Gatekeeping Under Time Pressure:
An Empirical Study of Hospital Admission Decisions in the Emergency Department

(Authors’ names blinded for peer review)

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

Key words: gatekeeper systems; routing; server behavior; uncertainty; health care: hospitals; service operations; econometrics

1. Introduction

Many service settings (e.g., health care, call centers, maintenance) are characterized by the presence of multiple service tiers, with customers commencing service at a low-cost entry level (e.g., an emergency department (ED), a general enquiries help-desk, a local repair shop) from where they can be referred to a more specialized and hence more costly level of service (e.g., acute hospital bed, complaints desk, engineering department) if necessitated by the complexity of their needs. The upstream server (e.g., ED physician, telephonist, technician) in such a setting assumes a dual role: They will service simple requests themselves while, at the same time, acting as a gatekeeper to downstream specialist units, thereby ensuring that customers receive the appropriate service intensity for their needs (Shumsky and Pinker 2003).

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

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

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

Our empirical study is based on over 650,000 patient attendances to the busy ED of a UK-based teaching hospital over a seven year period. ED physicians act as gatekeepers to expensive acute inpatient beds, responsible for restricting access to the main hospital to only those patients whose immediate treatment needs are too complex to be met by staff working within the ED itself. Despite the fact that one in 10 medical diagnoses are estimated to be wrong (Graber 2013), with errors in the diagnostic process the leading cause of internal investigation and malpractice claims in the ED (Cosby et al. 2008), ED physicians often experience high caseloads and must make decisions under considerable time pressure (see Section 3). As a consequence, unnecessary admission and inappropriate discharge decisions can occasionally occur: For the ED in our study, the error rate among admitted patients is estimated at 16.1% versus an error rate among discharged patients of 1.3%\(^1\). The high rate of admission errors relative to discharge errors suggests that when faced with an uncertain decision ED physicians err on the side of caution and adopt a “safety-first” principle, preferring to minimize the risk that their patient leaves untreated over the risk of an incorrect admission that is costly for the hospital and may impede the service received by the other hospital patients, but is safe for the patient at hand (Roy 1952). To confirm this behavior, we study empirically how an increase in system congestion – which reduces the time available for diagnosis and so increases uncertainty – affects the rate of avoidable admissions to acute hospital beds and inappropriate discharges from the ED. We find that for every one standard deviation increase in ED busyness, ED physicians *increase* the rate at which they admit patients to the hospital by 7.7%. At the same time, there is also a *reduction* in the rate of errors in discharge by 3.3%. Thus, when faced with additional diagnostic uncertainty ED physicians adjust the rate of admissions in order

\(^1\) An admission error is defined to be any patient admitted to the hospital and subsequently discharged within 24 hours with no treatment provided, while a discharge error is defined to be a patient treated in the ED and sent home who returns to the ED within seven days and is at that point admitted. See Section 4.2 for more detail.
to avoid a higher chance of an unlikely but potentially ‘catastrophic’ error in discharges. Moreover, they adjust beyond the rate that would be necessary to preserve prevailing rates of discharge error, which is suggestive of risk aversion. Importantly, this is precisely the opposite behavior to that which is desirable from a system perspective: Demand surges in the ED lead to more admission errors and therefore an amplification of the surge in the hospital (a bullwhip-type effect), with potentially negative implications for the other patients under their care.

Having established the undesirable behavioral effect of congestion on admission errors, the second part of this paper studies what can be done to mitigate the adverse impact of diagnostic uncertainty on performance (as measured by overuse of specialist resources) in gatekeeping systems such as this. We offer one potential solution: to allow the gatekeeper to classify a referral candidate as “unresolved” prior to making the referral decision, and offloading them instead to an intermediate second-stage gatekeeper who assumes responsibility for deciding whether a referral is necessary. Since these unresolved cases are more homogenous than the average patient arrival (with unambiguous referrals and non-referrals having already been filtered out by the first-stage gatekeeper), they can be looked after by a more specialized workforce and additional resources can be invested to acquire information to increase diagnostic accuracy and reduce referral errors. As it happens, this two-stage gatekeeping process already exists in the context of our study hospital by way of the presence of a clinical decisions unit (CDU). The CDU is a stand-alone unit attached to the ED into which a patient can be referred for further monitoring, diagnostic evaluation, and/or treatment. Beds in the CDU are of lower intensity and cost than acute beds in the main hospital, but patients are able to stay up to 24 hours (rather than 4 hours in the ED) and it is generally staffed by more experienced clinicians than the ED. The CDU thus provides ED physicians with an alternative to discharge or hospital inpatient admission that can be leveraged when it is unclear whether or not the patient should be admitted. A comparison of this two-stage process to the traditional gatekeeping set-up is shown in Figure 1.

After accounting for non-random assignment of patients to the CDU using appropriate sample selection methods, we show that patients routed through the CDU are 12.2% less likely to be
admitted in error than patients admitted directly by ED physicians, while being no more likely to be discharged in error. Moreover, we find that patients admitted directly from the ED are 8.2% more likely to have a specialty transfer during their hospital stay than patients admitted via the CDU, indicating fewer routing errors within the hospital for CDU patients.

Our study provides empirical support that intermediate “semi-specialist” gatekeeping units can help alleviate the trade-off between speed and quality in multi-tier service systems (see e.g. Anand et al. 2011, Alizamir et al. 2013): While it might appear desirable to incentivise gatekeepers to make referral decisions faster when the system is congested, to increase throughput and ensure that customers receive prompt service, this can reduce the time available for accurate diagnosis and lead to increased referral errors which can erode the benefits of higher throughput and lead to worse outcomes and higher system costs. An example in point is the introduction of a waiting time target in the NHS in 2004, requiring that 98% (adjusted later to 95%) of patients be admitted or discharged within four hours of arrival to the ED. This target led to faster decision-making in the ED and reduced waiting times. However, it also coincided with a 30% increase in hospital admission rates, at a multi-billion pound cost to the UK healthcare system (NAO 2013). Our findings suggest that a more systematic use of a two-tier gatekeeping system, emphasizing the gatekeeping role of CDUs, might have moderated the unintended negative effects of the waiting time target.

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

2. Literature Review

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

Most relevant to our study is extant literature on gatekeeping systems. Such systems are comprised of two service tiers, with the server in the first tier referred to as a ‘gatekeeper’ because of the dual nature of their role, either ‘self-treating’ the customer or else, if too complex, referring them to a more costly but higher-skilled second tier ‘specialist’ (Shumsky and Pinker 2003). This service system has been studied mainly in the health economics literature – due to parallels with
systems of referrals between primary and secondary/tertiary care – with a focus on the conditions under which gatekeeping systems are preferable to direct access and the design of contracts to reduce information frictions (Mariñoso and Jelovac 2003, Brekke et al. 2007, González 2010). In the operations management literature, early modeling work has looked at how the system optimal rate of referrals between gatekeeper and specialist can be incentivized in both deterministic (Shumsky and Pinker 2003) and stochastic (Hasija et al. 2005) settings. This modeling framework has been extended to investigate e.g. outsourcing contract decisions (Lee et al. 2012) and the performance of security-check queues (Zhang et al. 2011).

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

Another stream of research focuses specifically on the classification problem. Both van der Zee and Theil (1961) and Argon and Ziya (2009) examine customer classification policies when there exists imperfect information about customer type (e.g. refer versus self-treat). While the classification threshold affects error rates in these papers, misclassification is not inherently affected by service times or effort. Alizamir et al. (2013), on the other hand, also examines the process of customer type identification, but with a server who can perform additional diagnostic testing to resolve type uncertainty. The more tests they perform, the better the accuracy of diagnosis at a
cost of increasing levels of congestion and waiting times for other customers. Similarly, Wang et al. (2010) study diagnostic centers in which servers trade-off the dual concern of accuracy and congestion given that misclassification costs are incurred by both the service provider and customer. They find that increases in capacity may increase congestion, extending the result from the centralized system in Hopp et al. (2007) to the decentralized system. We also expect classification thresholds and errors to depend on congestion levels in our ED setting and study this behavior empirically. We differ, however, in that we (i) are interested in the behavior of the server in response to varying levels of diagnostic uncertainty, rather than the system optimal response, and (ii) also study a mechanism that can be implemented to reduce rates of errors when faced with type uncertainty.

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

Finally, our work relates to recent empirical studies of health care and other service settings which have looked into the impact of organizational factors such as workload on service outcomes (see Freeman et al. (2016) for a recent overview), for example clinical safety (Kuntz et al. 2014), service times (KC and Terwiesch 2009), reimbursement (Powell et al. 2012) and sales performance (Tan and Netessine 2014). Also related is work on patient routing, with Kim et al. (2014) and
KC and Terwiesch (2012) showing, respectively, that high occupancy levels in the intensive-care unit (ICU) can reduce rates of ICU admission and increase early discharge propensity. While these behaviors preserve/free up capacity in the resource-constrained and expensive ICU (i.e. the ‘specialist’ resource) for higher priority patients, we find that when the ED (the first gatekeeping tier) is crowded this pattern may be reversed, with instead more patients being referred into acute inpatient beds (the second-tier ‘specialist’ resource in our context). Freeman et al. (2016) find a similar result in a maternity context. In the first empirical analysis of the two-tier gatekeeping system, they demonstrate that midwives (gatekeepers) refer high complexity patients to obstetricians (specialists) at higher rates in the presence of congestion. In contrast, we focus instead on the effect of diagnostic uncertainty on both referral (admission) and self-treatment (discharge) errors, rather than the one-sided case, as well as exploring a possible preventative measure.

3. Decision Making and Uncertainty in the Emergency Department

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

The ED is a highly time-pressured environment, with congestion and delays in care associated with e.g. higher complication rates and increased mortality (Bernstein et al. 2009, Huang et al. 2010, Sun et al. 2013). Despite this, there is an upwards global trend in ED attendances and ED crowding continues to worsen (Pines et al. 2011). In the US, for example, ED visits between 1997 and 2007 grew at almost twice the rate of population growth (Tang et al. 2010), while in England between
1997 and 2012 ED admissions grew by 47% compared to population growth of 10% over this period (NAO 2013). In fact, the ED is now the primary point of entry to the hospital, admitting more than half of non-obstetric cases (Greenwald et al. 2016). Consequently, mitigating ED crowding is a significant policy concern and countries have adopted a wide range of interventions designed to manage this problem. Examples include telephone advice centers, implementation of fast tracks, increases in capacity and staffing, changes in boarding practices, and, most relevant to our study, the use of observation units and clinical decisions units (for an overview of the various approaches adopted in different countries see Pines et al. 2011). Yet these approaches have met with only limited success; February 2016 statistics from England, for example, revealed that only 87.8% of patients were admitted, transferred or discharged within four hours of their arrival at the ED – significantly lower than the target of 95% and the lowest rate since records began (NHE 2016).

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

One characteristic of the ED context that might drive differences in outcomes as service times are compressed is the existence of high levels of clinical uncertainty and variation in diagnostic accuracy in emergency medicine (Sklar et al. 1991, Green et al. 2008). When service times are reduced, physicians have less time available to spend with each patient to perform diagnostic testing and to acquire the information necessary to make accurate and informed gatekeeping decisions (Smith et al. 2008, Alizamir et al. 2013). Decision density is also high, with ED physicians often caring simultaneously for multiple patients, which can lead to elevated cognitive loading. As a consequence they must regularly rely on heuristics and intuition, such as pattern recognition and
rule-based decision-making, when assessing patients’ needs (see Croskerry (2002) for an excellent overview of the types of heuristics employed by ED physicians). Significant time pressure and resource constraints prevalent in many EDs means that these cognitive shortcuts can result in higher than desired levels of costly but preventable errors (Leape 1994). For example, in a study of 100 cases of diagnostic error, Graber et al. (2005) found that cognitive factors contributed in 74% of cases. Two particularly costly errors in the gatekeeping context are referrals errors (type II errors) and self-treatment errors (type I errors). Inappropriate referrals use expensive inpatient beds that might otherwise be used for other types of activity (e.g. planned procedures); inappropriate discharges can lead to patients returning in a worse health state requiring more costly treatment and, potentially, the payment of compensation. As diagnostic uncertainty increases and servers have less information available to make gatekeeping decisions, therefore, we hypothesize higher rates of both of these types of error.

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

As an alternative to Hypothesis 1, as diagnostic uncertainty increases the referral threshold may be adjusted to prevent an increase in one type of error at the expense of a higher rate of the other. Previous modeling literature suggests that such an effect may occur if the cost incurred by the customer and/or the service provider from the ‘protected’ error is significantly greater than that of the ‘sacrificed’ error (Alizamir et al. 2013, Wang et al. 2010, Zhang et al. 2011). In gatekeeping settings, in addition to the provider and customer, the server may also associate costs with each of these types of error. Since the ultimate referral decision is made by the gatekeeper and not by
the provider or the customer, the gatekeeper’s chosen referral rate and the first-best referral rate for the service provider/customer may differ greatly. While various contracts have been proposed to align gatekeeper incentives with those of the provider (see Section 2), these neither adjust for the gatekeepers perceived or realized misclassification costs nor are such contracts typically used in practice. For example, ED physicians in the study hospital are salaried employees and their wages are not affected by the decisions that they take. There may, though, be other factors that these physicians consider when deciding whether to admit or discharge a patient. For example, medical errors have been shown to have a negative emotional impact on physicians (Christensen et al. 1992), can result in malpractice investigations and/or litigation (Studdert et al. 2006), and can also lead to reputation damage and peer disapproval (Leape 1994). The costs (whether financial or otherwise) that physicians associate with these concerns will affect how they respond to uncertainty.

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

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

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

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

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

4. Data Description and Variable Definitions

The data for our study is comprised of detailed information relating to 651,044 ED attendances over a period spanning seven years from December 2006 through December 2013, as well as matching inpatient records for all of those patients admitted from the ED into the hospital during this period. (All 8,527 observations (obs.) from the final month, December 2013, are dropped since data entry may not have been completed fully.) The ED we study is the largest in the region and has experienced increasing demand pressure over recent years, with attendances up by 4.2% year-on-year from 215 ED visits per day on average in the first year of our sample to 274 per day in the
final year. On average 29.1% of patients who arrive at the hospital are admitted to an inpatient bed, with admissions and discharges increasing at approx. the same rate over the sample period (by 4.7% per annum (p.a.) for admissions versus 4.1% for discharges).

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

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

4.1. Referral to the CDU

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

4.2. Admission and Discharge Errors

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

An admission error (or ‘false admission’) occurs when a patient is admitted to an acute hospital bed despite that admission being unnecessary or excessive to their needs. These patients block beds and use expensive specialist resources and time, with unnecessary hospital admissions estimated to
Table 1  Descriptive statistics and correlation table.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean (1)</th>
<th>Mean (2)</th>
<th>Mean (3)</th>
<th>Mean (4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Admission error</td>
<td>429,313</td>
<td>4.71</td>
<td>4.68</td>
<td>5.02</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>CDU = 0</td>
<td>CDU = 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2) Discharge error</td>
<td>429,313</td>
<td>0.95</td>
<td>0.87</td>
<td>1.83</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>CDU = 0</td>
<td>CDU = 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(3) Specialty change</td>
<td>125,228</td>
<td>22.01</td>
<td>22.54</td>
<td>17.12</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>CDU = 0</td>
<td>CDU = 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(4) CDU admission</td>
<td>420,313</td>
<td>8.17</td>
<td>0.00</td>
<td>100.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>CDU = 0</td>
<td>CDU = 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(5) ED busyness</td>
<td>429,313</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td></td>
<td>All</td>
<td>CDU = 0</td>
<td>CDU = 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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

A discharge error (or ‘false discharge’), on the other hand, occurs when a patient who should have been admitted to the hospital is instead discharged from the ED. These patients often come back in a more serious state, requiring a higher intensity of care than would otherwise have been needed if correctly admitted. Pope et al. (2000), for example, found risk-adjusted mortality for patients with acute myocardial infarction who were inappropriately discharged from the ED to be 1.9 times higher than for hospitalized patients. A patient is recorded as a discharge error if after discharge from the ED or CDU they re-attend the ED within 7 days and are at that point admitted to an inpatient bed in the hospital. The rate of discharge errors in the full sample is 0.9% and is 1.3% for the subset of 304,085 discharged patients. Note that the high rate of admission errors relative to discharge errors is already suggestive of physicians overweighing the low probability of a discharge error and taking the cautious approach of admitting patients when faced with uncertainty.

While not all patients we class as admission errors and discharge errors may be true errors (e.g., a patient may require admission for observation according to medical guidelines, or a patient may

\[2\] Authors’ calculations based on total hospital spend in 2012-13 of £12.5 billion on ED admissions (NAO 2013), with 49% of ED admissions staying less than 48 hours (NAO 2013), and estimated 10% of ED admissions for short-term care being avoidable (Denman-Johnson et al. 1997).
be discharged and re-visit the hospital for a problem unassociated with their initial visit), our study investigates how misclassification rates change under different organizational conditions.

4.3. Routing errors

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

4.4. ED Congestion

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

To generate this measure, we first determine which patients’ ED visits overlapped with the period from arrival to one hour post-arrival of patient \(i\), and calculate the sum of those overlapping periods \(Queue_{ED,i}\). It is well known that busyness levels in EDs vary across the day, on weekdays and weekends, in different seasons, and change over time. Since some of this is predictable and staffing can be partially set to meet demand, we will adjust \(Queue_{ED,i}\) also to account for these differences. We achieve this by employing a variation on the approach used in Kuntz et al. (2014) and Berry Jaeker and Tucker (2016) which establishes an approx. upper bound on the available capacity. We estimate this upper bound using quantile regression to predict the 95th percentile level of occupancy at hour \(h\). The dependent variable in this regression is the time-weighted average occupancy level over every hour \(h\) starting midnight on 1st January 2007 and ending midnight on 31st December 2013. (Note that all dates dropped during the data cleaning process, as described in Appendix A, are also removed here.) We estimate this model with independent variables: (i) year, (ii) quarter of the year, (iii) time, split into six four-hour windows per day (e.g., midnight to 4a.m., etc.), (iv) a binary variable equal to one if a weekend and zero otherwise, (v) the interaction between (iii) and (iv), and (vi) the interaction between (v) and a binary variable equal to one if the date was between the years 2011 to 2013, and zero otherwise. The fitted values from this model then provide us with our estimate of capacity for each hour \(h\), \(Queue_{ED,i}^{95th}\). ED congestion, \(Occ_{ED,i}\), is then equal to \(Queue_{ED,i}\) divided by \(Queue_{ED,i}^{95th}\), where \(h_i\) is the hour of arrival of observation \(i\). Finally, we normalize this by subtracting it’s mean, \(\mu(Occ_{ED,i})\), and dividing through by it’s standard deviation, \(\sigma(Occ_{ED,i})\), to form \(zOcc_{ED,i}\). Plots of \(zOcc_{ED,i}\) are provided in Figure 3.
4.5. Control Variables

In addition to the primary variables described above, we also have available and derive a large number of control variables that allow us to account for heterogeneity in the patient population and in the hospital that may be correlated with the dependent variables, and/or with the main independent variables of interest. These are reported in Table 6 in Appendix B, and capture temporal factors, differences in diagnosis and condition, contextual factors (e.g., arrival method), and attributes of the assigned physician. Any factors not reported in our data that might be correlated with the primary independent variables (and so through omission may bias the results) will be accounted for using appropriate empirical methods to be described in Section 5.1.

5. Models and Results I: Response to Diagnostic Uncertainty

We are interested in how ED physicians respond when faced with additional uncertainty. As mentioned in Section 4.1, this requires us to initially study only those referral decisions made by ED physicians directly, i.e. not those cases referred into the CDU. Identification is complicated, however, by the fact that the patients who are passed to the CDU in our sample may be inherently different from those for whom the physician makes the referral decision themselves. While we account as far as possible for these differences with our set of controls (reported in Table 6), there may still exist factors unobservable to us, the researchers, but observable to the physician (e.g., fitness level, medical history) that influence whether or not the physician leverages the CDU option. Not accounting for this endogeneity could lead to biased coefficient estimates and invalidate our findings. In this section, we describe the empirical approach we adopt to resolve this.

5.1. Econometric Specification

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

\[ CDU^*_i = \delta_0 + X_i \delta_1 + Z_i \delta_2 + zOccED_i \delta_3 + \epsilon_i^s, \]  
\[ CDU_i = 1[CDU^*_i > 0], \]  

where \( \epsilon_i^s \sim \mathcal{N}(0, 1) \), \( CDU^*_i \) is a latent variable, the vector \( X_i \) contain the set of all controls (reported in Table 6), the vector \( Z_i \) contains the set of instrumental variables (to be described in Section 5.2), \( CDU_i \) is the observed dichotomous variable that indicates whether the patient was sent to the CDU, and \( 1[\cdot] \) is the indicator function. The second stage (outcome) equation takes the form

\[ AdmErr^*_i = \beta_0 + X_i \beta_1 + CDU_i \beta_2 + zOccED_i \beta_3 + \epsilon_i^\beta, \]  
\[ AdmErr_i = 1[AdmErr^*_i > 0], \]  

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

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

Rather than estimate the first and second stage models described above individually, instead, we estimate them simultaneously with a Heckman probit sample selection (heckprob) model using full information maximum likelihood (Maddala 1983). The heckprob model allows us to estimate the effect that ED congestion has on our outcomes for only those patients who were admitted or discharged directly by an ED physician (rather than by a physician in the CDU) – which is the effect we are interested in – while also allowing us to account for the fact that the rate of referrals into the CDU may also differ as the ED becomes congested. To achieve this we censor the outcome
variable $AdmErr_i$, $DischErr_i$ or $SpecChg_i$, whenever $CDU_i = 1$, set $\alpha_2, \beta_2, \gamma_2 = 0$ in the outcome equation, remove ED length of stay (see Table 6) from the control vectors $X_i$ and $W_i$ (since ED busyness is very likely to affect length of stay in the ED), and then estimate the selection and outcome equations simultaneously under the assumption that their errors $(\epsilon_i^s, \epsilon_i^\gamma)$, $(\epsilon_i^s, \epsilon_i^\rho)$ or $(\epsilon_i^s, \epsilon_i^\rho)$ are jointly distributed according to the standard bivariate normal distribution with unit variances and correlation coefficients $\rho^s$, $\rho^\gamma$ or $\rho^\rho$ which are estimated as parameters in the models.\(^3\) We claim that ED physicians adjust the rate at which they admit patients to the hospital (rather than simply making more mistakes in general) when faced with higher levels of diagnostic uncertainty due to shortening service times if as the system becomes more congested there is an increase in the rate of false admissions (i.e., $\beta_3 > 0$) without a similar increase in the rate of false discharges (i.e., $\alpha_3 \leq 0$) or referrals to the wrong specialty (i.e., $\gamma_3 \leq 0$).

5.2. Instrumental Variables

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

The first IV is the CDU admission propensity of the assigned physician. This is equal to the physician’s average rate of CDU referrals over the previous twelve months relative to the rate expected given the case-mix of patients they treated. A patient assigned to a physician who is more predisposed to admit patients to the CDU will be more likely to be sent there themselves, satisfying the relevance condition. Furthermore, since we already control for the physician’s admission, discharge and, where relevant, transfer propensity in the selection and outcome equations (see Table 6), the physician’s predisposition to admit patients to the CDU should not affect the error rates other than through the CDU admission decision itself, satisfying the validity condition.

\(^3\)Traditionally, Heckman sample selection models are used when the outcome is not observed in the case of non-selection (for example, if we had no further information about those patients admitted to the CDU). In our case, however, we observe the outcome both when the ED physician makes the referral decision and when it is made in the CDU. It is possible, therefore, for us to estimate the coefficients under both regimes (i.e., when the referral decision is made by either the ED or a CDU physician). This estimation can be made jointly using an endogenous switching regression model, or instead by estimating both sides of the equation separately by “tricking” the Heckman selection model to do so, as described in Lee (1978). We employ this trick by censoring the dependent variable in the outcome equation ($AdmErr_i$, $DischErr_i$) depending on whether $CDU_i$ takes the value zero or one. Censoring when $CDU_i = 1$ allows us to estimate the effect of ED busyness on error rates made by ED physicians, while censoring when $CDU_i = 0$ allows us to estimate the effect on decisions made in the CDU instead. Joint estimation (not reported) results in nearly identical estimates of the coefficients and $\rho$. 
Table 2 Descriptive statistics and correlation table for the instrumental variables.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>All</th>
<th>CDU = 0</th>
<th>CDU = 1</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6) CDU busyness</td>
<td>429,313</td>
<td>12.84</td>
<td>12.97</td>
<td>11.31</td>
<td>0.07***</td>
<td>0.01***</td>
<td>-0.04***</td>
<td>-0.01***</td>
<td>0.03***</td>
</tr>
<tr>
<td>(7) Phys. CDU use</td>
<td>429,313</td>
<td>0.00</td>
<td>0.01</td>
<td>-0.06</td>
<td>0.01***</td>
<td>-0.00</td>
<td>0.01***</td>
<td>-0.02***</td>
<td>0.17***</td>
</tr>
</tbody>
</table>

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

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

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

5.3 Results

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

Given that ED busyness is significant in the selection equation (model (1) of Table 3) we must correct with the heckprob models for potential endogeneity to ensure that the coefficient of ED busyness in the outcome equations are not biased by this. Heckprob model coefficients are reported in Table 4. In heckprob (1e), (2e) and (3e) we identify the effect of ED busyness for only the subset of patients for whom the referral decision is made directly by an ED physician, i.e. censoring when \( CDU_i = 1 \). For completeness, in heckprob (1c), (2c) and (3c) we report this instead for only those patients admitted to the CDU, i.e. censoring when \( CDU_i = 0 \). Heckprob (1e) shows evidence of significant negative selection (\( \rho = -0.420, \) p-value < 0.001), meaning that patients who are not selected for admission to the CDU are less likely to be a false admission than a patient selected at random from the population, verifying the need to account for endogeneity.

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

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Base coefficient estimates using probit model specification.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) CDU</td>
</tr>
<tr>
<td>ED busyness</td>
<td>0.067*** (0.004)</td>
</tr>
<tr>
<td>CDU referral</td>
<td>−</td>
</tr>
<tr>
<td>CDU busyness</td>
<td>−0.035*** (0.002)</td>
</tr>
<tr>
<td>Phys. CDU rate</td>
<td>0.827*** (0.022)</td>
</tr>
<tr>
<td>N</td>
<td>429,313</td>
</tr>
<tr>
<td>Log-lik</td>
<td>−94,213</td>
</tr>
<tr>
<td>Pseudo-R²</td>
<td>0.224</td>
</tr>
</tbody>
</table>

Notes: All estimations made using a probit model specification; Robust standard error in parentheses; Likelihood ratio (Pr > χ²) < 0.0001 in all models.

*** p < 0.001, ** p < 0.01, * p < 0.05, † p < 0.10.
Table 4 Coefficient estimates to establish ED physicians’ response to increased uncertainty, using heckprob model specification.

<table>
<thead>
<tr>
<th>Decision made by ED physicians</th>
<th>Decision made in the CDU</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1e) AdmErr</td>
<td>(1c) AdmErr</td>
</tr>
<tr>
<td>(2e) DischErr</td>
<td>(2c) DischErr</td>
</tr>
<tr>
<td>(3e) SpecChg</td>
<td>(3c) SpecChg</td>
</tr>
<tr>
<td>ED busyness</td>
<td></td>
</tr>
<tr>
<td>0.036***</td>
<td>0.019</td>
</tr>
<tr>
<td>(0.005)</td>
<td>(0.017)</td>
</tr>
<tr>
<td>ρ</td>
<td></td>
</tr>
<tr>
<td>−0.420***</td>
<td>0.092</td>
</tr>
<tr>
<td>(0.035)</td>
<td>(0.127)</td>
</tr>
<tr>
<td>N</td>
<td>429,313</td>
</tr>
<tr>
<td>N uncensored</td>
<td>429,313</td>
</tr>
<tr>
<td>Log-lik</td>
<td>−152,824</td>
</tr>
</tbody>
</table>

Notes: All estimations made using the heckprob model specification; Robust standard error in parentheses; Likelihood ratio (Pr > χ²) < 0.0001 in all models.

"***p < 0.001, **p < 0.01, *p < 0.05, †p < 0.10.

that ED physicians are not simply becoming more error prone as they become busy (since we would expect a similar increase in both of these types of errors). This supports Hypothesis 2.

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

To give an idea of the scale of the effects, we convert coefficient estimates into average partial (marginal) effects (APEs) with 95% confidence intervals (CI95%). A one standard deviation increase in ED busyness increases the probability of admission to the CDU by 0.79%, CI95 = (0.70%, 0.88%), of being admitted by an ED physician in error by 0.36%, CI95 = (0.26%, 0.46%), and decreases the probability of being discharged in error by −0.03%, CI95 = (-0.07%, 0.00%). Compared with the average rate of CDU use, false admissions and false discharges reported in Table 1, this represents a relative increase (decrease) of approx. 9.7%, 7.6% and −3.5%, respectively. Thus moving from a low to high busyness state in the ED, i.e. from −2σ to +2σ, will have a surprisingly large impact, especially on CDU use and false admission rates. For example, assuming a cost of £500 per false admission, if all 651,044 patients had been treated in the ED in a high busyness state rather than low then over-referral by ED physicians would have cost the hospital approx. £4.7 million more.

5.4. Robustness to Endogeneity Concerns

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

6. The Two-Stage Gatekeeping System

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

The two-stage gatekeeping system enables gatekeepers to before making a referral decision judge whether sufficient information is available for accurate diagnosis and, if not, to pass the customer downstream to another gatekeeper who assumes responsibility for the referral decision (see e.g. Figure 1). The gatekeepers in this second service stage are more experienced than those in the first
stage and are allocated more resources and time in order to resolve uncertainty in diagnosis. Since only those customers for whom the original referral decision is ambiguous should be passed to this second gatekeeping stage, the more experienced (and costly) servers will be expected to spend little of their time with unambiguous cases. Moreover, since capacity in the first service stage is left unchanged, the increase in time available for resolving uncertainty will be allocated specifically to those customers who stand to benefit the most (i.e. those passed to the second stage). So long as gatekeepers in the first stage refer to the second stage only those customers for whom there exists high enough classification uncertainty, then fewer errors should be made in referral decisions than if those patients had instead been referred directly by the first-stage gatekeeper. This is similar in concept to that of complexity-augmented triage proposed in Saghaian et al. (2014b), which recommends first triaging patients who arrive at the ED on the relative complexity of diagnosis and then on their degree of urgency. Similarly, we expect that streaming patients based on residual uncertain should increase the overall accuracy of referral decisions.

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

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

In the rest of this section, we first describe how the ED–CDU interaction operates like the two-stage gatekeeping system, before introducing the variables that are used to capture diagnostic uncertainty and the physician’s profile.

### 6.1. The Clinical Decisions Unit

The clinical decisions unit (also known as an observation unit) is a dedicated area for emergency patients of low to moderate risk that exists separate to the main ED and general hospital units. The unit is designed to provide services such as further diagnostic evaluation, additional testing, and continuation of therapy for patients who require care beyond the initial level that can be provided in the ED (Ross et al. 2012). Patients admitted to the CDU are expected to have symptom complexes that can be resolved within a six-to-24 hour period, with further assessment determining whether inpatient admission is required at the end of their CDU stay (Hassan 2003). These units also typically benefit from the presence of specialist trained and more senior staff, as well as advanced diagnostic capacity. As a consequence, various advantages of such units have been identified in the literature, such as improved patient satisfaction and safety and shorter stays (see Cooke et al. 2003, for an excellent survey). It is also thought that making greater use of decisions units can result in considerable cost savings, estimated in one study at $3.1 billion per year (Baugh et al. 2012).
Thus, while generally it is believed that CDUs are an effective alternative to inpatient admission, to our awareness no studies have looked at the impact on transfer or discharge error rates, nor at those characteristics of patients and physician’s that influence the CDU’s effectiveness.

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

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

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

7.1. Empirical Specification

The empirical approach that we adopt is similar to that described in Section 5.1, except that rather than use a heckprob model we estimate the models instead with a recursive bivariate probit (biprobit) model, again with full information maximum likelihood (Maddala 1983). These models have the same error structure as the heckprob model but differ in that censoring is not performed and \( \alpha_2, \beta_2, \gamma_2 \) are left as free parameters to be estimated in the models. We first ask whether there is evidence that (and, if so, the extent to which) decoupling the gatekeeping decision and allowing ED physicians to, when they are uncertain, pass on the referral decision to a second gatekeeping stage can help to reduce the overuse of specialists and referral of patients to the wrong specialists. This would be confirmed by coefficients \( \beta_2 < 0 \) and \( \gamma_2 < 0 \) in the respective outcome equations.
We are also interested in if there is any evidence of a change in discharge errors, estimated by $\alpha_2$, when patients are routed through the CDU.

### 7.2. Results: The CDU Effect

Looking first at the question of whether patients admitted to the CDU have lower admission and discharge error rates and less chance of being referred to the wrong specialist, we find in Table 5 evidence of positive correlation in each of the three respective biprobit models, with estimated correlation coefficients $\rho = 0.292$ ($p$-value $< 0.001$), $\rho = 0.074$ ($p$-value $= 0.038$), and $\rho = 0.095$ ($p$-value $= 0.033$), respectively. This suggests that patients selected for admission to the CDU are more likely to be a false admission, false discharge, and to require specialty transfer than an ‘average’ patient who visits the ED. This is consistent with expectation: patients admitted to the CDU should be more complicated than the average ED arrival, else this more expensive service would be being used inappropriately. These biprobit model estimates provide strong evidence that patients admitted to the CDU are significantly less likely to (i) result in false admission (coef. $= -0.740$, $p$-value $< 0.001$ in column (2o)) and to (ii) require a transfer of specialty after admission (coef. $= -0.351$, $p$-value $< 0.001$ in column (3o)), with (iii) no corresponding increase in discharge errors reported (coef. $= 0.075$, $p$-value $> 0.10$ in column (3o)). This confirms our hypothesis that routing customers with unresolved diagnostic uncertainty through a two-stage gatekeeping system can help to significantly reduce the number of referral errors made in systems staffed by gatekeepers with a low tolerance for risk of non-referral errors, as well as helping as a secondary benefit to ensure that customers are referred to the correct specialist.

To see how much better admission decisions are when made in the CDU rather than by an ED physician, we convert coefficient estimates to average treatment effects (ATEs) and average treatment effects on the treated (ATTs). These results show that if no patients were referred
through the CDU the rate of admission, discharge and specialty routing errors would have been 5.72%, 0.92%, and 22.8%, respectively. These change to 1.45%, 1.12% and 15.2%, respectively, if all patients are instead routed through the CDU. Thus the CDU acts to significantly reduce admission and specialty transfer errors with little change in discharge errors. Moreover, the especially large and negative ATT for admission errors, −12.2%, suggests that ED physicians are especially good at routing into the CDU patients who they would have otherwise been admitted in error, and also that the CDU significantly reduces the rate of false admissions for these patients.

8. Managerial Implications and Conclusions
Our results suggest that decoupling the gatekeeping decision and allowing gatekeepers the possibility of referring patients on to a third party who assumes responsibility for the referral decision may help to improve system performance by reducing the rate of over-referrals. One possibility that we should consider, however, is that if the CDU were not present then resources used to operate the CDU could instead be redeployed to the ED. Since we have found that fewer false admissions occur when the ED is less busy, this increase in ED resourcing might perhaps be more beneficial than the presence of the CDU. To assess whether this is the case, we now perform a counterfactual analysis based around merging the CDU with the ED and pooling capacity.

Over our six year sample period the total number of hours spent by patients in the ED was 1.5m, with 326k hours spent by patients in the CDU. If the ED and CDU were merged, therefore, capacity in the ED would increase by approx. 21.7%. To estimate the effect that the increase in capacity would have on the rate or admission errors, we must adjust observed ED busyness downwards to account for the fact that more resources (e.g., physicians, nurses, treatment rooms) would have been available. To do this we multiply QueueED_{95th}^h, our measure of capacity based off of the 95th percentile of ED busyness in Section 4.4, by 1.217 and re-estimate OccED_i for all i. To ensure that the original and updated measures of ED busyness are on the same scale, we standardize using the original mean, \mu(OccED_i), and standard deviation, \sigma(OccED_i). On average this has the effect of reducing workload by ~0.64\sigma. Substituting the original values of zOccED_i for the updated values achieved through pooling ED and CDU capacity suggests there would be an approx. 0.23% reduction in false admissions. On the other hand, using the model estimated by biprobit (1o) in Table 5, had the CDU not been available then the average rate of admission errors would have been 5.72%, as compared to an average rate in the sample of 4.71%, a reduction of approx. 1.01%. This suggests that the pre-screening of patients that takes place in the ED prior

\[4\] Note that we take a very conservative view and assume that all of those patients who were treated in the CDU could have instead been relocated elsewhere in the hospital without any additional capacity needing to be installed, meaning that all resources from the CDU can be redeployed to the ED. We thus estimate an upper bound on the gains that could be achieved from pooling ED and CDU capacity.
to offload to the CDU allows resources in the CDU to be targeted at those patients who benefit the most from more specialized and higher intensity service, which confers advantages above and beyond those that would occur if those resources were allocated at random in the ED.

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

Our findings also provide an alternative strategy to adding capacity that can be implemented in multi-tiered service contexts to improve the accuracy at which customers are routed to the appropriate service provider. By allowing gatekeepers the option of passing the referral decision to a higher intensity gatekeeping tier (rather than directly to the even more costly specialist) when diagnostic uncertainty is high, the first-stage gatekeeper can focus on processing those more unambiguous cases while the second-stage gatekeeper searches for the appropriate treatment option for the more complex cases. An interesting possible direction for future research might be to translate our empirical findings to a modeling approach and to investigate the two-stage gatekeeping system analytically.

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

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

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

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

Appendix B: Control variables
In Table 6 we describe the variables used as controls in the models.

References
### Table 6  Table of controls.

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporal (T.)</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Categorical (6)</td>
</tr>
<tr>
<td>Daily time trend</td>
<td>Continuous</td>
</tr>
<tr>
<td>Month</td>
<td>Categorical (12)</td>
</tr>
<tr>
<td>School break</td>
<td>Categorical (7)</td>
</tr>
<tr>
<td>Day of week</td>
<td>Categorical (7)</td>
</tr>
<tr>
<td>Window of arrival x weekend</td>
<td>Categorical (24)</td>
</tr>
<tr>
<td>Diagnosis related factors (D.)</td>
<td></td>
</tr>
<tr>
<td>Diagnostic category</td>
<td>Categorical (17)</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Categorical (46)</td>
</tr>
<tr>
<td>Affected region of body</td>
<td>Categorical (7)</td>
</tr>
<tr>
<td>Procedure performed</td>
<td>Categorical (22)</td>
</tr>
<tr>
<td>Contextual factors (C,)</td>
<td></td>
</tr>
<tr>
<td>Mode of arrival</td>
<td>Categorical (8)</td>
</tr>
<tr>
<td>ED visits, last year</td>
<td>Continuous</td>
</tr>
<tr>
<td>ED visits, last month</td>
<td>Continuous</td>
</tr>
<tr>
<td>Admissions per ED visit, last year</td>
<td>Continuous</td>
</tr>
<tr>
<td>Admissions per ED visit, last month</td>
<td>Continuous</td>
</tr>
<tr>
<td>Zero ED visits, last year</td>
<td>Binary</td>
</tr>
<tr>
<td>Zero ED visits, last month</td>
<td>Binary</td>
</tr>
<tr>
<td>Physician related factors (P.)</td>
<td></td>
</tr>
<tr>
<td>Admission errors</td>
<td>Continuous</td>
</tr>
<tr>
<td>Discharge errors</td>
<td>Continuous</td>
</tr>
<tr>
<td>Admission errors x Discharge errors</td>
<td>Continuous</td>
</tr>
<tr>
<td>Physician category</td>
<td>Categorical (14)</td>
</tr>
<tr>
<td>Operational/other factors (O.)</td>
<td></td>
</tr>
<tr>
<td>Length of stay in ED</td>
<td>Categorical (13)</td>
</tr>
<tr>
<td>Hospital congestion</td>
<td>Continuous</td>
</tr>
</tbody>
</table>

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


