Why are patients waiting longer for emergency care in England?

Using applied economics to answer real-world operational challenges

January 2019

Summary: What drives winter A&E performance

The problem

Many providers did not meet the A&E four-hour maximum waiting time standard in winter 2017/18. To help improve next winter, we wanted to understand what was behind this.

Our analysis

We have undertaken detailed analysis of what drives A&E performance using daily data from winters 2016/17-17/18. We have used econometric analysis which allows us bring together all the factors and isolate the effect of each on A&E performance. Our key findings are shown below.

Demand

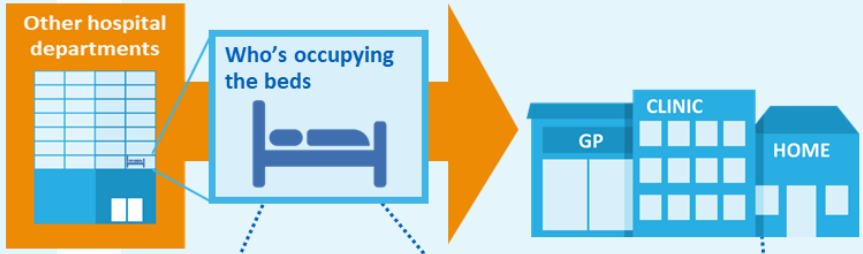
Winter often sees higher numbers of patients with different types of illnesses

A&E department

The A&E departments need to respond to these additional patient needs in winter

In the hospital

Maintaining capacity in the hospitals ensures new patients can be admitted from the emergency department, avoiding bottlenecks.



Key factors

Admissions

High **daily** and **uneven hourly** admissions decrease performance

Flu

1/3rd of the growth in emergency admissions came from flu in winter 17/18

Workforce

Senior doctors make quicker decisions so patients wait for less time

Resilience

How well hospitals **respond** to and **recover** from pressure affects performance

Bed occupancy

Rising occupancy reduces performance, with accelerating effects above **92%**

Long-stay patients

Long-stay patients can decrease performance by reducing bed flexibility

Discharges

Discharges **on the day** and **previous days** improve performance

Introduction

In recent years, patients have been waiting longer for treatment in Accident and Emergency (A&E) departments in the National Health Service (NHS). In England the NHS constitutional standard stipulates that 95% of patients attending A&E should be seen, treated and admitted or discharged within four hours. Performance against this four-hour standard – 'A&E performance' – is a prominent metric that hospitals are measured against, often featuring in newspaper headlines and politicians' list of concerns.

After more than five years of deterioration, A&E performance reached a new low of 77% in winter 2017/18, a striking symptom of the substantial pressure that hospitals are struggling to cope with. But A&E performance is not just a performance management metric. A performance of 77% means that almost one in four patients in A&E was waiting over four hours to be dealt with. Longer waiting times have a real impact on patients: they are associated with higher patient mortality and worse outcomes[1]. Accordingly, there has been a strong focus on improving A&E performance.

Identifying which factors have a link to A&E performance is crucial to understanding where to target interventions to effectively recover performance. The decline in performance has come at a time of real challenge for the NHS. Money is tight, capacity to treat patients is reaching its limit, but demand on front-line services continues to grow at pace.

Hospitals do have influence over waiting times in A&E, and so there is practical value in understanding the relationship between A&E performance and those short-term operational factors that are within their control. But they make up just one part of a much larger, complex system of health and social care delivery. We therefore need to understand the extent to which A&E performance is also determined by factors outside the hospital, which may require more fundamental, long-term changes to how we deliver health and social care.

This issue lends itself well to economic frameworks and analysis. The delivery of urgent and emergency care is highly complex, influenced by a range of factors that

[1] Gruber J, Hoe T, Stoye G. Saving Lives by Tying Hands: The Unexpected Effects of Constraining Health Care Providers, NBER Working Paper No. 24445 2018

can be volatile and interconnected, especially during high pressure periods. The specific effects of each factor may not always be clear when looking at headline metrics. Other analytical methods used to analyse this complex system, such as operational research and data science, can observe factors that are correlated. However, the absence of a hypothesis-led approach limits the ability to interpret and apply the results. We have therefore developed an economic framework, and applied econometric analysis, to isolate and identify the factors that had the most effect on winter A&E performance in England in 2016/17 and 2017/18.

This type of question is not usually tackled from an economic perspective, having often been perceived to be an operational issue. However, we believe that there is a real opportunity for economic frameworks to add value and to take a different perspective. Fundamentally, although emergency care is a complex system, it can be modelled using a supply and demand framework. It is likely that economists can apply similar techniques to model other complex systems and produce insights that make a real impact on policy. We believe this makes this work even more important, by demonstrating the value of using economics in non-traditional settings.

This work builds on the existing literature, including our 2015 report A&E delays: why did patients wait longer last winter? which established the importance of admissions and bed occupancy in influencing A&E performance. The King's Fund analysed hospital capacity in more detail (NHS hospital bed numbers: past, present, future), and focused on increasing bed shortages. Our analysis adds to this evidence base by looking at daily data on supply side factors in the hospital, including identifying tipping points past which A&E performance deteriorates rapidly.

We find that daily bed occupancy and the types of patients in the beds, especially long-stay patients, are strongly related to A&E performance. On the demand side, we find that higher levels of emergency admissions and instances of flu have an adverse effect on A&E performance.

Returning to shorter waiting times in A&E will not just require improvements in operational factors within A&E departments and hospitals, but a wider rebalancing of demand and supply across the whole health sector. Indeed, as set out in the recent Long Term Plan, there are widescale plans to increase capacity outside of the hospital – for example in the community, care homes and GPs – to rebalance demand away from A&E departments and across the wider system.

Method

Economic framework

There is no explicit price for emergency care in England: healthcare is free at the point of use. This matters because in many other markets prices balance supply and demand. Individuals only buy goods or services that cost less than or equal to the amount they value them at.

Within emergency care, demand and supply are instead balanced through waiting times, which operate as the 'shadow price' of emergency care. Patients seeking care decide whether it is worth the cost of their time spent waiting. Using this economic framework, we are able to analyse A&E performance as being determined by a range of supply and demand factors:

Supply side:

- **Bed occupancy** – a measure of supply available in the wider hospital, defined as the proportion of available beds that are occupied at a set time each day. We include further detail on beds by also looking at:
  - **Long-stay patients** – how many beds are occupied by patients that have been in hospital for 21 days or more
  - **Discharges** – a comparison of the number of patients leaving hospital beds and the number coming in each day
- **Workforce** – the types of staff working in the A&E department
- **Resilience** – other local factors that affect how well a hospital is able to deal with surges in demand, but that are difficult to quantify.

Demand side:

- **Emergency admissions** – our key measure of demand, capturing the number of patients who are admitted into hospital as an emergency[2].
- **Flu** – capturing the number of beds occupied by flu patients.

Notably, these include factors outside the A&E department. This is because any imbalances in supply and demand in the rest of the hospital cause bottlenecks that affect patients waiting to be admitted from the A&E department.

[2] In line with previous research, we use admissions rather than attendances at the A&E department, to better capture demand for the overall emergency care

Econometric model

The complex nature of the emergency care system, combined with the large range of volatile data available to observe it, makes it extremely difficult to determine what is driving performance. There are large differences in A&E performance between providers, often driven by static, local factors that are difficult to quantify (e.g. culture or leadership) – which makes analysis of aggregated, sector-level data misleading. The econometric approach we have used ensures these complexities are properly considered.

We brought together a large dataset, covering 137 providers (all those with a type 1 A&E department, which treat the most severe and life-threatening emergency patients[3]). We analysed the data for each day between 1 December and 28 February in both 2016/17 and 2017/18. We used a range of sources: daily Sitrep data for bed occupancy, Hospital Episode Statistics (HES) for patient demand and Electronic Staffing Records (ESR) for workforce. This allowed us to identify operational factors, patient characteristics and substantive staff in A&E departments. Then, by applying econometric analysis, we looked at the effects of many factors at the same time and determine which are statistically significant and most influenced performance. We controlled for quality, size of A&E department and patient characteristics; all our findings are in addition to the effect of these.

The econometric technique we use is a pooled ordinary least squares (OLS) model. This allows us to identify general characteristics of A&E performance, applicable across all the providers in the sample, to inform national policy and support. We tested the robustness of our approach by comparing it to other techniques – panel fixed effects and fractional response models - and found they yielded very similar results. This suggests that our findings are robust across a range of specifications and assumptions.

Our model was estimated using 2016/17 winter data, and then applied to data from winter 2017/18. We found the results to be very similar across both winters, adding confidence in our approach and findings. This suggests that the factors we identify are reasonably stable over time and will hold for next winter too – allowing us to better direct our support to providers.

[3] Type 1 A&E departments provide a consultant-led 24-hour service

Findings

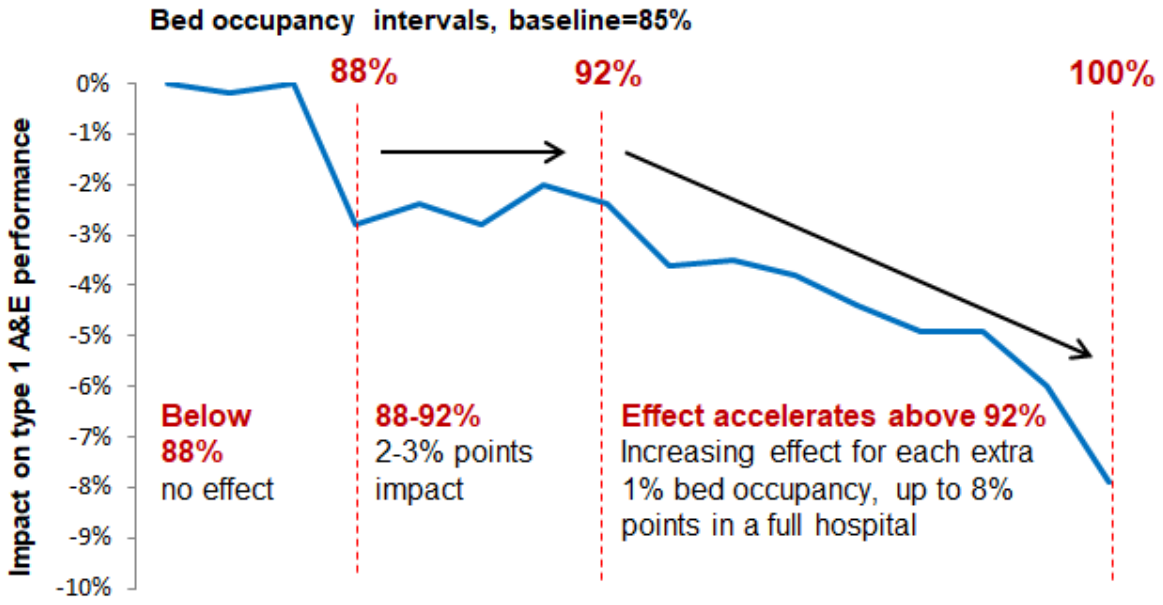
Each factor that influenced A&E performance is outlined below, grouped into supply side factors (in the hospital and the A&E department) and demand side factors.

Supply side factors in the hospital

Bed occupancy

When bed occupancy is high, A&E departments can spend a lot of time looking for available beds for patients who need to be admitted. These patients then wait longer in A&E and are more likely to breach the four-hour standard.

Figure 1: Impact on A&E performance of changing bed occupancy from 85% to each interval[4]



Bed occupancy is strongly associated with A&E performance across both winters.[5] We found significant disruption starts around 87% to 88% bed occupancy, indicated by a fall of 2 to 3 percentage points in A&E performance (see Figure 1 below). There is a tipping point above 92% bed occupancy, where the effect on A&E performance accelerates. A full hospital – 100% bed occupancy – has an A&E

[4] Figure 1 shows the effect of bed occupancy on type 1 A&E performance. This is broadly indicative of both winters.
[5] Bed occupancy was included in the analysis as discrete intervals to allow us to identify tipping points. We use general and acute beds only

performance 8 percentage points worse than another hospital with the same characteristics but bed occupancy of 85%.

Spot markets for emergency care?

This type of non-linear tipping point is a common feature observed in spot markets. Energy consumers tend not to be exposed to the real-time variations in the wholesale price of energy, and so price signals do not lead to customers adjusting their consumption in response. At the same time, energy generation capacity tends to be fixed and slow to respond. This makes both short-run energy demand and supply highly price inelastic, which manifests in very volatile prices and significant strain on energy systems.

In A&E departments this same situation takes place: patients do not have visible information on real-time waiting times (the shadow price of care) and often have no choice over the timing of their use of A&E, making their demand very inelastic to waiting times. The supply of available beds is also very inelastic to waiting times, with extremely limited options to expand capacity at short notice. As in energy, this results in a system characterised by volatile waiting times fluctuations, with extreme spikes in waiting times as bed occupancy approaches 100%.

Some hospitals can operate well with high bed occupancy while maintaining good performance, by compensating with the other factors we've identified below (for example through higher resilience or a more senior workforce) or others we have not analysed (e.g. by narrowing the time between beds becoming available and being filled).

This finding of a tipping point at 92% is particularly important because it updates previous analysis from 1999, which found a tipping point above 85% bed occupancy in two hospitals. This research has been widely used to inform hospital decisions and national policy. Since the 1999 research was completed, there have been substantial changes in the NHS, including a large rise in demand and restructuring of emergency care pathways. It is therefore important that this widely used statistic is updated to reflect the challenges hospital providers current face.

Long-stay patients

Hospitals with more long-stay patients – those who have been in hospital for 21 days or more – have additional bed pressures. Long-stay patients have two effects: on bed occupancy and bed flexibility. The effect on bed occupancy is already captured in the results reported in the section above. Long-stay patients affect bed flexibility by reducing the proportion of beds that providers can quickly free up to urgently accommodate new patients who need admitting for treatment.

We know that length of stay is influenced by case-mix and population demographic differences, so some providers will justifiably have more patients who stay for longer. To account for this, we take the number of patients who have been in hospital for seven days or more, and then look at the proportion of those who have been in for 21 days or more. These long-stay patients may have a particularly large effect because they account for a disproportionately large amount of bed capacity – despite making up just 4% of patients they account for around 40% of total bed days.[6]

Our analysis finds a 10-percentage point increase in the proportion of 7+ day patients who are long-stay (21+ day) patients is associated with a 2 to 3 percentage points decline in A&E performance. This effect assumes bed occupancy remains steady.

While we identified a strong relationship with long-stay patients, our model did not find a significant link for delayed transfers of care (DTOCs) or stranded patients. Given that DTOCs and stranded patients mainly affect performance through bed availability and flexibility, it is likely that the bed occupancy and long-stay patient metrics will have captured most of this effect.

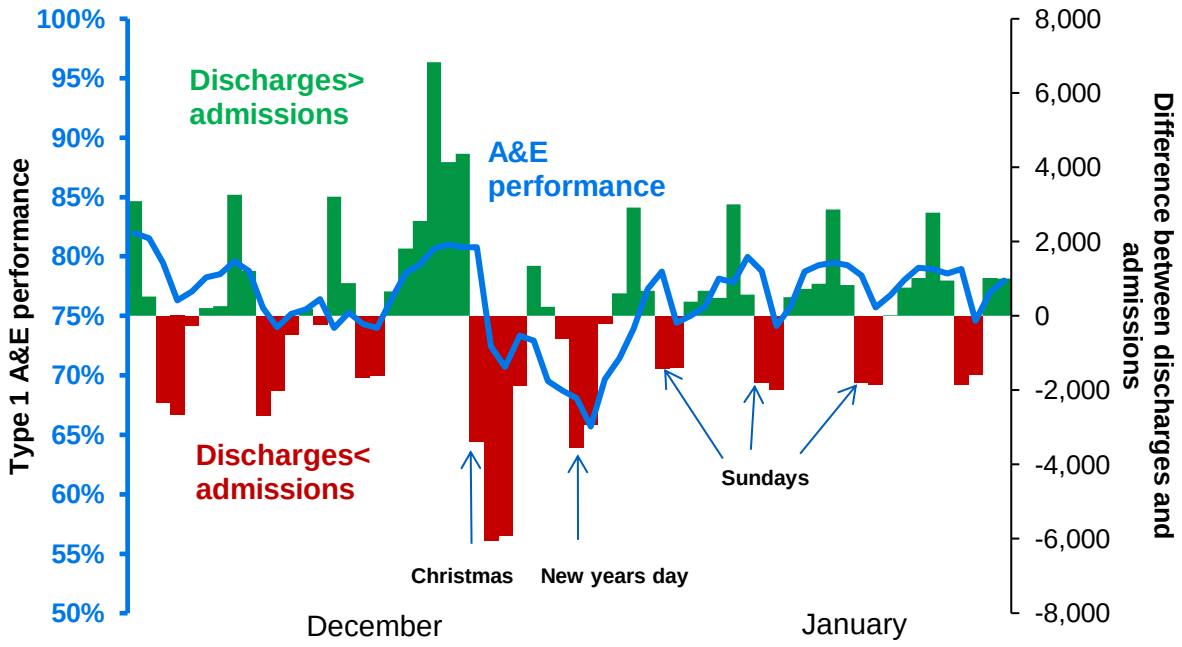
Discharges

Discharges play a key role in maintaining patient flow. Short term imbalances in supply and demand for inpatient hospital beds can occur because supply cannot be stored; it must be available at the time that emergency patients need it. We find that lower levels of discharges (relative to the number of non-elective admissions) leads to poorer A&E performance (see Figure 2). This not only affects A&E performance on the same day, but also in the two subsequent days – the effect of pressure, the

[6] HES data FY 2016/17. Patients admitted to an acute provider, excluding mental health specialties and patients with a length of stay more than 365 days.

difference between admissions and discharges, accumulates over time. If a hospital has a 10% pressure increase on Saturday and Sunday, when Monday arrives it is already under strain of having extra patients in beds, which means its performance would on average be between 0.7% to 0.8% points worse. If it is not able to increase discharges on the Monday and has another 10% pressure increase, this would reduce performance by a further 0.4% points.

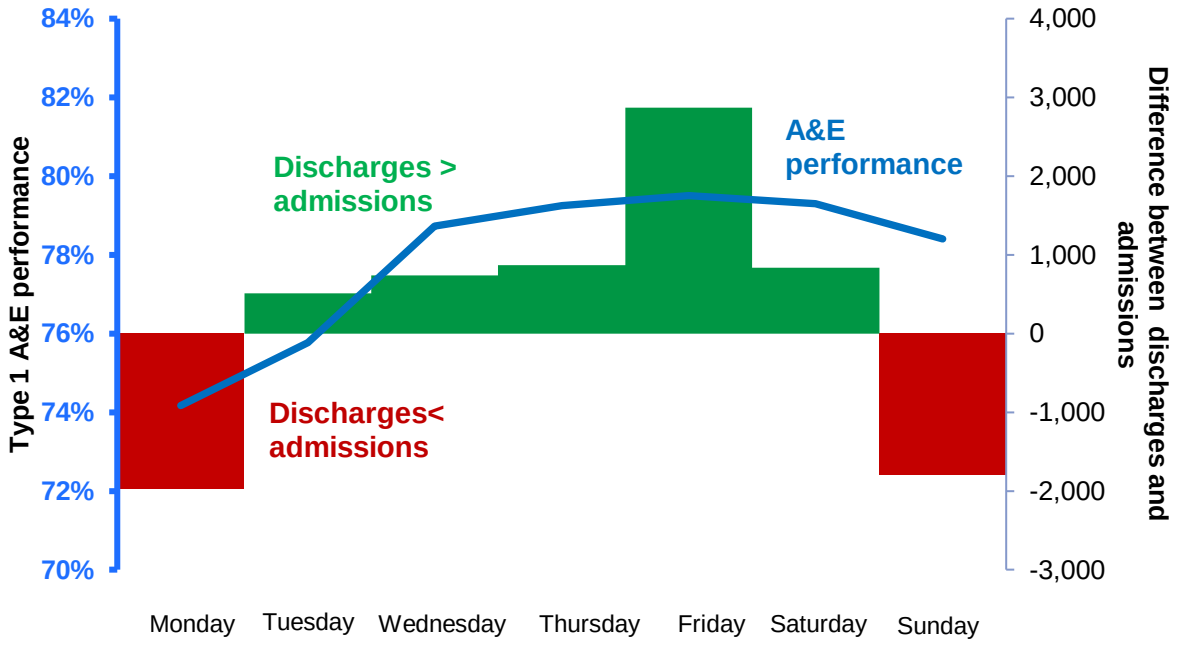
Figure 2: National winter 2017/18 pressure in the system



Source: HES data

Figure 2 shows how the differences between discharges and admissions build pressure over time during winter 2017/18, which is strongly related to A&E performance. Pressure decreased in the run up to Christmas then experienced the largest increase following Christmas. For the rest of winter, there is a clear weekly pattern of pressure increasing on Sunday and Monday then reducing from Wednesday to Friday.

Figure 3: National winter 2017/18 pressure in the system – one week in mid-January



Source: HES data

This variation in discharges helps to explain the differences in A&E performance on each day of the week. For example, lower weekend discharges reduce the available bed base for the hospital's busiest days (Monday and Tuesday) contributing to poorer performance on these days. Figure 3 illustrates this pattern for a typical week in mid-January.

Supply side factors in the A&E department

What happens in the A&E department itself is also crucial for A&E performance. A&E departments need to run efficiently to manage the increased volume of patient flow in winter, and the degree to which they can respond affects performance. Our model identified two factors that determine how well A&E departments perform:

- the type of workforce they have – which can determine how quickly patients can be seen, treated and admitted or discharged.
- Resilience - driven by medium-term, institutional capacity (leadership, culture, operational processes), which we capture by analysing how well providers respond to surges and deteriorating conditions over the whole winter

Workforce

We found that a 1 percentage point increase in the proportion of A&E staff who are senior doctors[7] increased A&E performance by 0.1 to 0.2 percentage points. These senior staff are often responsible for making rapid treatment decisions, drawing on their long experience, which improves patient flow.

Our analysis may well underestimate the effect of workforce on A&E performance. We are limited to monthly workforce data for substantive staff only, from ESR data. This means that we cannot observe how much daily or shift-level fluctuation in staff resources affects performance, or what effect staff vacancies or temporary staff may have.

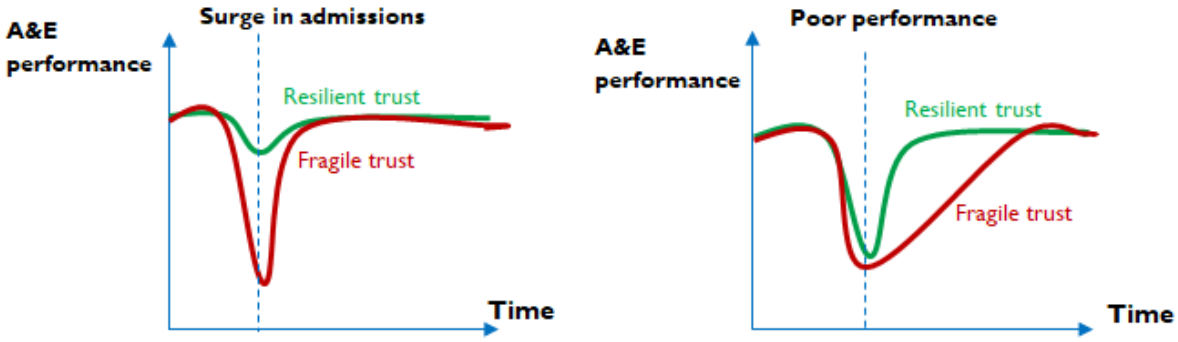
Resilience

Measuring resilience helps us identify the differences in performance between providers, which are often driven by static, local factors that are difficult to quantify. We define resilience in two ways.

First, how much an A&E departments' performance falls after a surge in admissions (Figure 4, chart left). We split the 137 providers into five categories based on their resilience, with the most resilient providers experiencing half the dip in A&E performance than the least resilient. This resilience factor does not appear to be related to the other operational factors – more resilient providers do not seem to have lower bed occupancy or more senior workforce for example. This could therefore be picking up factors that we cannot measure, such as managerial capacity, culture or leadership.

[7] Senior doctors are defined as registrars and above.

Figure 4: Provider resilience from a surge in admissions and bounce-back from poor performance



The second resilience factor measures how many days it takes a provider to bounce back from poor A&E performance (Figure 4, chart right). The most resilient providers were able to bounce back from dips in A&E performance by the next day. The least resilient providers took up to three days to recover. Comparing resilient providers with less resilient providers suggests that the ability to bounce back is driven by better operational capacity – e.g. lower bed occupancy and fewer long-stay patients.

Demand side factors

Admissions variation

Demand is the inflow into the emergency care pathway. Our previous research showed that admissions affect A&E performance more than the total number of people turning up at A&E (known as attendances),[8] although high numbers of attendances do affect safety and staff workload. More admissions stretch the capacity in the A&E department and in other hospital departments that need to find beds to admit patients to. We have extended this evidence by looking at two admission-based measures for demand in our model: daily admissions variation from the 90-day winter average, and hourly admissions variation. These account for both how many admissions there were in a day, and how spread out those admissions were during that day.

In line with previous work on A&E performance, we find higher than average daily admissions have an adverse effect on A&E performance. A 10-percentage point

[8] 7 A&E delays: why did patients wait longer last winter? https://www.gov.uk/government/publications/ae-delays-why-did-patients-wait-longer-last-winter

increase in admissions was associated with between a 0.2 to 1 percentage point decrease in A&E performance.

Hourly admissions variation can be hard for hospitals to accommodate if they do not have spare capacity. A smoother stream of admissions can give hospitals more time to free up beds by discharging patients gradually through the day. Providers with the highest hourly variation in emergency admissions had on average 3% points worse A&E performance than the lowest variation providers.

Patients under 10, over 60, and those referred by GPs have higher hourly admissions variation than other patient groups. This is driven by a greater number of admissions from these groups in the evening.

Flu

Flu affects individual patients, but also has externalities in hospitals due to its contagiousness. High rates of flu add to demand pressures, worsen patient flow and can spread infection to staff. In winter 2017/18, flu-related non-elective hospital admissions were over three times higher than the three previous winters[9]. This accounted for about a third of the emergency admissions growth between winter 2016/17 and 2017/18. Patient flow is affected because patients with flu generally stay for longer and are isolated to minimise contagion, which reduces the flexibility of beds.

It is difficult to quantify the precise effect of flu on A&E performance. We have identified flu patients as those with flu as their primary or secondary diagnosis in HES data. Recording of diagnoses in HES can be variable, but our initial analysis suggests that a one percentage point increase in the proportion of general and acute beds occupied by flu patients decreased A&E performance by between 1.0% and 1.2%. This effect assumes bed occupancy and emergency admissions remain stable. Although this is a large effect, only five percent of days in any hospitals across the two winters analysed exceeded 1.8% beds occupied by patients with flu. A one percentage point increase therefore represents a large change in flu cases.

[9] Data from HES. Patients with flu are identified as those whose primary or secondary diagnosis code is J09-J12.

Conclusions

By bringing together high frequency, daily data on A&E performance and a wide range of operational factors, we have been able to isolate and identify those factors that had a significant effect on A&E performance during the winters of 2016/17 and 2017/18. We have shown a strong link between bed occupancy and A&E performance, including updated analysis of where there are tipping points above which A&E performance deteriorates more quickly. And through our analysis of long-stay patients we have shown that the flexibility of the bed base is an important factor.

Now we know what is driving the deterioration in A&E performance we can confidently take steps to address it. Capacity, measured by beds and senior workforce, is clearly crucial. This points to a need for longer-term solutions, as hospitals will quickly reach the limit on their capacity without additional funding and support from community care. The NHS's recent Long Term Plan aims to support this. With a focus on using additional funding to expand capacity out of hospitals, the Long Term Plan will allow patients to be treated elsewhere and discharged from hospitals more quickly – directly influencing the bed occupancy and long-stay patients factors in our model.

This analysis has also supported the wider work to prepare the NHS for winter. In winter 2018/19 hospital providers were asked by NHS Improvement and NHS England to reduce the number of long-stay patients in acute hospitals by 25%. This accompanies ongoing work to reduce bed occupancy to the 92% tipping point and expedite discharges.