Naming & Shaming: The impacts of different regimes on hospital waiting times in England and Wales*

Timothy J. Besley[†] **Gwyn Bevan**[‡] **Konrad B. Burchardi**[§]

May, 2009

Abstract

Improving accountability in public services has been a central objective of many public sector reforms in recent years. Chief among these have been efforts to generate observable performance measures as a basis for monitoring performance. This paper examines a natural experiment in regimes applied to waiting list targets for hospital admissions in England and Wales. Prior to 2001, each country had similar policies, organisational structures for hospital care, and levels of resources. After 2001, the principal difference between the countries were the consequences for hospitals that failed to meet targets for waiting times: in England, failure resulted in sanctions in a process of 'naming and shaming', but in Wales, failure was perceived to result in extra resources. We use hospitals in Wales as a 'control group', to examine the effect of 'naming and shaming' in England. We found that this policy did indeed reduce waiting times in England as compared with Wales. However, there is some evidence there was in England, initially, some shuffling of prospective patients to meet specific targets which increased mean waiting times.

JEL Classification: H11, I12, I18.
Keywords: hospital waiting times, targets.

*We are grateful to Oliver Bevan for constructing the database and to staff in the Welsh Assembly Government for help in the supply of data from Welsh hospitals. We thank Oliver Bevan and Carol Propper for comments on an earlier draft, and seminar participants at Imperial College London for their feedback.

[†]LSE, Department of Economics, E-mail: t.besley@lse.ac.uk

[‡]LSE, Department of Management, E-mail: r.g.bevan@lse.ac.uk

[§]LSE, Department of Economics, E-mail: k.b.burchardi@lse.ac.uk, Tel: +44 (0)20 7852 3534, Fax: +44 (0)20 7955 6951

1 Introduction

Improving accountability in public services has been a central objective of many public sector reforms in recent years. Chief among these have been efforts to generate observable performance measures as a basis for monitoring performance which has been introduced in various forms by governments of most OECD countries to shift the focus from inputs to outcomes.[1] However, such efforts are not without controversy. Measurable performance criteria do not always reflect things which matter to consumers. Worse still, this can result in effort being directed away from desirable goals towards meeting the target as suggested in the multi-tasking model of Holmstrom and Milgrom (1991).

This paper examines the impacts of a regime of 'naming and shaming' for failure to achieve waiting list targets, which applied on a massive scale, to hospitals in the National Health Service (NHS) in England only. By the year 2000, responsibility for running the NHS in England, Scotland, and Wales was devolved to governments of each country (devolution was largely stalled in Northern Ireland). The NHS in each UK country received massive increases in funding, of 5 per cent in real terms over the six years from 2001-02 (Smee, 2005). Only the government in England, however, sought to change the system of perverse incentives that had developed across the different countries: from one that ignored success and rewarded failure to one that celebrated success and penalised failure. This was done through the radical and controversial system of annual 'star rating' of NHS organisations, between 2001 and 2005, which 'named and shamed' those that 'failed', which were zero rated; and offered 'earned autonomy' to the 'high-performing' three-star organisations. In Wales and Scotland the system of perverse incentives continued alongside the very different regime at work in England. The policy differences that have emerged following devolution offer a natural experiment to evaluate their impacts. Propper, Sutton, Whitnall, and Windmeijer (2008a,b) have compared performance in reducing hospital waiting times in England and Scotland. This paper compares that performance in England and Wales.

We estimate a difference-in-differences model of the proportion of people on the waiting list for different times in Wales and England at the level of a hospital trust.[2] Trust fixed effects allow us to control for sources of unobserved heterogeneity and we also control for common shocks through the inclusion of year dummy variables. We exploit the fact that the timing and nature of the treatment in England and Wales is different to identify the effect of the target on waiting lists. The results show that targets were indeed effective in bringing down the waiting times in England, where, for a NHS trust, with a median number of patients waiting in June 1999, the estimated effect of the 18-, 15- and 12-month targets is to have reduced the numbers of patients waiting longer than the

[1] For Canada, for example, see (Perrin, 2002).

[2] Earlier papers have highlighted differences at the national level in performance on waiting times (Alvarez-Rosete, Bevan, Mays, and Dixon, 2005; Bevan, 2006; Bevan and Hood, 2006a,b). Hauck and Street (2007) undertook a detailed analysis across three English hospital trusts and one Welsh hospital trust close to the English-Welsh border. Propper, Sutton, Whitnall, and Windmeijer (2008a) estimated difference-in-differences models of the proportion of people on the waiting list who waited over 6, 9 and 12 months in England and Scotland. All these analyses strongly suggest that the policy of star ratings did reduce waiting times in England.

targeted time to zero. The 9-month target is estimated to have reduced the number of patients waiting between 9 and 12 months by 67%.

This paper contributes to an emerging body of empirical literature that examines the consequences of applying performance measurement to public services. There are four points to make of this literature. First, despite the growth of the industry of performance measurement, literature reviews consistently highlight the paucity of rigorous evaluations (Marshall, Shekelle, Davies, and Smith, 2003; Rosenthal and Frank, 2006; Fung, Lim, Mattke, Damberg, and Shekelle, 2008; Burgess and Ratto, 2003). Second, the mere publication of performance measurement is often found to have had little impact: in US health care (Marshall, Shekelle, Davies, and Smith, 2003; Fung, Lim, Mattke, Damberg, and Shekelle, 2008); central government (Norway and the UK), primary schools (Norway), urban policy (the Netherlands), nuisance control (in a US city) (van Dooren and de Walle, 2009). Third, there have been a number of studies of linking performance measurement with financial incentives, which can affect incomes of organisations, teams or individuals or a mix of these: examples include tax collection (Brazil), job training (US), teachers (US and Israel), health care (US and UK) (Heckman, Heinrich, and Smith, 1997; Heinrich, 2002; Rosenthal and Frank, 2006; Burgess and Ratto, 2003; Campbell, Reeves, Kontopantelis, Middleton, Sibbald, and Roland, 2007; Doran, Fullwood, Reeves, Gravelle, and Roland, 2008). Fourth, for health care, Hibbard (Hibbard, Stockard, and Tusler, 2003, 2005; Hibbard, 2008) has argued that, another form of high powered incentives is generated by careful design of public reporting that inflicts reputational damage; there is, however only limited evidence from US health care to test this hypothesis (Bevan and Hamblin, 2009). Pawson (2002) gives numerous examples of 'naming and shaming' across diverse elds of public policy in the EU, US and UK. He examined the evidence of effectiveness of six examples: the Car Theft Index (CTI) for car manufacturers (UK); non-payers of local taxes (UK); hospital report cards (US); Sex Offender Registration (Megan's Law - US); the Healthcare Financing Administration's Mortality Studies (US); the Toxic Releases Inventory (TRI)(US). He argues that the evidence shows that of these six, only the Car Theft Index was effective. But Fung and ORourke (2000) produced more recent evidence on the TRI than that used by Pawson (2002): over its first eight years, the TRI reduced releases of toxic chemicals by 45 per cent. Fung and O'Rourke argued that there was strong evidence that the TRI had proved to be much more effective than traditional command and control because it resulted in public pressure being focused on the worst polluters.

The remainder of the paper is organized as follows. The next section gives the background policy context for the analysis. It outlines the common development of the organisation and governance of the NHS in the UK, the different regimes that then developed in England and Wales after devolution, and what is known about their impacts. Section three outlines the data and methodology used while section four presents the results. The concluding comments are in section five.

2 Background and Context

The NHS in the UK was created in 1948 to provide universal coverage financed by taxation, largely free at the point of delivery in a publicly-organised system of functional units (acute hospitals) and units defined territorially (for example, care for the mentally ill, ambulances, primary care, dentistry), which broadly allowed clinical autonomy to medical professionals in their decisions on treating patients Klein (2006). Periodic reorganizations changed the boundaries, names and nature of those sub-units, but not the system's other abiding characteristics. Reorganization in the 1970s created health authorities in England and Wales in hierarchical structures that were responsible for planning services for defined resident populations and running hospitals and community health services.[3]

From its inception, the prevailing view was that the NHS was staffed by publicly spirited workers who needed no incentives, sanctions or rewards – see Le Grand (2003) for further discussion. But this view began to change over time. For example, Enthoven (1985) claimed serious problems with the hierarchical organization of the NHS and its lack of incentives. He described the NHS as being in a 'gridlock': 'caught in the grip of forces that make change exceedingly difficult to bring about' (p. 9), the fundamental problem being that 'the system contains no serious incentives to guide the NHS in the direction of better quality of care at reduced cost' (p .13). He recommended the introduction of incentives by requiring providers to compete in an 'internal market'.

This view was ultimately influential in shaping the Thatcher government's pursuit of reform. In response, the so-called 'internal market', based on the principle of provider competition, was implemented between 1991 and 1997, with a funding system that promised that 'money would follow the patient'(Secretaries of State for Health, Wales, Northern Ireland and Scotland, 1989; Bevan and Robinson, 2005; Klein, 2006). This led to the reorganization of the health authorities into purchasers, which contracted for hospitals and community health services.

In spite of these reforms, waiting times remained a problem.[4] In response, the government in England announced new policies in 2000, with an objective of cutting maximum waiting times for elective admission from 18 months to 6 months by the end of 2005 (Secretary of State for Health, 2000). The principal policy instrument for delivering this transformation was the system of 'star ratings', which applied to acute hospitals from 2001 to 2005 (Department of Health, 2001, 2002; Commission for Health Improvement, 2003a,b; Healthcare Commission, 2004, 2005). This process gave each organization a score from zero to three stars based on performance against a small number

[3]Different legislation applied the same principles with the creation of Health Boards in Scotland and Health and Social Service Boards in Northern Ireland.

[4]Failing providers do not exit the market (Tuohy, 1999; Enthoven, 1999; Secretary of State for Health, 2000; Bevan and Robinson, 2005). It has also proven difficult to create an effective demand-side either by commissioning services through purchasing organisations or patient choice. The evidence from two systematic reviews (Marshall, Shekelle, Davies, and Smith, 2003; Fung, Lim, Mattke, Damberg, and Shekelle, 2008) of the literature on the effects of publishing information on hospital performance found that patients did not respond as consumers to use evidence on hospital performance to switch from poor to good hospitals.

of 'key targets' and a larger set of targets and indicators in a 'balanced scorecard'. Organizations that failed against 'key targets', and were 'zero-rated', were 'named and shamed' as 'failing', and their chief executives were at risk losing their jobs: this happened to six chief executives of the 12 trusts given 'zero rating' in 2001 and four of these improved their rankings in the following year's star ratings (Beverley and Haynes, 2005). Organizations that performed well on both the 'key targets' and the 'balanced scorecard', and achieved the highest rating of three stars, were rewarded by being publicly celebrated for being 'high performing' and granted 'earned autonomy' (Bevan and Hood, 2006a,b).

In the models used for star ratings, the 'key targets' were most important. To justify the claim that star ratings offered a rounded assessment of performance, key targets were supplemented by a wider set (about forty) targets and indicators in a so-called 'balanced scorecard'. Within the star ratings of acute trusts and PCTs, reducing hospital waiting times was of overriding importance; failure to deliver these targets could result in being zero-rated. For acute trusts: six of the nine key targets were for waiting times (the other three were achieving a financial balance, hospital cleanliness, and improving the working lives of staff): and one of the three domains in 'balanced scorecard' was the 'patient focus', which was also dominated by waiting time targets. The star rating for Primary Care Trusts also included three key targets for waiting times. Table 1 gives the targets for waiting for elective admission in England showing how these became more demanding over the five years of star rating.

The application of targets became more explicit as the system developed. In the first year (2000/01) the 18-month target applied at end of March only. In the second year (2001/02) the targets were set were that 'no patients waiting more than 18 months for inpatient treatment' and 'fewer patients waiting more than 15 months for inpatient treatment'. From the third year, failure was defined in terms of the number of breaches and these for each year were as follows

- 2002/03: the sum of the number of patients waiting longer than 15 months at the end of each the first 11 months of 2002/03 plus the number of patients waiting longer than 12 months at the end of March 2003;
- 2003/04: the sum of the number of patients waiting longer than 12 months at the end of each the first 11 months of 2003/04 plus the number of patients waiting longer than 9 months at the end of March 2004;
- 2004/05: the sum of patients waiting more than 9 months at each month from April 2004 to March 2005.

The 'star rating' system succeeded in conveying to those who worked in the NHS that reducing waiting times mattered by 'naming and shaming' those that failed. The evidence from the US is that systems of performance assessment that are designed to inflict reputational damage on poorly performing hospitals have an impact where markets do not (Hibbard, Stockard, and Tusler, 2003, 2005;

Chassin, 2002; Bevan and Hamblin, 2009). Hibbard identified the four requisite characteristics for a system to inflict damage: these are that it be a ranking system, published and widely disseminated, easily understood by the public, and followed up by future reports. The 'star rating' system satisfied all these characteristics (Mannion, Davies, and Marshall, 2005).

In contrast with England, following devolution, the government in Wales initially abandoned targets for waiting times (Hauck and Street, 2007), and when these were introduced from 2001, a report from the Auditor General for Wales (2005, p. 36) observed that, although waiting times were 'an important part of the Welsh Assembly Government's overall health policy. Waiting time targets have been set out in a variety of documents and not always been clearly and consistently articulated or subject to clear and specific timescales'. We rely on that report for understanding of the changing policy in Wales on reducing waiting times.

Targets for waiting times were used in Wales more as an aspiration, in the hope that managers would respond. These targets were adjusted to reflect variations in local circumstances, with some Trusts allowed a number of breaches, which were not publicized, so people on these waiting lists would have been misled to expect treatment within the relevant waiting time target (Auditor General for Wales, 2005, p. 35). The system of reporting performance in Wales from 2003/04 was through targets specified through the Service and Financial Framework (SaFF) but there was confusion over the relative priority of the various SaFF targets (although Trusts perceived financial and waiting time targets to be more important than others); which was exacerbated by the large number of targets Trusts were expected to achieve (104 in 2003-04, although these were reduced to 40 in the following year) (Auditor General for Wales, 2005, p. 39). There was a website that indicated to the public likely waiting times by specialty, hospital, and specialist (Health of Wales Information Service, 2006), but there was no equivalent system to star ratings in Wales. There was no ranking system, no attempt to inform the public about hospitals' performance against targets through regular reports. Whereas in England, the governments' response to the problem of long waiting times was to set ambitious targets, in Wales, targets were set to reflect existing poor performance. This is illustrated in Table 2, which gives the targets in place in 2005, the final year of star ratings in England. The Auditor General for Wales (2005, p. 17) also commented on the contrast between the ambitious target set in England for a pathway-based maximum waiting time of 18 weeks from GP referral to treatment, to be achieved by 2008, whereas the Welsh Assembly Government had 'no similarly clear strategy outlining how it intends to reduce target waiting times over the medium term'.[5]

In addition to the reforms described here that affected the operation of the NHS, Figure 1 shows

[5]A complication in comparing performance in England and Wales is that from 1 April 2004, the government in Wales introduced the 'second offer scheme' for patients on the inpatient and day case waiting list if they had waited, or were likely to wait, over 18 months, or would breach the specific targets for particular treatments. This scheme paid for such patients to be treated at alternative providers (private hospitals in Wales or hospitals in England) at no charge to the hospital for these patients as the costs were paid from made central funds. This scheme was extended in June 2004, so that, by March 2005, it would guarantee an offer of treatment by an alternative provider for those waiting over twelve months (Auditor General for Wales, 2005, p. 9).

the substantial trend increases in funding (£s per capita) for both England and Wales were similar over the seven years beginning in 2000-01. The principal difference between the NHS in England and Wales was in the governance regime.

There is now a large theoretical literature looking at why organizations that cohere around a public service motive may be different from standard private organizations run to maximize profit. Here is not the place to review that literature in detail. However, it is useful to outline how some of the ideas in that literature affect the interpretation of the results developed here.

A key difficulty in achieving accountability and improving incentives in public services is the difficulty of measuring the 'quality' of the output in a relevant sense. Public services generally run on the basis of some kind of non-profit mission as discussed in Besley and Ghatak (2003) where mission is defined by Wilson (1989, p. 95), as a culture 'that is widely shared and warmly endorsed by operators and managers alike'. This measurement problem leads government to develop broad measurable indicators which are then used to regulate the performance of public service providers. Some aspects of accountability can then be tied directly to such measurable indicators.

Since Baker (1992) and Holmstrom and Milgrom (1991), it has been appreciated in the theoretical literature that care needs to be taken in using imperfect performance indicators to regulate the operation of organizations. Using high-powered incentives for observable performance can be problematic in this context. Even if you get more of what you are rewarding, as you would expect, it is essential that this does not come at the expense of poorer performance on other, harder-to-measure, dimensions. This effort diversion is frequently referred to as 'gaming' in the literature on public sector performance : Smith (2005) offers a typology; its problematic existence has been recognised in empirical studies with financial incentives (see for example, Heinrich, 2002; Doran, Fullwood, Reeves, Gravelle, and Roland, 2008; Burgess and Ratto, 2003); and Bevan and Hood (2006a,b) and Bevan and Hamblin (2009) have shown how 'naming and shaming' also resulted in gaming.

3 Data and Methodology

This section discusses the data and the way in which we use these to construct a test for the impact of waiting time targets on the length of waiting times.

3.1 Data

We obtained data on the distribution of waiting times for each NHS trust in Wales and England.[6] The data is a snapshot of the hospitals' waiting lists on the last day of each financial quarter of the NHS. The length of the waiting time is classified in 7 different 3-month bands ('waiting between 0 and 3', 'between 3 and 6 months' etc. with the highest being 'waiting more than 18 months') and

[6]The data can be downloaded at www.performance.doh.gov.uk/waitingtimes/index.htm and www.statswales.wales.gov.uk/ReportFolders/ReportFolders.aspx. We accounted for mergers by summing the data for the merged hospitals prior to the merger.

our data consists of the number of patients waiting in each of those bands.[7] It covers 28 quarters in the period from the first quarter of the financial year 1999/2000, corresponding to end of June 1999, to the last quarter of the financial year 2005/2006, corresponding to end of March 2006.

Figure 2 presents the type of data we have for one specific NHS trust. It shows for every of the 28 quarters covered the 7 data-points in our dataset.

The waiting list statistics are patients waiting to be admitted either as a day case or ordinary admission. The principal difference in definitions between Wales and England is that, in Wales, all referrals are included whatever the source, whereas in England, only referrals from medical and dental general practitioners are included (Auditor General for Wales, 2005, pp. 50-53).

To get a feel for what the data show, figures 3-9 present the sum of patients waiting per region for each of the waiting bands. Table 3 presents the mean and median of the number of patients waiting per trust in each waiting band for the 9 regions in our data, i.e. Wales and the 8 English regions. It is evident from these figures that waiting lists fell in line with the targets and that the gap with Wales opened over the period, suggesting that the targets did have an impact on hospital policy in England.

3.2 Methodology

To evaluate the effect of the English regime of 'naming and shaming' for failure to achieve targets for waiting times for hospital admission we make use of the fact that around the time when this regime was introduced in England, the targets which were introduced into the NHS in Wales were without a regime of 'naming and shaming'. In section 4 we will come back to this when we assess the robustness of our main results. We hence believe the Welsh NHS to be a suitable control group for evaluating the 'treatment' of the English NHS. The effect of each target can then be identified by running for each waiting band $w = 0, 3, 6, 9, 12, 15, 18$ a simple difference-in-difference specification of the form

$$y_{itw} = \beta_w \cdot \text{target}_{itw} + \delta_{wi} + \gamma_{wt} + \eta_w \cdot t_{\text{Wales}} + \epsilon_{itw} \tag{1}$$

where γ_{wt} is a set of time dummies, δ_{wi} a set of NHS trust dummies and target_{itw} a dummy being 1 if hospital i is in a region where at time t a target for waiting category w existed and y_{itw} is the number of patients waiting in band w in hospital trust i at time t. To allow for a potentially different underlying trend in the Welsh waiting lists over the period studied and exploit only the discontinuity of the targets we include a linear Welsh time trend t_{Wales} in (1).[8]

[7] For example, in any hospital we have data on the number of patients waiting between 9 and 12 months on the last day of any financial quarter.

[8] As a robustness check, we will also consider the Welsh targets after the introduction of the second offer scheme in March 2004 as treatment (see table 7). All of our main results are robust to this different specification.

4 Results

4.1 Core Results

Focusing first on the effect of the waiting time targets on the targeted waiting category, we present results from the specification in equation (1). The raw data show that there were negligible numbers of patients in the English NHS waiting longer than each target, i.e. more than 18, 15, 12 and 9 month, when the respective targets came into force. Table 4 presents the coefficient estimates of the effect of each of these targets in the English NHS on the number of patients waiting over 18 months, between 15 and 18 months, between 12 and 15 months and between 9 and 12 months, respectively. They are all negative, significant and large in magnitude. In an English hospital with a median number of patients waiting in June 1999, the first three targets are estimated to have reduced the numbers of patients waiting longer than the targeted time to zero. The 9-month target is estimated to have reduced the number of patients waiting between 9 and 12 months in an English NHS trust by 67%, again compared to the median number of patients waiting in June 1999.

Table 5 presents similar regressions to table 4, but previous treatments are now included. This allows us to evaluate how progress towards achieving targets was made already before the target came into force. Column (4) suggests that the number of patients waiting between 9 and 12 months in the English NHS decreased already significantly at the times when the 12 and 15-month targets were enacted, so in the two years before the 9-month target actually came into force. Including this early treatment effect, the 9-month target's estimated effect is to have achieved that no patients were waiting more than 9 months in a median English NHS trust. The early treatment effect in anticipation of the announced target can as well be observed for the 12-month target. Column (3) shows that the number waiting longer than 12 months already dropped in the two years prior to the 12-month target coming into force. However, for the earlier 15-month target no significant prior drop in the waiting list is estimated.

This finding makes sense in a world of increasingly demanding targets. The 18-month target had been in place for the NHS in England since 1995 (NHS Executive, 1995). What was new about the regime was that sanctions applied for failure to meet that target in 2001. Experience of hitting (or missing) that target would likely have made it clear that systemic changes would be necessary to continue to meet targets in future.

4.2 Evidence of Gaming of patients on waiting lists?

The results presented suggest that the targets were effective in reducing long waits. They suggest as well that hospitals early on managed their waiting lists to fulfil the later targets. At least in the beginning, hospital managers have tried to game targets by shuffling patients across different categories of waiting times, too. The results of table 5 provide evidence for this. They show that the reduction in long waits came at the expense of an increase in the numbers of patients waiting

for shorter time periods. In particular, columns (5)-(7) show that the number of patients waiting between 3 and 9 months significantly increased both after the introduction of the 15- and the 18-month target. For the interpretation of these numbers it is important to recall that we use census rather than discharge data. If the hospitals had reacted to the targeting regime by treating additional patients once they waited for 9 months, and who would have waited even longer prior to the targeting regime, the numbers of patient waiting less than 9 months should not change. The increase in the waits between 3 and 9 months hence shows that patients who, in the absence of the targeting regime, would have waited 0 to 6 months, were now left waiting until their waiting time approached the maximum allowed.

Again the later targets, i.e. the 12- and the 9-month target, did not have this effect. Their coefficient estimates are not significantly positive in columns (5)-(7). But they are generally non-negative and never significantly negative. This indicates that while the number of close-to-9-month waits did not increase further, the hospitals were not able (or had no incentive) to cut back the previously increased level of close-to-9-month waits either. Further, the estimated overall effect of the targeting regime, measured by the sum of the effects of the four targets, is to have increased the number of patients waiting in all three categories below 9 months waiting time. A t-test for this sum is significant least at the 5% level for all three categories.

The exception to this rule is the effect of the 9-month target on the number of patients waiting between 6 and 9 months. However, considering the early treatment effects outlined above, this might well be driven by the later to be introduced 6-month target.[9]

Taken together the results of table 5 seem to suggest that the targets were effective in reducing long waits, but this was done, at least in part, not by treating more patients, but by prioritising the treatment of patients waiting for a long time and leaving other patients, who were only waiting for a short time, wait longer.

One possible metric to evaluate the overall effect of the targets would be its effect on the mean-waiting time. Table 7 presents different specifications of how the mean-waiting time changed with the introduction of the four targets under study. The table shows that mean waiting times did indeed decrease after the introduction of the 15- and 12-month targets. However, they may have increased after the introduction of the 9-month target.

In order to get a feel for the magnitude of the gaming one might ask what the effect of the targets on the mean waiting time would have been, had there not been an increase in the number of patients waiting e.g. up to 9 months rather than being treated within 6 or 3 months. We construct this hypothetical mean waiting time from table 5. Specifically, we calculate this as the mean-waiting time implied by the predicted values of the regressions in table 5 after having set the coefficients in columns (5) through (7) to zero. This reflects the thought-experiment that targets had no effect on the distribution of below-9-month waits. The result of this comparison is summarized in figure

[9]This was to be achieved by December 2005 under the new regime of 'annual healthchecks', the successor to star ratings.

10, which plots the actual mean waiting time in England alongside the hypothetical mean waiting time.[10] This calculation suggests in the first two years of the star rating regime (quarters 8 to 16), the mean-waiting time would have been up to one month shorter had there not been any change in waiting at other time lengths; this fell to six months in the third year (quarters 16 to 20), and subsequently to zero in the fourth year. This does suggest that there was initially some gaming of the targets, which had a material impact on average waiting times, but this declined and ceased altogether by the fourth year.

4.3 Did the targets have other detrimental effects?

While this exercise is constructive, it does not get at wider possibilities for redeploying resources to meet the targets which had detrimental effects on patient care – this would be the classic multi-tasking behavioural response. There is, however, little evidence that the shorter waiting times in England were offset by detrimental effects on performance in other dimensions of quality of care for which we have data. Hauck and Street (2007) report results of a detailed analysis of four hospitals (three in England and one in Wales) which were close to the border and serving similar populations over the period from 1997/98 to 2002/03. They found that, in the English hospitals, there were increases in activity and low or declining mortality rates; but, in the Welsh hospital, there was no increased activity and high and rising mortality rates. Figures 11 and 12 gives national comparisons for hospital mortality and activity (Finished Consultant Episodes) that include the period of star ratings (2001-02 to 2005-06). Over that period, Figure 11 shows that, in England, there was a steady trend of declining mortality; and, in Wales, no such trend and no reduction. (This statistic is a crude indicator as it does not account for deaths outside hospital.) There are good reasons why in some cases (e.g., cancer waiting times) that shorter waiting times will reduce mortality. Figure 11 shows that, in England, there was a steady trend of increasing activity; and, in Wales, no such trend and no increase.

This finding is in line with the findings of Leatherman and Sutherland (2003) who examined cross-country comparisons of all available indicators of quality of care in the two countries and found that Wales performed worse on most of these (for example, higher mortality rates from: causes considered amenable to healthcare, coronary heart disease, stroke and diabetes). An audit by the Royal College of Physicians (2006) highlighted the much worse organisation of stroke care in Wales, which meant that patients treated in Wales were more likely die from stroke, or if they survived would have higher levels of disability than in England or Northern Ireland (Royal College of Physicians, 2006).

[10] Since the hypothetical mean waiting time is calculated from the predicted values of the regressions in table 5, we present as well the mean waiting time calculated from the predicted values without setting any coefficients to zero. This follows the actual mean waiting time closely.

5 Conclusion

This paper has exploited a natural experiment between two regimes for hospital waiting time targets in which failure resulted in 'naming and shaming' in England was perceived to be rewarded in Wales. Using Wales as a control group, we found that 'naming and shaming' did reduce the time that patients waited. In fact, such waiting has all but been eliminated by the use of targets combined with real sanctions for hospital chief executives. Given that the identification proposed here is quite clean, it is reasonable to argue that what we have found is a behavioural effect at the hospital level. It shows that targets with sanctions – part of the 'naming and shaming' regime that has been used in recent years to improve public services in England – has had an impact: that increased funding together with 'naming and shaming' meant that the performance of the NHS in England (as measured by waiting times) was transformed; and the absence of 'naming and shaming' meant that no similar transformation took place in Wales. And that providers in Wales were able to use the extra funding to extract provider rents, particularly as we were unable to find any hard evidence of changes in England in response to the regime there were detrimental to patient welfare as compared to Wales.

Propper, Sutton, Whitnall, and Windmeijer (2008b), in their comparison of England and Scotland, found that the target regime in England led to a shift in the distribution of waiting times towards meeting targets and reductions in the numbers of patients with long waiting times in the tail of the distribution; in contrast, in Scotland, the distribution of waiting times shifted in the opposite direction increasing the number of longer waits and reducing the number that waited below the target set for England. They found little evidence of gaming of patients on waiting lists in England: no evidence of re-ordering of patients on lists to meet targets; some evidence of waiting list manipulation (patients were removed, temporarily and permanently, from waiting lists but no evidence that this manipulation was harmful to the health of patients). Their analysis of evidence of patient quality suggests that this improved in England relative to Scotland (for three measures of mortality, 30 day mortality rates for all admissions, all emergency admissions and emergency admissions for acute myocardial infarction).

Our paper began by emphasising the sparsity of good empirical evidence to evaluate the industry of performance measurement; the disappointing findings of much of this empirical literature, that the whilst this industry creates jobs for those involved in the production of measures of performance, it seems to have little effect on those whose performance is measured; and theoretical reasons and empirical evidence of gaming when measures of performance are linked to high-powered incentives. Our analyses, and those by Propper, Sutton, Whitnall, and Windmeijer (2008a,b), of the natural experiment of the target regime in England in comparison with Wales and Scotland is vital empirical evidence of the effectiveness of linking performance measurement with high-powered incentives in the form of 'naming and shaming'. Like Propper et al we have found little evidence of gaming using the routinely available data. So why has the harsh regime of 'star rating', that was introduced in England only, transformed performance in England relative to Wales and Scotland with so little evidence of

adverse outcomes? We agree with the explanations suggested by Propper et al: only limited measures of quality are available; as waiting times were widely recognised as a weakness of the NHS, this could become accepted as a mission by its employees; and the target regime was a long-term commitment to reduce waiting times requiring systemic change that could be implemented because of the massive increases in funding. We emphasise a further explanation: that waiting time targets were a crucial but not the only part of performance measurement by 'star ratings'. Other indicators and results of inspections determined 'star ratings' and thus would curb gaming by reducing effort on these other measured elements of quality of care. These indicators included, for example, 30-day hospital mortality rates (which included deaths outside hospital), hospital re-admission rates, and results from surveys of patients and staff. Results of inspections included hospital cleanliness and those by the Commission for Health Improvement into implementing systems of governance to improve and assure the quality of care: as these were based on visits they stood a good chance of discovering flagrant diversion of effort away from quality of care to meet waiting time targets (Bevan and Cornwell, 2006). Nevertheless, one of the weaknesses of 'star ratings' was the absence of a systematic audit to counter gaming (Bevan and Hood, 2006a,b; Bevan and Hamblin, 2009).

References

Alvarez-Rosete, A., G. Bevan, N. Mays, and J. Dixon (2005): "Effect of diverging policy across the NHS," *BMJ*, 331(7522), 946–950.

Auditor General for Wales (2005): *NHS waiting times in Wales. Volume 1 - The Scale of the problem*. The Stationery Office, Cardiff.

Baker, G. P. (1992): "Incentive Contracts and Performance Measurement," *Journal of Political Economy*, 100(3), 598–614.

Besley, T., and M. Ghatak (2003): "Incentives, Choice, and Accountability in the Provision of Public Services," *Oxford Review of Economic Policy*, 19(2), 235–249.

Bevan, G. (2006): "Setting Targets for Health Care Performance: Lessons from a Case Study of the English NHS," *National Institute Economic Review*, 197(1), 67–a–79.

Bevan, G., and J. Cornwell (2006): "Structure and logic of regulation and governance of quality of health care: was OFSTED a model for the Commission for Health Improvement?," *Health Economics, Policy and Law*, 1(4), 343–370.

Bevan, G., and R. Hamblin (2009): "Hitting and missing targets by ambulance services for emergency calls: impacts of different systems of performance measurement within the UK," *Journal of the Royal Statistical Society*, 172(1), 1–30.

Bevan, G., and C. Hood (2006a): "Have targets improved performance in the English NHS?," *British Medical Journal*, 332(7538), 419–422.

——— (2006b): "What's measured is what matters: targets and gaming in the English public health care system," *Public Administration*, 84(3), 517–538.

Bevan, G., and R. Robinson (2005): "The Interplay between Economic and Political Logics: Path Dependency in Health Care in England," *Journal of Health Politics Policy and Law*, 30(1-2), 53–78.

Beverley, C., and J. Haynes (2005): *Franchised Trusts*. Health Management Specialist Library: Management Briefing, NeLH Health Management Specialist Library.

Burgess, S., and M. Ratto (2003): "The Role of Incentives in the Public Sector: Issues and Evidence," *Oxford Review of Economic Policy*, 19, 285–300.

Campbell, S., D. Reeves, E. Kontopantelis, E. Middleton, B. Sibbald, and M. Roland (2007): "Quality of Primary Care in England with the Introduction of Pay for Performance," *New England Journal of Medicine*, 357(2), 181–190.

Chassin, M. R. (2002): "Achieving And Sustaining Improved Quality: Lessons From New York State And Cardiac Surgery," *Health Affairs*, 21(4), 40–51.

Commission for Health Improvement (2003a): *NHS Performance Ratings. Acute Organisations, Specialist Organisations, Ambulance Organisations 2002/03*. The Stationery Office, London.

——— (2003b): *NHS Performance Ratings. Primary Care Organisations, Mental Health Organisations, Learning Disability Organisations 2002/03*. The Stationery Office, London.

Department of Health (2001): *NHS performance ratings acute trusts 2000/01*. Department of Health, London.

——— (2002): *NHS performance ratings acute trusts, specialist trusts, ambulance trusts, mental health trusts 2001/02*. Department of Health, London.

Doran, T., C. Fullwood, D. Reeves, H. Gravelle, and M. Roland (2008): "Exclusion of Patients from Pay-for-Performance Targets by English Physicians," *New England Journal of Medicine*, 359(3), 274–284.

Enthoven, A. C. (1985): *Reflections on the Management of the NHS*. Nuffield Provincial Hospitals Trust, London.

——— (1999): *In Pursuit of an Improving National Health Service*. Nuffield Trust, London.

Fung, A., and D. ORourke (2000): "Reinventing Environmental Regulation from the Grassroots Up: Explaining and Expanding the Success of the Toxics Release Inventory," *Environmental Management*, 25(2), 115–127.

Fung, C. H., Y.-W. Lim, S. Mattke, C. Damberg, and P. G. Shekelle (2008): "Systematic Review: The Evidence That Publishing Patient Care Performance Data Improves Quality of Care," *Annals of Internal Medicine*, 148(2), 111–123.

Hauck, K., and A. Street (2007): "Do targets matter? A comparison of English and Welsh National Health priorities," *Health Economics*, 16(3), 275–290.

Health of Wales Information Service (2006): *Waiting Times Information*.

Healthcare Commission (2004): *2004 Performance Rating*. The Stationery Office, London.

——— (2005): *NHS performance ratings 2004/2005*. Healthcare Commission, London.

Heckman, J., C. Heinrich, and J. Smith (1997): "Assessing the Performance of Performance Standards in Public Bureaucracies," *The American Economic Review*, 87(2), 389–395.

Heinrich, C. J. (2002): "Outcomes-based Performance Management in the Public Sector: Implications for Government Accountability and Effectiveness," *Public Administration Review*, 62(6), 712–725.

Hibbard, J. H. (2008): "What Can We Say about the Impact of Public Reporting? Inconsistent Execution Yields Variable Results," *Annals of Internal Medicine*, 148(2), 160–161.

Hibbard, J. H., J. Stockard, and M. Tusler (2003): "Does Publicizing Hospital Performance Stimulate Quality Improvement Efforts?," *Health Affairs*, 22(2), 84–94.

——— (2005): "Hospital Performance Reports: Impact On Quality, Market Share, And Reputation," *Health Affairs*, 24(4), 1150–1160.

HM Treasury (2004): *Public Expenditure Statistical Analyses 2004*. The Stationery Office, London.

——— (2006): *Public Expenditure Statistical Analyses 2006*. The Stationery Office, London.

Holmstrom, B., and P. Milgrom (1991): "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, & Organization*, 7, 24–52.

Klein, R. E. (2006): *The New Politics of the National Health Service (5th ed)*. Radcliffe Press, Oxford.

Le Grand, J. (2003): *Motivation, agency and public policy: Of knights and knaves, pawns and queens*. Oxford University Press, Oxford.

Leatherman, S., and K. Sutherland (2003): *The Quest for Quality in the NHS: A mid-term evaluation of the ten-year quality agenda*. The Stationery Office, London.

Mannion, R., H. Davies, and M. Marshall (2005): *Cultures for Performance in Health Care*. McGraw Hill, Maidenhead.

Marshall, M. N., P. G. Shekelle, H. T. O. Davies, and P. C. Smith (2003): "Public Reporting On Quality In The United States And The United Kingdom," *Health Affairs*, 22(3), 134–148.

NHS England (2000-2006): *Hospital Episodes Statistics - Headlines*. The Health and Social Care Information Centre, Leeds.

NHS Executive (1995): "Revised and expanded Patients Charter: implementation," *Health Service Guidelines HSG(95)13*.

NHS Wales (2000-2006): *PEDW Headline Figures (Old Definition)*.

Pawson, R. (2002): "Evidence and Policy and Naming and Shaming," *Policy Studies*, 23(3), 211 – 230.

Perrin, B. (2002): *Implementing the Vision: Addressing Challenges to Results-Focused Management and Budgeting*. OECD, Paris.

Propper, C., M. Sutton, C. Whitnall, and F. Windmeijer (2008a): "Did 'Targets and Terror' Reduce Waiting Times in England for Hospital Care?," *The B.E. Journal of Economic Analysis & Policy*, 8(2 (Contributions)).

——— (2008b): "Incentives and Targets in Hospital Care: Evidence from a Natural Experiment," The Centre for Market and Public Organisation 08/205, Department of Economics, University of Bristol, UK.

Rosenthal, M. B., and R. G. Frank (2006): "What Is the Empirical Basis for Paying for Quality in Health Care?," *Medical Care Research Review*, 63(2), 135–157.

Royal College of Physicians (2006): *National Sentinel Stroke Audit*. Royal College of Physicians, London.

Secretaries of State for Health, Wales, Northern Ireland and Scotland (1989): *Working for patients [CM 555]*. HMSO, London.

Secretary of State for Health (2000): *The NHS plan [CM 4818-I]*. The Stationery Office, London.

Smee, C. (2005): *Speaking Truth to Power: Two Decades of Analysis in the Department of Health*. Radcliffe Press, Oxford.

Smith, P. C. (2005): "Performance Measurement in Health Care: History, Challenges and Prospects," *Public Money & Management*, 25(4), 213–220.

Tuohy, C. (1999): *Accidental Logics. The Dynamics of Change in the Health Care Arena in the United States, Britain and Canada*. Oxford University Press, New York.

van Dooren, W., and S. V. de Walle (eds.) (2009): *Performance Information in the Public Sector. How it is used.* Palgrave Macmillan, Basingstoke.

Wilson, J. Q. (1989): *Bureaucracy: What Government Agencies Do and Why They Do It*. Basic Books, New York.

Appendix A

Figure 1

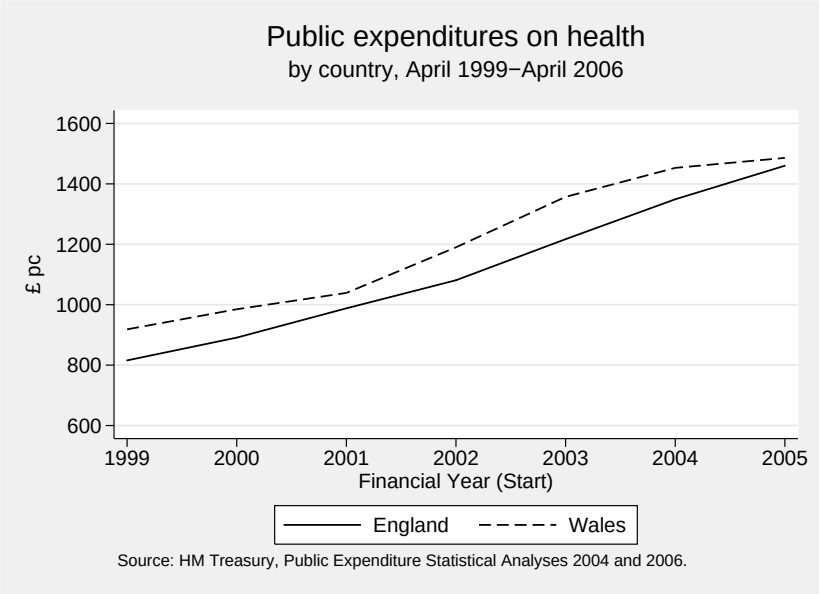


Figure 2

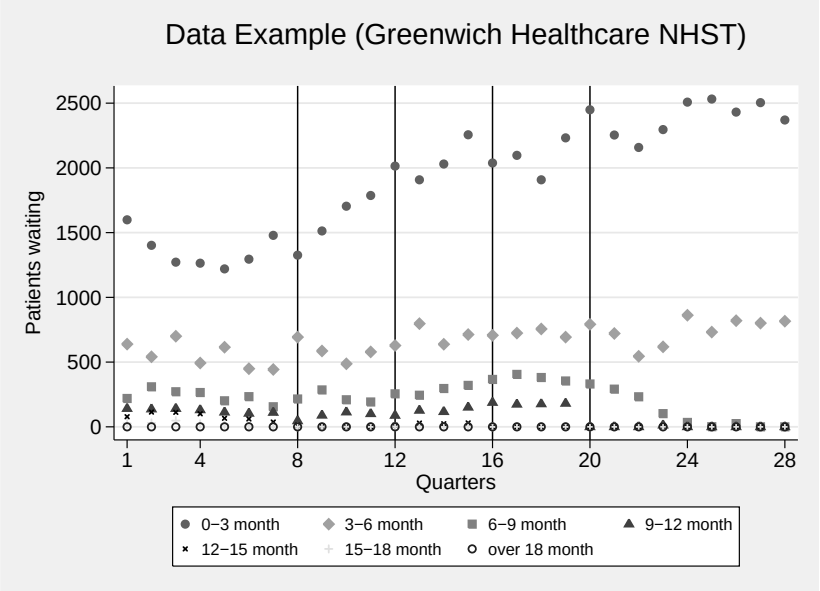
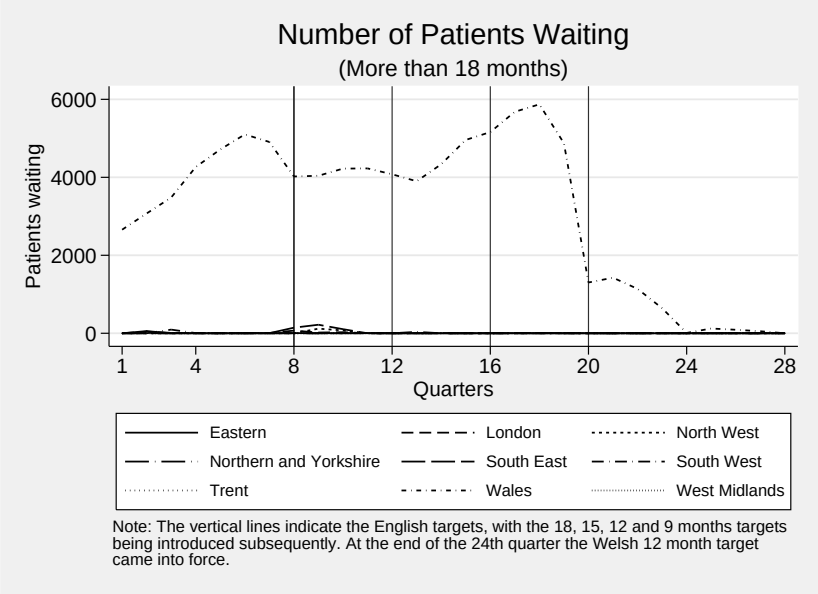
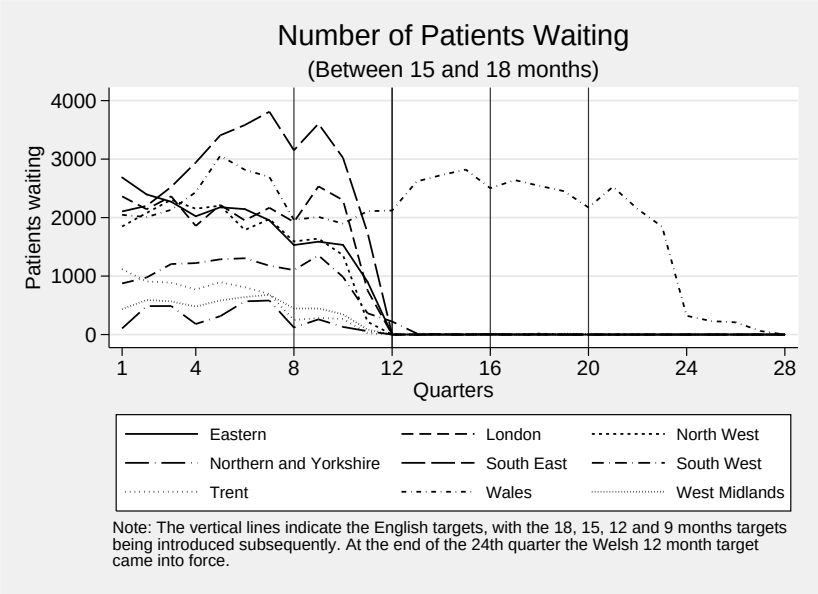


Figure 3



Note: The vertical lines indicate the English targets, with the 18, 15, 12 and 9 months targets being introduced subsequently. At the end of the 24th quarter the Welsh 12 month target came into force.

Figure 4



Note: The vertical lines indicate the English targets, with the 18, 15, 12 and 9 months targets being introduced subsequently. At the end of the 24th quarter the Welsh 12 month target came into force.

Figure 5

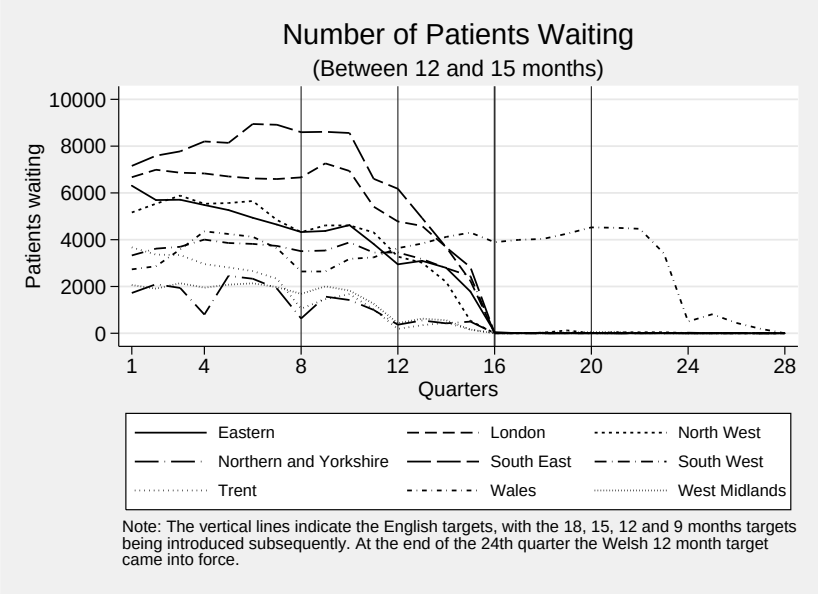


Figure 6

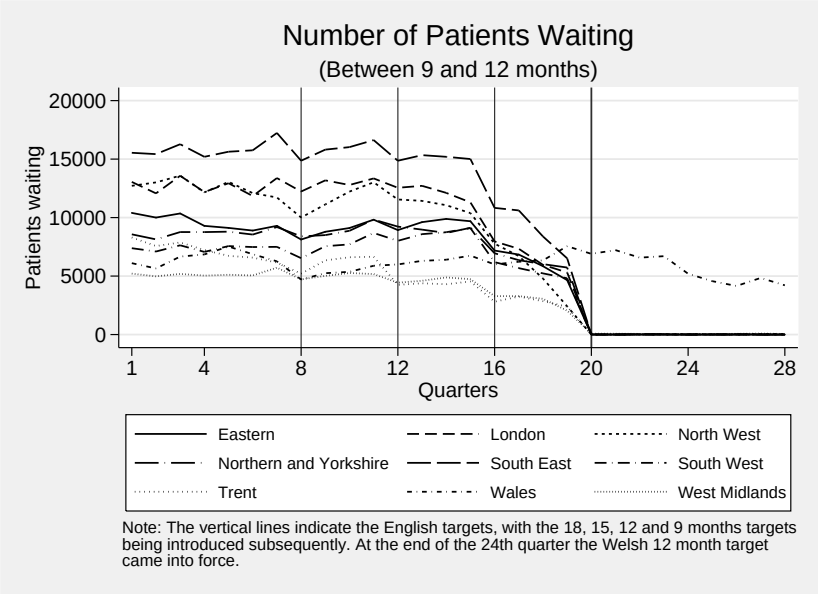


Figure 7

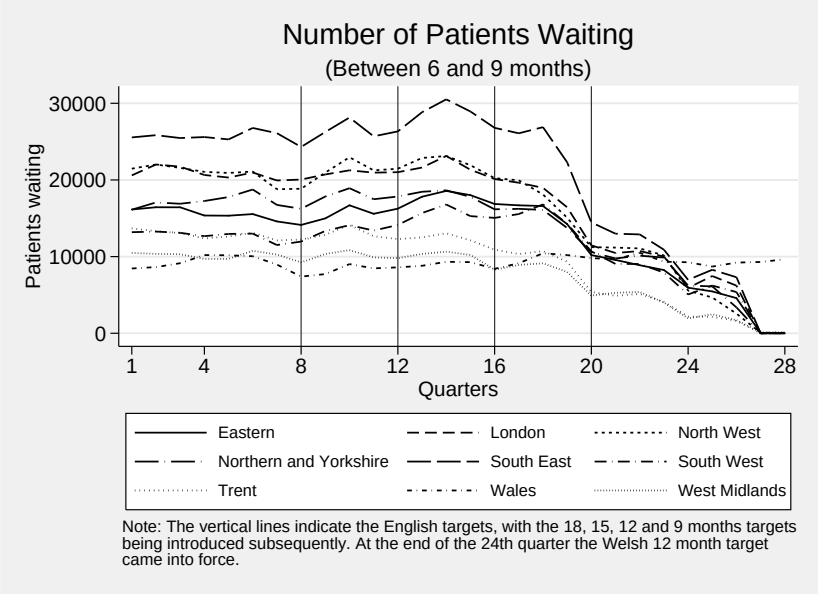


Figure 8

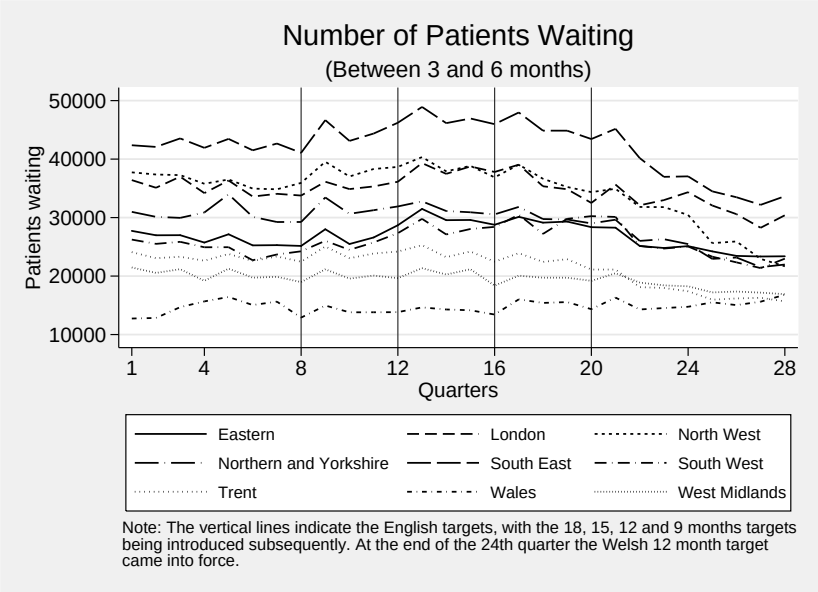


Figure 9

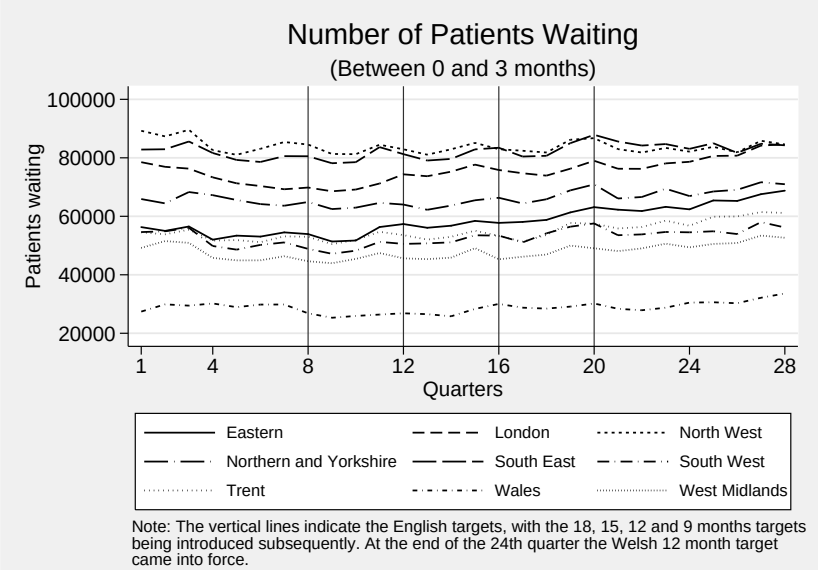


Figure 10

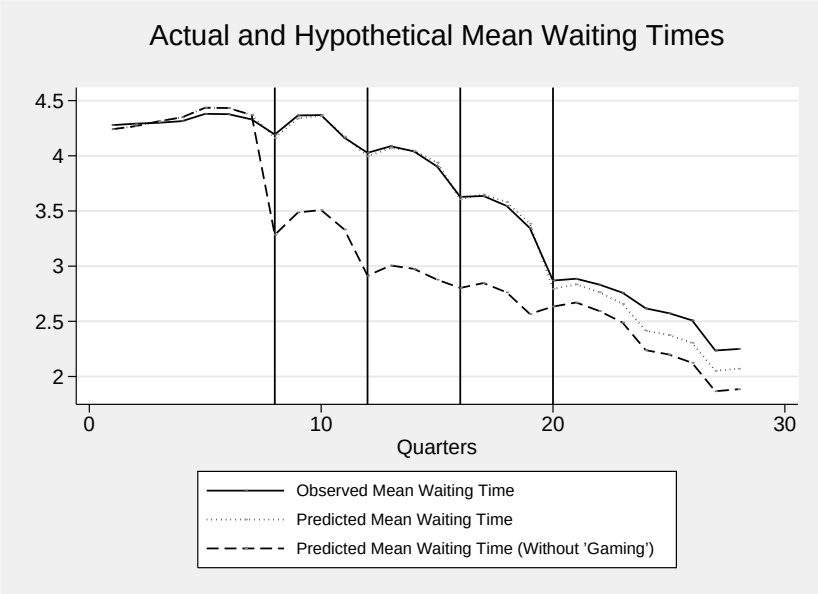


Figure 11

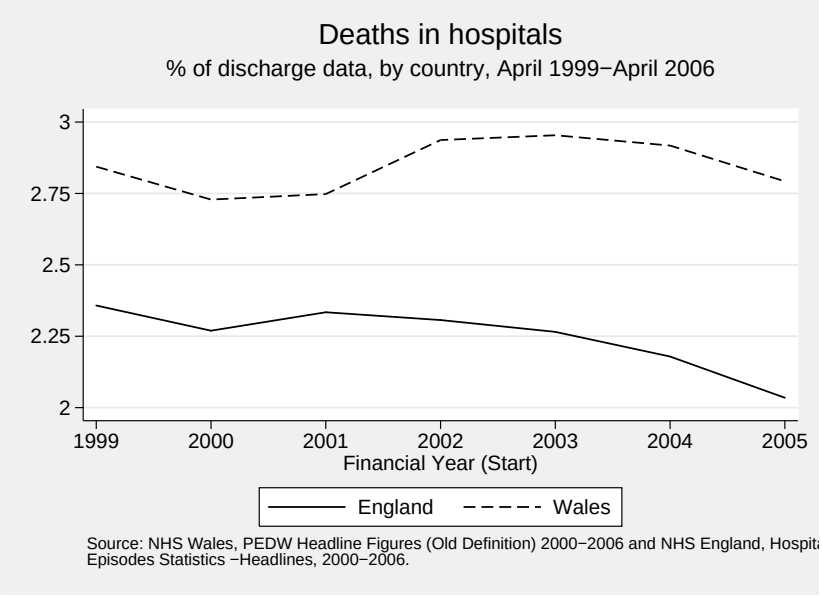
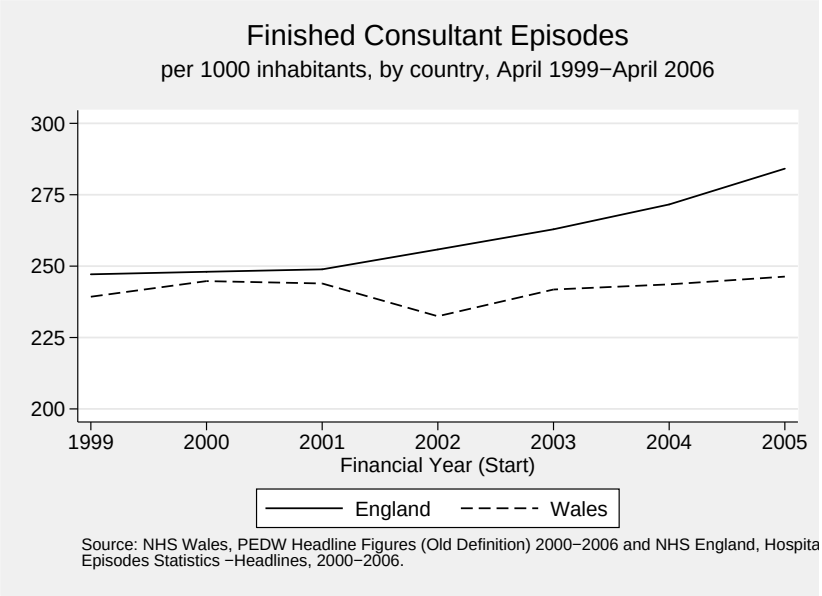


Figure 12



Appendix B

Table 1
Targets for waiting for elective admission in England

Year	*at start of year (months)*	*at end of year (months)*
2000/01	18	18
2001/02	18	15
2002/03	15	12
2003/04	12	9
2004/05	9	9

Sources: Department of Health (2001, 2002); Commission for Health Improvement (2003a), Healthcare Commission (2004, 2005).

Table 2
Waiting time targets for England and Wales in 2005

Type of waiting	*England (weeks)*	*Wales (weeks)*
For first outpatient appointment	13	78
For inpatient / day case treatment	26	78

Source: Auditor General for Wales (2005a: 15).

Table 3
Descriptive statistics (end of June 1999)

	trusts	*mean (median) number of patients waiting per trust*						
		<3m	3m-6m	6m-9m	9m-12m	12m-15m	15m-18m	>18m
England	159	3340	1554	863	510	227	73	0
		(3156)	(1466)	(796)	(406)	(153)	(33)	(0)
Eastern	18	3128	1540	897	578	351	149	0
		(2856)	(1493)	(796)	(468)	(252)	(74)	(0)
London	28	2804	1300	736	466	238	84	0
		(2962)	(1291)	(648)	(369)	(151)	(67)	(0)
North West	22	4057	1715	976	578	235	84	0
		(3590)	(1676)	(939)	(574)	(161)	(45)	(0)
North. and Yorkshire	17	3877	1822	948	504	101	6	0
		(3372)	(1699)	(1006)	(395)	(38)	(0)	(0)
South East	25	3313	1695	1022	622	286	84	0
		(3112)	(1563)	(1049)	(591)	(240)	(54)	(0)
South West	18	3034	1458	733	410	185	49	0
		(2694)	(1468)	(803)	(272)	(107)	(18)	(0)
Trent	14	3895	1721	978	593	262	80	0
		(3376)	(1457)	(925)	(539)	(134)	(42)	(0)
West Midlands	17	2894	1263	617	306	121	26	0
		(2666)	(1258)	(620)	(237)	(69)	(14)	(0)
Wales	12	2284	1062	704	509	228	171	221
		(1862)	(855)	(533)	(387)	(166)	(87)	(47)

Table 4
Estimates of effects of targets

	patients waiting			
	(1) over 18m	(2) 15m to 18m	(3) 12m to 15m	(4) 9m to 12m
18 month	-192.0*			
	(92.6)			
15 month		-191.4**		
		(45.0)		
12 month			-193.9**	
			(20.5)	
9 month				-273.2**
				(36.1)
N	4682	4682	4682	4682
R^2	0.76	0.66	0.65	0.76

Notes: * significant at 5%; ** significant at 1%; standard errors in parentheses are clustered at trust level; all regressions include trust and time dummies and a Wales-specific time trend.

Table 5
Targets and Past Targets

	patients waiting						
	(1) over 18m	(2) 15m to 18m	(3) 12m to 15m	(4) 9m to 12m	(5) 6m to 9m	(6) 3m to 6m	(7) below 3m
18 month	-192.0*	-55.5	-110.7**	29.1	369.0**	399.7**	300.7*
	(92.6)	(28.8)	(42.3)	(48.2)	(41.2)	(54.5)	(152.1)
15 month		-177.1**	-277.4**	-148.5**	187.2**	302.5**	142.8
		(38.7)	(46.9)	(39.0)	(31.5)	(51.0)	(82.3)
12 month			-217.6**	-252.4**	30.4	108.7*	-14.0
			(27.7)	(32.2)	(35.4)	(48.9)	(83.7)
9 month				-284.9**	-225.3**	91.7	233.2*
				(44.3)	(53.6)	(77.5)	(102.0)
N	4682	4682	4682	4682	4682	4682	4682
R^2	0.76	0.66	0.66	0.76	0.84	0.91	0.96

Notes: * significant at 5%; ** significant at 1%; standard errors in parentheses, the standard errors are clustered at trust level; all regressions include trust and time dummies and a Wales-specific time trend.

Table 6
Estimates of effects of targets on mean waiting times

	mean waiting time		
	(1)	(2)	(3)
18 month	-0.037	-0.037	0.176
	(0.255)	(0.165)	(0.259)
15 month	-0.738**	-0.738**	-0.583**
	(0.194)	(0.156)	(0.149)
12 month	-0.886**	-0.886**	-0.731**
	(0.162)	(0.162)	(0.144)
9 month	0.377	0.377*	0.629**
	(0.193)	(0.188)	(0.209)
Wales t-trend	Yes	Yes	No
s.e.	Cluster(trust)	Robust	Cluster(trust)
R^2	0.85	0.85	0.85
N	4682	4682	4682

Notes: Standard errors in parentheses, * significant at 5%; ** significant at 1%; all regressions include trust and time dummies.

Table 7
Wales considered treatment for $t > 24$

	patients waiting						
	(1) over 18m	(2) 15m to 18m	(3) 12m to 15m	(4) 9m to 12m	(5) 6m to 9m	(6) 3m to 6m	(7) below 3m
18 month	-147.1*	-12.7	15.4	103.7**	213.2**	171.9**	181.8
	(63.3)	(16.4)	(22.7)	(28.9)	(28.6)	(38.0)	(111.6)
15 month		-130.8**	-181.7**	-91.0**	79.6**	148.4**	61.9
		(25.2)	(31.0)	(25.2)	(20.5)	(36.4)	(50.0)
12 month			-131.9**	-202.6**	-91.2**	-73.3	-108.5
			(16.2)	(19.8)	(25.6)	(39.0)	(72.8)
9 month				-289.2**	-274.1**	6.3	190.4*
				(41.3)	(51.6)	(71.4)	(84.8)
N	4682	4682	4682	4682	4682	4682	4682
R^2	0.77	0.68	0.67	0.76	0.84	0.91	0.96

Notes: * significant at 5%; ** significant at 1%; standard errors in parentheses, the standard errors are clustered at trust level; all regressions include trust and time dummies and a Wales-specific time trend.