 2009, Berry Jaeker and Tucker 2016, Chan et al. 2016), queue abandonment (Batt and Terwiesch 2015), and reimbursement (Powell et al. 2012). Of specific relevance is the work on patient routing into specialist services. In two studies of intensive care units (ICUs), KC and Terwiesch (2012) and Kim et al. (2014) show that ICU staff block admissions and discharge patients prematurely when their specialist unit becomes congested. While this behavior does not avoid deterioration of system performance, as evidenced by increased ICU readmission rates, it does ration access to congested services to the most needy patients. In contrast to these studies, we focus on upstream congestion faced by gatekeepers who refer patients to specialist services (acute hospital beds in our case). We find that the rationing pattern observed in KC and Terwiesch (2012) and Kim et al. (2014) is reversed upstream: When gatekeepers become busy, they refer *more* patients than necessary to the specialist unit, thus increasing congestion downstream. This behavior has been observed elsewhere. For example, Freeman et al. (2016) show that midwives who act as gatekeepers to specialist obstetricians refer high-complexity patients to obstetricians at higher rates in the presence of congestion. Gorski et al. (2017) show that hospital admissions rates from the ED increase with congestion. Building on these studies, we provide evidence that error-avoidance behavior specific to healthcare – an emphasis on avoiding missed referrals in the interest of a "safety first" principle – will naturally lead to increased unnecessary referrals under congestion, and then show that a second gatekeeping stage can reduce both forms of error, and thereby mitigate the over-admission phenomenon.

3. Setting Description

3.1. The emergency department

The ED at our study hospital is visited, on average, by 250 patients per day and operates in a manner similar to the majority of hospitals in the US, UK and worldwide. Patients self-present

or arrive by ambulance with a variety of complaints and symptoms, some of which can be easily managed in the ED (e.g., wound suturing, casting, splinting), while others are complex and clearly require admission to the hospital for specialized, longer-term care (e.g., hip fracture, multiple trauma, heart attack, stroke). Many patients, however, present with symptoms that could either be caused by a minor ailment or be the sign of a more serious and even life-threatening condition (e.g. chest or abdominal pain). These patients require careful diagnosis before an admission or discharge decision can be taken.

After a patient arrives, the degree of urgency is assessed by a triage nurse. Unless the patient needs immediate attention, they register and wait to be seen for further assessment by an ED physician. The physician may order diagnostic tests (e.g. blood tests, imaging) and may consult a specialist in the hospital. If, after assessment, the physician determines that the patient requires a level of care beyond that which can be provided in the ED, she can admit the patient to an acute bed in the hospital. Otherwise, after treating the symptoms, the patient will be discharged and may be advised to arrange an outpatient, ambulatory or primary care follow-up appointment. ED physicians thus act as gatekeepers to expensive hospital inpatient beds and ration access by admitting only those patients whose needs cannot be met in the less resource-intensive ED setting (Blatchford and Capewell 1997).

Congestion in EDs – referred to as crowding in the medical literature – is common and has worsened over the past decade as capacity has failed to keep pace with the growth in attendance (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). Mitigating ED congestion is a significant policy concern and countries have adopted a wide range of interventions designed to manage the problem (Boyle et al. 2012). 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 decision units (Pines et al. 2011). Yet these approaches have met with only limited success. In England, February 2016 statistics, 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). Emergency departments are therefore a fitting context from which to study gatekeeping behavior under congestion.

3.2. Gatekeeping challenges in the emergency department

While playing a crucial role as gatekeepers to expensive inpatient beds, ED physicians must make decisions under time-pressure and often there exists significant uncertainty in the medical diagnosis. Consequently, they may on occasion make an error in diagnosis. Graber (2013), for example, estimates that one in ten medical diagnoses made in EDs are inaccurate, with errors in the diagnostic process being the leading cause of internal investigations and claims of malpractice within the ED (Kachalia et al. 2007, Cosby et al. 2008). When physicians are exposed to elevated levels of congestion, even less time is available for diagnosis, increasing diagnostic uncertainty and the likelihood that an error will be made. It thus follows intuitively that gatekeeping referral errors (i.e. an avoidable hospitalization or wrongful discharge) will be affected by the congestion level in the ED through the diagnostic process.

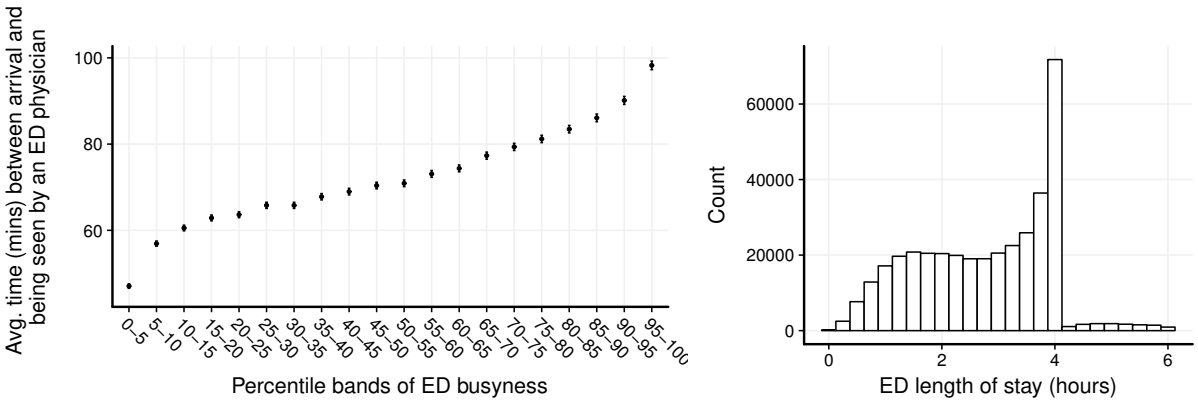
To demonstrate how ED congestion translates into reduced time available for diagnosis, we plot in Figure 2 (left) the mean time between ED arrival and the patient's first contact with a physician as a function of ED congestion. Each point in the plot corresponds to one of 20 percentile bands of ED congestion of width 5%. (Note that ED congestion is adjusted for differences across hours of the day and various other time-related factors using a method described in Section 5.1). As congestion in the ED increases, the average time between a patient's arrival and their first contact with a physician increases from under 50 minutes to over 95 minutes.

In England, the time pressure caused by congestion is further exacerbated by the government's 4-hour waiting-time target, which requires that 95% of ED patients must leave the ED within four hours of arrival. Towards the end of our study period, failure to achieve this target in any month attracted a fine of £200 per breach (NHS 2013), which could amount to between £75,000 (5% breaches) and £300,000 (20% breaches) per month for the ED in our study hospital. As a consequence, 4-hour target breaches were taken seriously, as shown in Figure 2 (right). This meant that any delay in the start of treatment effectively translated into a direct reduction of the time available to spend with the patient.

Together, the plots in Figure 2 show that the average available service time before breaching the 4-hour target is shortened by nearly 25% as ED congestion varies from the first to the last percentile band in the graph, falling from approximately 190 minutes to approximately 145 minutes. We expect this congestion induced shortening of service times to have a direct effect on the ED physicians' referral behavior in the study hospital.

In addition to a significant degree of time pressure and diagnostic uncertainty, the ED context offers a further characteristic that is relevant to the phenomenon that we wish to study: When ED

Figure 2 **(Left) Mean time between patient arrival at the ED and being seen by an ED physician as a function of ED congestion, with 95% confidence bands; (Right) Histogram of ED length of stay.**



physicians are faced with a gatekeeping decision under uncertainty, they do not weigh admission errors and discharge errors equally. Instead they typically adopt a "safety-first" principle, preferring to minimize the risk of a wrongful discharge (i.e. that the patient leaves untreated) over the risk of an avoidable hospitalization (Roy 1952). To quote a physician from our study hospital: "No-one has ever been sued for admitting a patient to a hospital."

3.3. The clinical decision unit

The clinical decision unit (CDU) (also known as an observation unit) is a dedicated bedded area that is separate from the main ED but is organizationally integrated with the ED and staffed by ED physicians and nurses. 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 that can be resolved within 6-24 hours, with further assessment determining whether inpatient admission is required at the end of their CDU stay (Hassan 2003). Various clinical and operational advantages of CDUs have been documented in the literature, including improved patient satisfaction, safety and length of stay (see Cooke et al. 2003, for an excellent survey) as well as considerable cost savings, estimated by Baugh et al. (2012) at $3.1 billion per year in the US. However, the benefit of a CDU to regulate admission and discharge error rates in the presence of congestion has, to our knowledge, not yet been examined.

4. Hypothesis Development

4.1. Gatekeeping under congestion

ED physicians are well aware of the level of congestion in their ED, both through direct visual cues and from information provided by IT systems that show, for example, the list of waiting patients with their registration details and triage information. Following Hopp et al. (2007), we assume that

ED physicians will exercise a degree of discretion over the time they spend with their patients. They reduce service times when the system becomes congested since the opportunity cost of time spent with their current patient increases against the alternative of completing the service and reducing the length of the queue.

Our interviews in the study hospital confirmed the view that ED physicians trade off the amount of time spent with an individual patient with throughput concerns. When service times are reduced in response to increased congestion, physicians have less time available to assess a patient and acquire all of the information necessary to make accurate gatekeeping decisions (Smith et al. 2008, Alizamir et al. 2013). In addition, as congestion increases, ED physicians will have to care for more patients simultaneously (KC 2014), leading to increased decision density and cognitive overload. The work of ED physicians relies on intuition and heuristics (Croskerry 2002) and cognitive overload can render these cognitive shortcuts ineffective, resulting in high levels of preventable errors (Leape 1994). For example, in a study of 100 cases of diagnostic errors Graber et al. (2005) found that cognitive factors contributed in 74% of cases.

In summary, ED physicians are examples of *congestion-sensitive gatekeepers* who (i) are aware of congestion levels in the gatekeeping system, and (ii) adjust their service to trade off the time spent with individual customers against improved system throughput. As congestion increases, congestion-sensitive gatekeepers put more emphasis on increasing throughput and therefore reduce the time taken to assess a customer's needs. This, together with the reduction in available service time shown in Figure 2, leads to increased diagnostic uncertainty when gatekeeping decisions are made and therefore to more gatekeeping errors. This negative congestion effect is amplified when these gatekeepers operate in multitasking environments, as congestion will increase the need to parallel process customers, which can lead to cognitive overload and make gatekeepers more error-prone.

Hypothesis 1. *As system congestion increases, congestion-sensitive gatekeepers make more errors in their referral decisions.*

4.2. Trading off referral errors

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 also to reputation damage and peer disapproval (Leape 1994). The costs (financial or otherwise) that physicians associate with these concerns will affect how they trade off false positives (avoidable hospitalization) and false negatives (wrongful discharge) in their gatekeeping decision. The overtreatment phenomenon in healthcare suggests that medical professionals, when faced with

uncertainty, will more often choose to do more rather than less (Gawande 2015). Specifically, unnecessary referral to specialists occurs more frequently than missed referrals (Bunik et al. 2007). The threat of litigation is frequently cited as a cause of this phenomenon, and medical professionals have been shown to refer patients more frequently to higher intensity care when they perceive a risk of undertreatment (Shurtz 2013). We can therefore assume that ED physicians associate a wrongful discharge as having a higher cost than an avoidable hospitalization.

Gatekeepers who have asymmetric disutilities for the two types of errors can trade them off against one another. If they reduce the thresholds for one type of decision, say for specialist referrals, they will make more unnecessary referrals. At the same time, however, their rate of missed referrals will be reduced because in cases of doubt they are now more likely to refer. As congestion increases, the overall error propensity increases, in accordance with Hypothesis 1. It is therefore rational for the gatekeeper to reduce the decision threshold for the less severe error to protect from an increase in the rate of the more severe error. This leads to a relative increase of the less severe error rate.

Hypothesis 2. *If congestion-sensitive gatekeepers weigh one type of referral error more heavily than the other, the proportion of the more heavily weighted error as a percentage of total gatekeeping errors will fall with system congestion.*

In our context, since we expect ED physicians to weigh the cost of a wrongful discharge more heavily than an avoidable hospitalization, we therefore anticipate an increase in the proportion of avoidable admissions relative to wrongful discharges.

4.3. The two-stage gatekeeping system

In congested gatekeeping systems, the speed-quality trade-off becomes a fundamental concern: Should the gatekeeper spend more time to assess the patient at hand before taking a referral decision, thereby increasing decision quality, or should she take a decision earlier, with less information, in the interest of speed and increased throughput? Much of the gatekeeping literature is concerned with the incentivization of gatekeepers to make such trade-off decisions in the interest of maximizing overall system performance (Shumsky and Pinker 2003). From a system design perspective, however, the question arises how one can shift the entire frontier so that gatekeepers can make fewer referral errors.

There are two interventions that could achieve this: (1) an increase in system capacity, and (2) the replacement of less experienced by more experienced gatekeepers, who can make more accurate referral decisions with less information. While both approaches would shift the speed-quality frontier, they each come at a cost: increasing capacity requires the hiring of staff; more

experienced gatekeepers demand higher wages. In addition, it is not obvious that these changes would have as much of an impact as desired. The more experienced gatekeepers would spend a significant proportion of their time working "below the top of their license" by assessing 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 staff. Similarly, there is no guarantee that any increase in capacity would be used only to attend to those patients whose diagnosis is unresolved, as e.g. servers may add discretionary components to the service of unambiguous cases (Hopp et al. 2007, Debo et al. 2008).

A better approach, therefore, would be to *improve the match between customers with heterogeneous needs with gatekeepers with heterogeneous experience and resources.* This is the basic idea behind the two-stage gatekeeping system.

Figure 1 illustrates the difference between a single-stage and a two-stage gatekeeping system. In a single-stage gatekeeping system, the front-line gatekeeper is required to take a binary decision to either refer the customer to a specialist or finish the service for the customer herself. In a two-stage gatekeeping system, the front-line gatekeeper has an additional decision option. When she realizes that the referral decision is beyond her capacity because, e.g., the residual uncertainty is too high, she can refer the patient to a second-stage gatekeeping unit. This second-stage gatekeeping unit can then be provided with resources that are more focused on the needs of these more complex types of cases. Examples include extra time, specialized testing equipment, or specially trained or experienced gatekeepers. The second stage gatekeeper, who is not exposed to the congestion-induced time pressure faced by the first-stage gatekeeper, can then make a better-informed referral decision at a later stage. Since a two-stage gatekeeping system offers a better match of gatekeeper experience and other service resources with the difficulty of the gatekeeping decision, we expect it to reduce gatekeeping errors.

Hypothesis 3. *A two-stage gatekeeping system reduces both types of gatekeeping errors.*

The concept of a two-stage gatekeeping system is akin to but also different from that of complexity-augmented triage proposed in Saghafian et al. (2014), which recommends first triaging patients who arrive at the ED on the relative complexity of diagnosis and then on their degree of urgency. While the triage process operates at the front-end of a queuing system, focusing on channeling customers to the most appropriate queue, the gatekeeping process is part of the service provision at the back end. The two-stage gatekeeping system therefore embeds the complexity assessment within the first-stage gatekeeping process and provides an additional decision option for the server, i.e. to refer a patient to the second stage.

Table 1 Descriptive statistics and correlation table.

		Mean			Correlation table			
	N	All	CDU = 0	CDU = 1	(2)	(3)	(4)	(5)
(1) Total gatekeeping errors (%)	373,663	5.30	5.14	6.84				
(2) Short-stay obs. admission (%)	373,663	4.25	4.17	5.04	0.89^{***}			
(3) Wrongful discharges (%)	373,663	1.05	0.97	1.80	0.44^{***}	-0.02^{***}		
(4) CDU admission (%)	373,663	9.77	0.00	100.00	0.02^{***}	0.01^{***}	0.02^{***}	
(5) ED congestion	373,663	−0.01	−0.01	−0.00	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 congestion equal to 1.01, 1.01 and 1.02 for 'All', 'CDU = 0' and 'CDU = 1', respectively; Correlation coefficients significant with $^{***}p < 0.001$, else $p > 0.05$.

5. Data Description and Variable Definitions

The data for our study is comprised of detailed information relating to 651,041 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. 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 approximately the same rate over the sample period (by 4.7% per annum (p.a.) for admissions versus 4.1% for discharges).

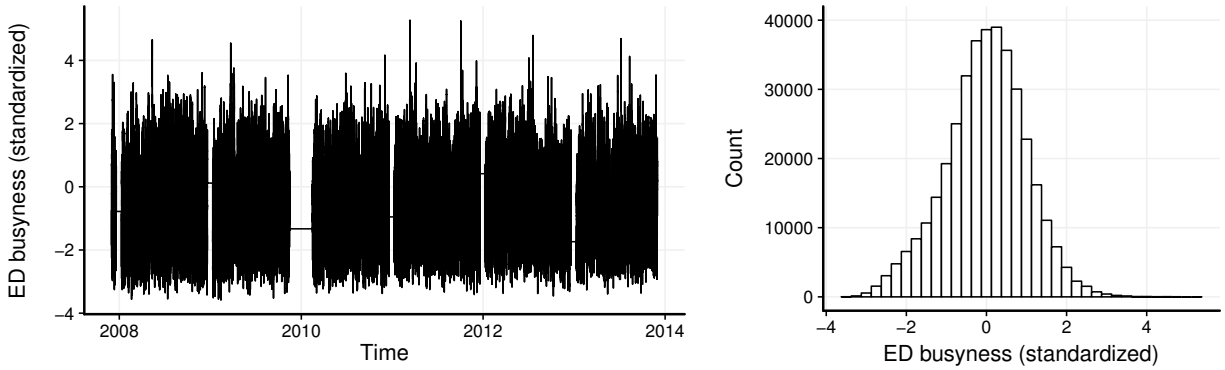
To prepare the data for analysis, we performed an initial cleaning round to ensure, as far as is possible, that our results are not affected by various data or time-related confounds. This included dropping: (i) ∼8.5k obs. from December 2013, since data entry may not have been completed; (ii) ∼11.5k obs. with missing or incomplete data; (iii) ∼17k obs. for patients who left against medical advice, died in the ED or were transferred to another hospital; and (iv) ∼127k obs. corresponding to children under the age of 16, who cannot be admitted to the CDU. We then use this data set to generate various variables of interest (see later), before: (v) excluding ∼60k obs. from the first year of data, the warm-up period for generating these variables; and (vi) removing dates close to public holidays when demand and staffing patterns vary significantly. This process is described in full in Appendix A. After this, we were left with 373,663 observations to take forward for analysis.

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

5.1. ED congestion

Our main independent variable of interest is the level of ED congestion that patients experience when they arrive in the ED. To generate this measure for patient i, we first determine which

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



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, denoted by $QueueED_i$.

It is well known that levels of congestion in EDs vary throughout 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 should adjust $QueueED_i$ to account for these differences. In other words, we are only interested in a variation of congestion levels that cannot be explained by seasonal predictors. We achieve this by employing a variation on the approach used in Kuntz et al. (2015) and Berry Jaeker and Tucker (2016) which establishes an approximation of available capacity. Specifically, we estimate capacity using quantile regression to predict the 95th percentile level of occupancy at hour h, $QueueED_h^{95th}$. The dependent variable in this regression is the average occupancy level at every hour h, starting from midnight on 1st January 2007 and ending at 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 is between January 2011 and December 2013, and zero otherwise.

The fitted values from the quantile regression model provide us with our estimate of capacity at each hour h, $CapacityED_h = \widehat{QueueED}_h^{95th}$. ED congestion, $EDCong_i$, can then be expressed as the ratio of observed occupancy to estimated capacity, i.e. $QueueED_i/CapacityED_{h_i}$, where h_i is the hour of arrival of patient i. Finally, we normalize by subtracting the mean, $\mu(EDCong_i)$, and dividing through by the standard deviation, $\sigma(EDCong_i)$, to form $zEDCong_i$. Plots of $zEDCong_i$ are provided in Figure 3.

5.2. Admission and discharge errors

The two dependent variables of interest in our analysis capture errors made in referral (admission) and non-referral (discharge) decisions by ED physicians. An admission error (or 'avoidable hospitalization') 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. We estimate that avoidable hospitalizations cost the NHS in England over £600 million in the 2012-13 financial year.[1]

While we cannot observe avoidable hospitalizations directly, we observe short-stay observational admissions, defined as patients who are discharged within 24 hours of being admitted to the hospital from the ED or CDU without treatment or procedure performed. The second of these conditions is met if there is no OPCS-4.6 (HSCIC 2013) intervention or procedure code – the UK equivalent of the American Medical Association's CPT coding system – associated with the post-admission inpatient record. The average rate of short-stay observational admissions for the full sample of 373,663 visits is 4.3% and for the 116,125 visits which resulted in admission is 13.7%.

While most avoidable hospitalizations will be among these short-stay observational admissions, we are not arguing that all short-stay observational admissions are unnecessary. However, avoidable hospitalization does occur (Denman-Johnson et al. 1997, Burgess 1998, Cooke et al. 2003). Therefore, although the rate of short-stay observational admissions is misleading as a measure of admission errors, we believe that a *change* in short-stay observational admission rates, e.g. in response to ED congestion, will be largely caused by a commensurate change in avoidable hospitalizations. Formally, we assume that the rate of short-stay observational admissions, r_{ObsAdm}, is the sum of the rate of necessary hospitalizations, r_{Adm}, and the rate of avoidable hospitalizations, r_{AdmErr}, and that r_{Adm} does not vary with the ED congestion level c, while the other two rates do vary with c. Under this assumption $r'_{\mathrm{AdmErr}}(c) = r'_{\mathrm{ObsAdm}}(c)$, i.e. an estimated change in short-stay observational admission rates estimates the change in the rate of avoidable hospitalizations. Moreover, $\frac{r'_{\mathrm{AdmErr}}(c)}{r_{\mathrm{AdmErr}}(c)} = \frac{r'_{\mathrm{ObsAdm}}(c)}{r_{\mathrm{AdmErr}}(c)} \geq \frac{r'_{\mathrm{ObsAdm}}(c)}{r_{\mathrm{Adm}}+r_{\mathrm{AdmErr}}(c)} = \frac{r'_{\mathrm{ObsAdm}}(c)}{r_{\mathrm{ObsAdm}}(c)}$, i.e. estimated relative changes in short-stay observational admission rates will be conservative estimates of relative changes in avoidable hospitalizations.

Note that both congestion and case-mix in the ED, and hence the rate of necessary hospitalizations, vary both by day of the week, time of the day, or seasonally over the year, so that the

[1] 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).

assumption that the rate of necessary hospitalizations is independent of congestion c is not tenable in general. However, we believe the assumption is justified for our measure of congestion levels $c = zEDCong$ (see Section 5.1), which is already adjusted for temporal factors that explain variation in ED congestion. We discuss this point further in Section 6.4.

Discharge errors (or 'wrongful discharges') are somewhat easier to observe in our data. These patients often come back to the ED 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 among patients with acute myocardial infarction who were inappropriately discharged from the ED to be 1.9 times higher than among hospitalized patients. We record ED patients as a wrongful discharge if, after discharge 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 wrongful discharges in the full sample is 1.0% and is 1.5% for the subset of 257,538 discharged patients.

Note that Hypothesis 1 states that the overall rate of gatekeeping errors increases with ED congestion c. The total error rate r_{TotErr} is the sum of the rate of avoidable hospitalizations r_{AdmErr} and wrongful discharges r_{DisErr}. As before, we use the change in the rate of short-stay observational admissions in lieu of avoidable hospitalizations, i.e., $r'_{\text{TotErr}}(c) = r'_{\text{AdmErr}}(c) + r'_{\text{DisErr}}(c) = r'_{\text{ObsAdm}}(c) + r'_{\text{DisErr}}(c)$. This also means that the estimated relative change in our measure is a conservative estimate of the relative change in total gatekeeping errors, i.e., $\frac{r'_{\text{TotErr}}(c)}{r_{\text{TotErr}}(c)} = \frac{r'_{\text{AdmErr}}(c)+r'_{\text{DisErr}}(c)}{r_{\text{AdmErr}}(c)+r_{\text{DisErr}}(c)} \geq \frac{r'_{\text{ObsAdm}}(c)+r'_{\text{DisErr}}(c)}{r_{\text{Adm}}+r_{\text{AdmErr}}(c)+r_{\text{DisErr}}(c)} = \frac{r'_{\text{ObsAdm}}(c)+r'_{\text{DisErr}}(c)}{r_{\text{ObsAdm}}(c)+r_{\text{DisErr}}(c)}$.

5.3. Control variables

In addition to the primary variables described above, we also have available a large number of control variables that allow us to account for heterogeneity in the patient population and in the hospital, which may be correlated with the dependent variables, and/or with the main independent variables of interest. These are reported in Table 2 and capture patient demographics, 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 variables of interest (and so through omission may bias the results) will be accounted for using appropriate empirical methods as described in Section 6.1.

A control to be highlighted when discussing our empirical strategy is those variables that capture the historic short-stay observational admission, wrongful discharge, or total error rates of the assigned physician. These account for the fact that particular physicians may have a greater propensity to make errors than others, and approximately speaking are calculated as the average

Table 2 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
Patient and diagnosis related factors ($\mathbf{D}_i$)		
Age bands	Categorical (10)	The age of the patient, split into 10-year age bands (e.g., 10-20, 20-30, 100+)
Gender	Binary	A variable equal to one if the patient is male, else zero
Triage category	Categorical (7)	The triage level assigned to the patient on ED arrival
Initial severity assessment	Categorical (7)	The nature of the patient's condition (e.g., minor injuries, requires resuscitation, etc.)
Reason for ED visit	Categorical (32)	The reason for the ED episode (e.g., fall, burn, road traffic accident, etc.)
Contexual factors ($\mathbf{C}_i$)		
Mode of arrival	Categorical (8)	The mode of transport used to get to the hospital (e.g., helicopter, 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 proportion of hospital admissions to ED visits in the previous 12 months
Admissions per ED visit, last month	Continuous	The proportion 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 ($\mathbf{P}_i$)		
Historic physician error rate	Continuous	The short-stay observational admission, wrongful discharge, or total gatekeeping errors propensity of the assigned physician, calculated as in Appendix B
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 ($\mathbf{O}_i$)		
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 5.1

Notes: 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.

case-mix adjusted rates of each of these errors made by each physician over the preceding year (see Appendix B for a full description of the calculation of these variables).

6. Models and Results I: Response to Congestion

We focus first on testing Hypotheses 1 and 2 (the effect of congestion on gatekeeping errors) and return to Hypothesis 3 (concerned with estimating the effect of the CDU) in Section 7. In testing the first two hypotheses the presence of the CDU is a confounding factor in the econometric analysis. Specifically, we would like to identify how ED congestion impacts referral decisions made directly by ED physicians, i.e. corresponding to only those patients not passed to the CDU. This means we want to study the upper half of the two-stage gatekeeping process shown in Figure 1. However, as congestion increases, so too might the rate at which ED physicians leverage the CDU option, and so too the composition of the patients for whom the physicians take the hospital admission or ED discharge decision themselves might change. While we account as far as possible for these differences with our set of controls (reported in Table 2), there may still exist factors unobservable to us, but observable to the physician (e.g., patient fitness level, medical history) that influence whether or not the physician leverages the CDU option. Thus, despite only 9.8% of patients being passed to the CDU, it will be necessary to ensure that our findings are not confounded by unobserved

differences in the patient case-mix arising from changes in CDU usage as congestion levels increase. In this section, we describe the empirical approach used to resolve this endogeneity concern.

6.1. 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 to the CDU. The second determines whether or not a patient is admitted as a short-stay observational admission and/or wrongfully discharged, 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

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

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

where $\epsilon_i^{\delta} \sim \mathcal{N}(0,1)$, CDU_i^* is a latent variable, the vector $\mathbf{X}_i$ contains the set of all controls (reported in Table 2), the vector $\mathbf{Z}_i$ contains the set of instrumental variables (to be described in Section 6.2), 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

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

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

where $\epsilon_i^{\beta} \sim \mathcal{N}(0,1)$, and where $ObsAdm_i^*$ and $ObsAdm_i$ are the latent and observed variables for short-stay observational admissions, respectively. The latent variable equation for wrongful discharges ($DischErr_i$) and total gatekeeping errors ($TotErr_i$) is the same as for short-stay observational admissions, with coefficient vector $\boldsymbol{\beta}$ replaced with $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, respectively.

Rather than estimate the first and second stage models described above individually, 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 correct for potential sample selection bias arising from the fact that (a) patients may not be assigned to the CDU at random and (b) the coefficients, and in particular the coefficient of interest, $zEDCong$, may vary depending on whether the patient was admitted or not to the CDU. This modeling approach is necessary as we are primarily interested in the effect of ED congestion on outcomes for those patients who were admitted or discharged by an ED physician directly (rather than by a physician in the CDU). In order to estimate the heckprob model we must: (1) censor the outcome variable $ObsAdm_i$, $DischErr_i$ or $TotErr_i$ whenever $CDU_i = 1$, (2) set $\alpha_2, \beta_2, \gamma_2 = 0$ in the outcome equation, and (3) 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.[2]

6.2. Instrumental variables

While the heckprob model can be estimated without instrumental variables (IVs), estimation is improved and coefficients are 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., be relevant), but not affect the rate of short-stay observational admissions, wrongful discharges, or total gatekeeping errors, and so do not appear in the outcome equation (i.e., be valid). We use two IVs, included in the vector $\mathbf{Z}_i$. Summary statistics for these IVs are available in Table 3.

The first IV is the CDU admission propensity of the assigned physician. This is calculated in the same way as the physician's historic error propensity (as mentioned in Section 5.3 and described in Appendix B), and approximately 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 that they treated. A patient assigned to a physician who is predisposed to admit patients to the CDU will be more likely to be sent there, satisfying the relevance condition. A potential issue with this IV is that being assigned to a physician who is more likely to admit to the CDU may also affect the likelihood of that patient being admitted or discharged in error, since physician rates of CDU referral and error may not be independent. To account for this we add a control for the physician's historical short-stay observational admission, wrongful discharge, or total gatekeeping error propensity in the outcome equations (again, refer to Appendix B for further detail). After this, the physician's predisposition to admit patients to the CDU should not be correlated with the residuals in the outcome equations, satisfying the validity condition.

Our second IV is the busyness of the CDU. Congestion in the CDU, $zCDUCong_i$, is calculated in the same way as ED congestion in Section 5.1, 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

[2] 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 ($ObsAdm_i$, $DischErr_i$ or $TotErr_i$) depending on whether CDU_i takes the value zero or one. Censoring when $CDU_i = 1$ allows us to estimate the effect of ED congestion 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.

Table 3 Descriptive statistics and correlation table for the instrumental variables.

		Mean			Correlation table				
	N	All	CDU = 0	CDU = 1	(1)	(2)	(3)	(4)	(5)
(6) Phys. CDU use	373,663	−0.09	−0.10	−0.00	-0.02^{***}	-0.02^{***}	0.01^{**}	0.14^{***}	-0.05^{***}
(7) CDU congestion	373,663	0.01	0.01	−0.05	0.01^{***}	0.01^{***}	−0.00	-0.02^{***}	0.17^{***}

Notes: Columns 'All', 'CDU = 0' and 'CDU = 1' report mean values for the full sample, subsample where $\text{CDU}_i = 0$ and subsample where $\text{CDU}_i = 1$, respectively; Correlation table column numbers correspond to: (1) Total gatekeeping errors, (2) Short-stay observational admission, (3) Wrongful discharge, (4) CDU admission, (5) ED congestion; Correlation coefficients significant with $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$.

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 when the CDU is busy we expect there to be fewer CDU admissions, satisfying the relevance condition. For patients who are not admitted to the CDU, the CDU congestion level should have no direct effect on their likelihood of being admitted for short-stay observation or being wrongfully discharged. However, to the extent that CDU congestion may be correlated with busyness in the main hospital, we control for this using the congestion level of the hospital (calculated in the same way as CDU congestion).

Hypothesis testing of the IVs to identify whether there are signs of over-, under- or weak identification provide strong evidence that the IVs are valid (p-values > 0.10), relevant (p-values < 0.001), and achieve a maximal relative bias significantly less than 10%, as desired (see Section EC.2 of the e-companion). Our results are also robust to using the IVs individually.

6.3. Results

Before presenting the full set of results, we start by reporting in Table 4 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 busier, they (1) increase the rate at which they refer patients to the CDU (coef. $= 0.069$, p-value < 0.001), (2) make more gatekeeping referral errors in general (coef. $= 0.017$, p-value < 0.001), (3) make more short-stay observational admissions (coef. $=$ 0.027, p-value < 0.001), and (4) make wrongful discharges (coef. $= -0.017$, p-value $= 0.026$). These responses to increasing levels of diagnostic uncertainty are consistent with Hypotheses 1 and 2, i.e. that physicians simultaneously become more error-prone and more cautious, admitting significantly more patients to the hospital for short-stay observation, thus reducing the relative rate of wrongful discharges. In the remainder of this section we investigate our hypotheses using the empirical strategy outlined in Section 6.2.

Given that ED congestion is significant in the selection equation (model (1) of Table 4) we must correct with the heckprob models for potential endogeneity to ensure that the coefficient of

Table 4 Base coefficient estimates using probit model specification.

	(1) CDU	(2) TotErr	(3) ObsAdm	(4) DischErr
ED congestion	0.069^{***}	0.017^{***}	0.027^{***}	-0.017^{*}
	(0.004)	(0.005)	(0.005)	(0.008)
CDU referral	–	-0.003	-0.042^{**}	0.148^{***}
		(0.012)	(0.013)	(0.019)
CDU congestion	-0.062^{***}	–	–	–
	(0.003)			
Phys. CDU rate	1.054^{***}	–	–	–
	(0.017)			
N	373,663	373,663	373,663	373,663
Log-lik	$-99{,}268$	$-67{,}858$	$-55{,}078$	$-20{,}464$
Pseudo-R^2	0.170	0.124	0.162	0.059

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

ED congestion in the outcome equations are not biased. Heckprob model coefficients are reported in Table 5. In heckprob (1e), (2e) and (3e) we identify the effect of ED congestion 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 $CDU_i = 0$. Examining first the estimated correlations coefficients, ρ, the negative and significant correlation coefficients in heckprobs (1e), (2e) and (3e) indicate that those patients who were not referred into the CDU are less likely to be a gatekeeping referral error, short-stay observational admission or wrongful discharge than a patient selected at random from the population. Conversely, the positive and significant correlation coefficients in heckprobs (1c) and (2c) are consistent with the positive selection of patients to the CDU: patients referred to the CDU are more likely to be cases of gatekeeping errors and short-stay observational admissions than a patient selected at random from the population. This is entirely consistent with our expectations, suggesting that ED physicians are able to identify and admit to the CDU patients who will benefit from being there.

After correcting for endogenous selection, we find evidence consistent with that of probits (2), (3) and (4) in Table 4. In particular, evidence from Table 5 suggests that when the ED is more congested, ED physicians are likely to make more gatekeeping errors (coef. $= 0.020$, p-value < 0.001 in heckprob (1e)) and are significantly more likely (coef. $= 0.028$, p-value < 0.001 in heckprob (2e)) to admit patients to the hospital for short-stay observation. At the same time, ED physicians become less likely (coef. $= -0.014$, p-value $= 0.094$ in heckprob (3e)) to discharge patients in error when the ED becomes congested. All of this evidence is consistent with ED physicians "overcorrecting" for the increased risk when congestion induces increased clinical uncertainty, by increasing the rate at which they admit uncertain cases.

Table 5 Coefficient estimates to establish ED physicians' response to increased congestion, using heckprob model specification.

	Decision made by ED physicians			Decision made in the CDU		
	(1e) TotErr	(2e) ObsAdm	(3e) DischErr	(1c) TotErr	(2c) ObsAdm	(3c) DischErr
ED congestion	0.020^{***}	0.028^{***}	$-0.014^{\dagger}$	$0.026^{\dagger}$	0.042^{**}	-0.026
	(0.005)	(0.005)	(0.008)	(0.013)	(0.015)	(0.020)
ρ	-0.220^{***}	-0.114^{*}	-0.179^{**}	0.196^{**}	0.292^{***}	-0.161
	(0.040)	(0.047)	(0.057)	(0.066)	(0.069)	(0.097)
N	373,663	373,663	373,663	373,663	373,663	373,663
N uncensored	337,144	337,144	337,144	36,519	36,519	36,519
Log-lik	−157,985	−147,023	−116,471	−108,056	−106,168	−102,434

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

While not the primary effect of interest, the impact of demand pressure in the ED on the likelihood of a patient who has already been admitted to the CDU being subsequently admitted to the hospital in error is interesting (coef. = 0.042, p-value = 0.004 in heckprob (1c)). This suggests that the ED exerts downstream pressure on the CDU when busy to release additional buffer capacity. In particular, as the ED becomes more congested, more patients are referred to the CDU (coef. = 0.069, p-value < 0.001 in probit (1) from Table 4) and from the CDU, more patients are admitted to the hospital, which frees up capacity to accept additional incoming patients. At the same time, the CDU appears to be effective at preventing avoidable hospitalizations (coef. = −0.042, p-value = 0.002 in probit (2) from Table 4) – an effect we explore in further detail in Section 7.

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). This shows that moving from a low (2σ below the mean) to high ($+2\sigma$ above the mean) congestion state in the ED will: increases in absolute terms the probability of admission to the CDU by 4.00%, $CI_{95} = (3.56\%, 4.45\%)$, total gatekeeping errors by 0.84%, $CI_{95} = (0.44\%, 1.25\%)$, being admitted for short-stay observation by 0.90%, $CI_{95} = (0.54\%, 1.26\%)$, and decreases the probability of being wrongfully discharged by −0.16%, $CI_{95} =$ (-0.34%, 0.03%). Compared with the average rate of CDU use and error rates reported in Table 1, this represents a relative increase (decrease) of approximately 51.5%, 17.9%, 24.2% and −15.0%, respectively. The congestion state of the ED thus has a surprisingly large impact on the decision taken by physicians in the ED. In fact, if we (conservatively) assume a cost of £1000 per avoidable hospitalization, then if all patients had been treated in the ED in a high state of congestion, over-referrals by ED physicians would have cost the hospital approximately £6m extra.

6.4. Robustness to endogeneity concerns

While we have argued above that the increase in avoidable hospitalization and decrease in wrongful 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 changes, e.g. more complex cases may arrive. This then might necessitate an increase in hospital admissions by ED physicians. We note that this is unlikely to be driving our results, since (a) even if there were more admissions, we should not expect an increase in short-stay observational admissions since the more risky patients should be less (rather than more) likely to be discharged promptly without treatment, and (b) if patients were becoming riskier, we would also expect an increase in the rate of wrongful discharge, rather than a decrease.

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

Having established that physicians in the ED become more cautious and over-admit patients to expensive acute inpatient beds when faced with increased congestion and diagnostic uncertainty – at a significant cost to the provider – we next look at what action might be taken to mitigate this effect. We would like to know whether the two-stage gatekeeping process reduces the high rate of gatekeeping errors in referrals of patients from the ED into acute inpatient beds. In this section we describe the method of estimation and present results. Before doing so, we briefly provide some statistics related to the CDU.

Of the 36,519 ED patients that are sent to the CDU, 35.2% are subsequently admitted, and the rest are discharged. 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 16.2 hours for a patient classed as a short-stay observational admission, 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 from the CDU, 14.3% are then identified to be short-stay observational admissions, similar to the 13.6% error rate for those admitted directly from the ED. This is despite the fact that patients admitted from the CDU are subject to considerably more diagnostic uncertainty and thus inherently more likely to be admitted in error. Further analysis (documented in Section EC.1 of the e-companion) indicates that the CDU is at least 42% faster at processing patients routed through it than if they had been admitted to a hospital inpatient unit. Thus, while referral through the CDU does extend the service episode, it does so less than if all patients were instead referred directly to the hospital. This is consistent with findings in the medical literature (e.g. Baugh et al. 2012).

7.1. Empirical specification

The empirical approach that we adopt is similar to that described in Section 6.1. Rather than use a heckprob model we estimate the models 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, to what extent) decoupling the gatekeeping decision and allowing ED physicians to, when uncertain, pass on the referral decision to a second gatekeeping stage can help to reduce total gatekeeping errors and avoidable hospitalizations. 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 the rate of wrongful discharges, estimated by α_2, when patients are routed through the CDU.

For patients who are admitted to the CDU, it is possible that admission and discharge decisions made in the CDU are affected when the CDU becomes busier, which might impact on error rates. To account for this, we include in the outcome equations an additional control variable that takes the value of zero when the patient is not admitted to the CDU and is equal to $zCDUCong_i$ otherwise. (In the heckprob models we did not need this additional control since those patients who are admitted to the CDU are censored in models (1e), (2e), and (3e), i.e. this control would always take the value 0 in the outcome equation.) This ensures that $zCDUCong_i$ is still valid as an instrumental variable. We also control for the duration of time that a patient spends in the ED and in the CDU. This allows us to isolate the direct impact of admission to the CDU, controlling for differences in the time that the patients spent under observation in the ED and/or the CDU.

7.2. Results

Table 6 shows evidence of positive correlation in all of the biprobit models, with estimated correlation coefficients $\rho = 0.327$ (p-value < 0.001), $\rho = 0.286$ (p-value < 0.001), and $\rho = 0.205$ (p-value < 0.001). This indicates that there are unobservables that, on average, make a patient more likely to be admitted to the CDU and also more prone to being a gatekeeping referral error, short-stay observational admission or wrongful discharge. 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 (1) result in a short-stay observational admission (coef. $= -0.563$, p-value < 0.001 in column (2o)) and (ii) be wrongfully discharged (coef. $= -0.216$, p-value $= 0.001$ in column (3o)). Unsurprisingly, these two types of

Table 6 Coefficient estimates for CDU impact.

	(1) TotErr		(2) ObsAdm		(3) DischErr	
	(1s) CDU	(1o) TotErr	(2s) CDU	(2o) ObsAdm	(3s) CDU	(3o) DischErr
CDU referral	–	-0.597^{***}	–	-0.563^{***}	–	-0.216^{**}
		(0.045)		(0.045)		(0.074)
CDU congestion	-0.076^{***}	–	-0.076^{***}	–	-0.076^{***}	–
	(0.003)		(0.003)		(0.003)	
Phys. CDU rate	0.998^{***}	–	0.998^{***}	–	0.999^{***}	–
	(0.017)		(0.017)		(0.017)	
ρ	0.327^{***}		0.286^{***}		0.205^{***}	
	(0.027)		(0.026)		(0.042)	
N	373,663		373,663		373,663	
Log-lik	−160,611		−147,620		−113,905	

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.
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.10$.

error (i.e. total gatekeeping errors) also jointly decrease (coef. $= -0.597$, p-value $= 0.001$ in column (1o)). 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 the risk of non-referral errors.

To see how much better admission decisions are when made in the CDU rather than by an ED physician, we convert coefficient estimates to average treatment effects (ATEs) and average treatment effects on the treated (ATTs). This shows that if no patients were referred through the CDU the rate of short-stay observational admissions and wrongful discharges would have been 5.23% and 1.16%, respectively. These change to 1.76% and 0.67%, respectively, if all patients are instead routed through the CDU. Thus the CDU acts to significantly reduce short-stay admissions (ATE $= -3.47\%$) and wrongful discharges (ATE $= -0.49\%$). The ATTs are even larger than the ATEs, taking values -9.3% and -1.2% for short-stay observational admission and wrongful discharge, respectively. This suggests that ED physicians are especially good at routing into the CDU patients who they would have otherwise been a gatekeeping error.

8. Counterfactual Analysis

Having established the usefulness of the CDU, the question of how large the CDU should be arises. After all, the resources in the CDU could be redeployed in the ED, increasing the ED's capacity and thereby reducing congestion. This could improve decision-making in the ED and lower the rates of avoidable hospitalization. Thus, even though we find that the two-stage gatekeeping system reduces gatekeeping errors, the effect of the second gatekeeping stage in the combined system is not obvious. To examine the combined system, we perform a counterfactual analysis.

Over the six year sample period, the total number of hours spent by patients in the ED and CDU was 1.5m and 326k hours, respectively. Hence, if the ED and CDU were merged, we assume that capacity in the ED could be increased by approximately 21.7%.[3] To simulate this capacity expansion in the data we multiply our measure of ED capacity from Section 5.1, $CapacityED_h$, by 1.217.

Any increase in ED capacity would translate into a reduction in ED congestion, as the resources (e.g., physicians, nurses, treatment rooms) consumed by the CDU would have been available for use by the ED instead. Re-calculating ED congestion using the equation from Section 5.1 gives $EDCong_i^* = QueueED_i/(1.217 \times CapacityED_{h_i})$. To ensure that the original and updated measures of ED congestion are on the same scale, we then standardize using the original mean, $\mu(EDCong_i)$, and standard deviation, $\sigma(EDCong_i)$. This shows that if the CDU had not existed and if the resources consumed by the CDU could have been redeployed in the ED then this would have had the effect of reducing congestion levels in the ED by approximately 0.64σ. Substituting the original values of $zEDCong_i$ for the updated values achieved through pooling ED and CDU capacity, $zEDCong_i^*$, into heckprob (1e), we estimate that the increase in ED capacity would have reduced the rate of short-stay observational admissions by 0.14 percentage points.

Recall that in Section 7.2 we estimated that the rate of short-stay observational admissions would rise from their observed level of 4.25% with the CDU to 5.23% if the CDU were closed. However, this ignored the possibility of redistributing resources from the CDU, if it were closed, to increase capacity in the ED. Accounting for this, we estimate instead that if the CDU were closed and resources could be redistributed to the ED, then the short-stay observational admission rate would equal 5.09% ($= 5.23\% - 0.14\%$) – still a substantial deterioration relative to the status quo of 4.25%. In summary, we estimate the net effect of the CDU in the study hospital, after accounting for the opportunity cost of its resources, to be a relative reduction of the short-stay observational admissions rate by 16.5% (i.e. from 5.09% to 4.25%).

While this analysis demonstrates the advantage of the CDU over a pooled ED resource in the study hospital, the question of how resources should be distributed between the ED and CDU to optimize patient flow and minimize referral errors remains. This requires analytical work that goes beyond the scope of this empirical paper and is left for future research.

[3] Note that we take a 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.

9. Conclusions

Our empirical study of gatekeeping under congestion provides two main insights. First, our data shows that congestion-sensitive gatekeepers, such as the ED physicians in our study, change the trade-off point on the speed-quality curve in the interest of speed when congestion increases, leading to more referral errors. However, rates of "missed" and "unnecessary" referrals do not necessarily increase proportionally. Specifically, when gatekeepers regard a missed referral as a more severe error than an unnecessary referral, as is the case in our empirical context, they protect the rate of the more severe error by lowering the threshold for specialist referrals when congestion increases. This increases the rate of avoidable referrals disproportionately and therefore causes excess false demand for the downstream specialist precisely at a time when the gatekeeper should ration access to specialists more stringently to protect the specialist unit from the upstream demand surge. The result is a demand amplification effect, where an upstream demand surge in the gatekeeping system causes a relatively larger downstream demand surge in the specialist unit. This has repercussions for the safety of the hospital as a whole (Kuntz et al. 2015, Eriksson et al. 2017)

To alleviate this problem, managers would benefit from systems designed to adjust the speed-quality trade-off without a significant increase in cost. Our second insight is that a two-stage gatekeeping system can provide such a Pareto improvement when there is considerable heterogeneity in the difficulty of the gatekeeping decision. This would allow system designers to better match the heterogeneous customer characteristics that become apparent during the first gatekeeping stage with more appropriate resources and gatekeeper characteristics at the second stage, leading to overall improved decision-making. Our study provides empirical evidence that a two-stage gatekeeping system can significantly reduce rates of both missed and unnecessary referrals.

While our study focuses on emergency care, the benefits of multi-stage gatekeeping are likely to extend to other industries and health contexts. For example, accurate detection and diagnosis of rare diseases in primary care takes, on average, seven years in the US and five years in the UK (Shire 2013). Such cases are costly as patients visit their primary care physician (PCP) multiple times, undergo multiple tests and see multiple specialists. Our results suggest that a potential solution may be to designate a subset of more experienced PCPs (with a track-record of identifying more complex diseases) as second-stage gatekeepers, allowing PCPs to refer patients to them. More generally, our findings demonstrate that such a two-stage gatekeeping system could reduce overuse of inappropriate specialist services while improving the accuracy of referral, a win-win for both the system and the customer.

References

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

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

Batt R, Terwiesch C (2015) Waiting patiently: An empirical study of queue abandonment in an emergency department. *Management Science* 61(1):39–59.

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.

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.

Boyle A, Beniuk K, Higginson I, Atkinson P (2012) Emergency department crowding: Time for interventions and policy evaluations. *Emergency Medicine International* .

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.

Eriksson C, Stoner R, Eden K, Newgard C, Guise J (2017) The association between hospital capacity strain and inpatient outcomes in highly developed countries: A systematic review. *Journal of General Internal Medicine* 32(6):686–696.

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`.

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

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

Gorski J, Batt R, Otles E, Shah M, Hamedani A, Patterson B (2017) The impact of emergency department census on the decision to admit. *Academic Emergency Medicine* 24(1):13–21.

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.

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.

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 (2014) Does multitasking improve performance? evidence from the emergency department. *Manufacturing & Service Operations Management* 16(2):168–183.

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 (2015) 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).

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

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.

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.

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, Iravani SMR, Cheng Y, Diermeier D (2017) Workload management in telemedical physician triage and other knowledge-based service systems. *Management Science (forthcoming)* .

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 (2014) 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.

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.

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.

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.

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.

Appendix A: Data Preparation

In this first section of the appendix we describe the method used 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 accurate matching we require that the ED departure and inpatient admission timestamps are within two hours of each other. This left us with 7,771 unmatched records corresponding to 7,496 patients. For the 7,249 patients for whom we are unable to match 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. We also drop 311 ED attendances where the patient was supposed to have been admitted to the CDU but no corresponding record exists, 351 attendances where the patient was not meant to be admitted to the CDU but a record exists of them being in the CDU, and 17 obs. with timestamps indicating ED discharge prior to arrival. This left 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 Appendix B). 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 period from one day before until one day after each public holiday. Given a temporary change in coding convention in December 2009 and January 2010 that made 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 were left with 479,678 obs.

The other data restrictions are as described in Section 5.

Appendix B: Physician-level Controls

In Section 6.2 we introduce the history of CDU use by the physician assigned to a patient as an instrumental variable. Here we elaborate on how this IV is calculated.

We would like to identify the propensity of a physician to admit patients to the CDU after controlling for observable differences in patient characteristics and assignment. To do this, we first estimate a probit model that takes the form

$$CDU_i^* = \delta_0 + \mathbf{T}_i\boldsymbol{\delta}_1 + \mathbf{D}_i\boldsymbol{\delta}_2 + \mathbf{C}_i\boldsymbol{\delta}_3 + \epsilon_i^\delta\,, \tag{5}$$

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

where $\mathbf{T}_i$, $\mathbf{D}_i$ and $\mathbf{C}_i$ specify the temporal, patient/diagnosis related and contextual controls outlined in Table 2, and where $\epsilon_i^\delta \sim \mathcal{N}(0,1)$, CDU_i^* is a latent variable and CDU_i is the observed dichotomous variable that indicates whether the patient was sent to the CDU.

This model gives the baseline risk of a patient being admitted to the CDU if treated by an 'average' physician. We then take the fitted values from the auxiliary equation, $\widehat{CDU_i^*}$, and estimate a random effects probit model of the form

$$CDU_{i_{pm}}^* = \delta_{pm} + \widehat{CDU_{i_{pm}}^*} + \epsilon_{i_{pm}}^\delta\,, \tag{7}$$

$$CDU_{i_{pm}} = \mathbb{1}[CDU^*_{i_{pm}} > 0]\,, \tag{8}$$

where $\epsilon^{\delta}_{i_{pm}}$, $CDU^*_{i_{pm}}$ and $CDU_{i_{pm}}$ are as defined before but for the subset of observations i assigned to physician p in the 12 month period $[m-12, m-1]$, indexed i_{pm}. The random intercept δ_{pm} then captures variation in CDU admission rates across physicians and within physicians over time. The value of the IV for a patient who arrives in month m and is assigned to physician p is then set equal δ_{pm}.

The controls in $\mathbf{P}_i$ of Table 2, which capture a physician's historic short-stay observational admission, wrongful discharge and total gatekeeping error rates, are calculated in the same way as for their CDU admission propensity.

e-companion to
"Gatekeeping Under Congestion: 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 avoidable 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 wrongful discharges, for which our analysis shows there is also an 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 avoidable hospitalizations or (2) are avoidable hospitalizations, and in addition those instead admitted to the CDU who are (3) discharged or are (4) admitted and subsequently not deemed to be an avoidable hospitalization or are (5) admitted and then classed as an avoidable hospitalization. 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 directly 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 avoidable hospitalizations (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 2. 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 approximately 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 approximately 42.0% $(= 1 - \frac{251,581}{434,096})$.

Appendix EC.2: 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.2.1. 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

short-stay observational admission, wrongful discharge 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.

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 avoidable hospitalization 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 wrongful 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 avoidable hospitalization model, and equal to 442.6 for the wrongful 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.2.2. 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 avoidable hospitalization biprobit model $\chi^2 = 1.18$, p-value$= 0.555 > 0.10$, for the wrongful discharge model $\chi^2 = 0.96$, p-value$= 0.618 > 0.10$. 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).