# UNIVERSITY *of* York

This is a repository copy of *The long-run effects of diagnosis related group payment on hospital lengths of stay in a publicly funded health care system:Evidence from 15 years of micro data*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/184509/

Version: Published Version

## White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

RESEARCH ARTICLE

Health Economics WILEY

# The long-run effects of diagnosis related group payment on hospital lengths of stay in a publicly funded health care system: Evidence from 15 years of micro data

**María José Aragón**  |  **Martin Chalkley**  |  **Noémi Kreif**

Centre for Health Economics, University of York, York, UK

**Correspondence**
María José Aragón, Centre for Health Economics, University of York, Heslington, York, YO10 5DD, UK.
Email: mariajose.aragonaragon@york.ac.uk

**Abstract**
Diagnosis Related Group (DRG) payment systems are a common means of paying for hospital services. They reward greater activity and therefore potentially encourage more rapid treatment. This paper uses 15 years of administrative data to examine the impact of a DRG system introduced in England on hospital lengths of stay. We utilize different econometric models, exploiting within and cross jurisdiction variation, to identify policy effects, finding that the reduction of lengths of stay was greater than previously estimated and grew over time. This constitutes new and important evidence of the ability of financing reform to generate substantial and persistent change in healthcare delivery.

**KEYWORDS**
DRG payment, econometric methods, healthcare

**JEL CLASSIFICATION**
C1, H1

## 1 | INTRODUCTION

The adoption of Diagnosis Related Group (DRG) payment systems for hospitals has been one of the most extensive reforms in the financing of healthcare over the last 30 years. Following the development of this system for the United States Medicare in the 1980s it has been adapted and developed across many national healthcare settings (Busse et al., 2011). The central element, which typically replaces either fee-for-service or fixed hospital-level budgets, is the classification of a patient's medical need and the setting of a fixed price that a hospital will receive for meeting the need of each patient within that classification.

A number of motivations for the introduction of such systems have been suggested, such as increased accountability or greater control over expenditure (Forgione et al., 2005; Lægreid & Neby, 2016) but economists have focused on the implied incentives for both the cost and quality of treatment that a hospital will deliver. Compared with fee-for-service payment a fixed price DRG system rewards effort to reduce cost but may reduce the return to high service quality. Compared with a fixed budget there is an incentive to increase activity and economize on costs so as to allow that expansion. Whilst economic theory provides insight into how payment will influence choices, the effect in practice is a matter for empirical investigation and the evaluation of the impact of DRG payment systems has been extensive (Street et al., 2011).

Payment systems are usually introduced across whole systems, not randomly assigned, and the fundamental challenge to empirical investigation is to establish a plausible causal effect in the absence of equivalent treatment and control groups. Approaches have ranged from simple before-and-after comparisons to more sophisticated difference-in-differences regression designs. A consistent characteristic of these investigations has been the limited time period used to measure the effect of payment reform, usually considering just a few years before and after the introduction of a policy. In spite of a substantial investment in this area of research the evidence regarding the real effects of a fundamental and very wide-ranging payment reform remains limited. As proxied by the length of time that patients stay in hospital the consensus is that switching to a DRG system reduces resources used in treatment, but there is little consensus across studies as to how much, and no evidence concerning the durability of effect (Street et al., 2011). There is little evidence of significant effects on service quality (Or & Hakkinen, 2011).

This paper presents the most comprehensive evidence to date on the effect of the adoption of a DRG system on healthcare resources. Our study concerns the adoption of a DRG system that was then called *Payment by Results (PBR)* in the National Health System (NHS) of England starting in 2003. Using an assembled and extensive long run data set comprised of the details of all inpatient treatments delivered in the hospitals of both England and Scotland (which did not adopt the DRG system) over a period of 15 years (covering 6 years before and 9 years after the policy introduction) we subject the data to a suite of empirical methods – difference-in-differences (DiD), synthetic control (SC) and interrupted time series (ITS) – to uncover the effect of the policy on lengths of hospital stay. Whilst Scotland is a small country relative to England, having around one 10 th of the population, the two share a common heritage in terms of their health systems which are almost exclusively publicly funded and rely predominantly upon publicly owned institutions within the context of a NHS. They utilize the same terminology and definitions in respect of hospital services and record data in a similar fashion. Over the study period per capita expenditure on hospital services is similar.

Our findings are of considerable relevance to policy-making in regard to hospital financing. We not only confirm previous evidence that the introduction of a DRG system reduces resource use through shorter hospital treatments, but also find that the effect is at least as substantial as previously reported and *increases over time*. This latter finding is new and indicates that a different financing system might have enduring effects on the management of healthcare resources. The long-term effects we estimate are very substantial and wide-ranging compared to the impact effects that have hitherto been evaluated.

The underlying hypotheses that we examine are first that the policy reform of fixed price payment for hospitals results in reduced lengths of stay, second that this reform takes time to have an effect, and third that the effect may either grow or diminish over time. The first hypothesis is widely suggested by the large literature on the use of DRG systems (summarized in Busse et al., 2011) and follows from the observation that fixed price systems endow hospitals with ownership of the financial impact of efficiency or other cost savings. Since shorter lengths of stay are associated with lower resource use and hence cost they are one mechanism by which hospitals can gain financially – in contrast to systems where they recover their costs but cannot retain any cost savings. The second hypothesis is suggested by a large and very diverse literature concerning policy change and implementation (see Cerna (2013) for a review and summary). Numerous theories of change indicate that organizations, such as hospitals, respond to a policy stimulus through a process of internal negotiation, learning and subsequent implementation, all of which evolve over time. The conclusion of this process is uncertain both in extent and timing. Our focus is on healthcare where there is a specific and important element of complexity in reconciling the interests of patients, clinicians, managers and policy makers (see Braithwaite (2018)). This suggests that prolonged implementation of change is likely to be the norm.

We contribute to two bodies of knowledge in economics. The first concerns the study of the effects of DRG systems in healthcare. Since the adoption of the DRG-based prospective payment system for Medicare in the United States from 1983, a considerable literature developed postulating and then testing for the effects of the transition from a cost-based reimbursement system to fixed-price DRGs. The subsequent adoption of similar payment systems in other countries resulted in the extension of this field of study to consider the impact of moving to a DRG system from alternative financing mechanism, especially fixed annual budgets for healthcare providers. A summary of this field can be found in Busse et al. (2011) who concludes that theoretical analysis has emphasized that a DRG system provides incentives for cost-saving when moving from cost-reimbursement and an incentive to increase activity when moving from fixed budgets. In the latter case it is similarly argued that since increased treatment within a given budget requires economizing on resources that there will be cost-saving effects as well. Concern has been raised that cost-saving might also imply compromising on service quality although early studies found little evidence in practice (Desharnais et al., 1987) and further theoretical analysis provides a basis for supposing that quality might be maintained as a means of maintaining treatment numbers, provided that prices are appropriately set (Chalkley & Malcomson, 1998; Ma, 1994).

The empirical investigation of these issues has typically focused on measures that proxy either the cost of treatment or the quality of care given. In the first category, by far the most prevalent measure used is the length of stay (LOS) of patients in hospital and that is the focus of this paper. An important caveat is that whilst LOS plausibly relates to the cost of a hospital treatment, other things equal, the relationship is not a direct one (Carey, 2015).

Meng et al. (2020) conducted a meta-analysis of studies evaluating the consequences of the introduction of DRG payment systems, worldwide (USA, UK, Australia, France, Korea, Mainland China and Taiwan). In order to minimize the bias among the analyzed studies, they only included papers where the study design allowed for causal impact estimates: controlled before-after analysis (CBA), and interupted time series (ITS) analyses where the start of the program was clearly defined and there were at least three pre-policy data points. Among the 18 studies included, 13 were CBA studies, and five were ITS studies. Their meta-analysis showed that DRG-based payment was associated with lower LOS (a relatively small 8.07% decrease, about 1.2 days for a 15-day stay), while it increased readmission rates by 1.36%.

The impact on LOS was driven by the five ITS studies, where meta-analysis showed that DRGs-based payment was associated with a significant drop in LOS of 10.76% with a 95% confidence interval (CI) of −18.54 to −2.98, but the effects decreased over time. The meta-analysis of CBA studies didn't show a significant decrease in LOS after the implementation of DRGs-based payment. This is consistent with the findings of Kahn et al. (1990) who showed that that DRGs-based payment initially reduced the LOS, but it appeared to stabilize after the initial decrease. The authors highlight the uncertainty in these estimates, due to the small number of studies under analysis, and also due to the potential effects of secular decrease in LOS, associated with the evolution of medical and social practices.

Palmer et al. (2014) review a wider range of studies on the effect of activity-based financing, however they don't apply the strict study design criteria of Meng et al. (2020). Their review focusses on different outcome measures, discharge to post-acute care (for which they find a significant increase), and they also found some evidence of increased readmission rate.

Overall LOS is clearly an important indicator and most studies have found some reduction in LOS associated with the adoption of DRG systems. Two conclusions can be drawn, first that the starting point in terms of healthcare financing matters, with studies that have focused on publicly financed healthcare and hospitals being given fixed budgets finding a smaller impact on LOS than studies where the originating payment system is explicitly cost-based. Second there is a paucity of evidence concerning whether any reduction in LOS is a permanent or transitory feature of the adoption of DRGs, and, if the former, whether the impact grows or diminishes over time. The present study directly addresses this second gap in knowledge by examining data over a 15 year period with a post-DRG period of 9 years. The most closely related work is that of Farrar et al. (2009) who study the same healthcare system as we investigate – the NHS in England and who use the same control country (Scotland) for a DiD specification. Compared with that study we utilize considerably expanded data, adopt a portfolio of empirical methods and produce therefore both more extensive evidence of an impact effect and the first evidence of a long run effect of the policy change. Unlike the Farrar et al. (2009) study which also provides evidence of an absence of any effect on quality of service, we focus exclusively on the impact on LOS. Where Farrar et al. (2009) find an 8%–18% reduction in LOS, our estimates find similar or larger impact magnitude of effect, which then grows substantially over time. After 10 years we associate the adoption of the DRG system with reductions between 20% and 70% in LOS, which even accepting the potentially small elasticity between LOS and cost reductions (as in Carey (2015)) would constitute a large magnitude of effect of DRG-based payment on hospital costs.

The second body of knowledge to which this study contributes is the econometric analysis of non-experimental data to establish the causal impact of policy interventions.

We compare the relative merits of policy evaluation methods that use long time-series built from micro data, while highlighting and critically assessing the assumptions the methods make. First, in the case of DiD, we explore the assumption that trends in LOS in England and Scotland have been parallel over the pre-policy period and find some evidence contrary to this assumption. Second, we attempt to construct a "Synthetic England", collapsing the micro data to aggregate (Healthcare Resource Group (HRG) specific and overall) time series for England and the 14 Scottish Health Boards (SHB) in Scotland. Here, we find that the traditional application of the SC approach has failed to find a good fit in the pre-policy period. Hence, we employ a recently suggested novel method, synthetic differences in differences that can adjust for the bias resulting from poor pre-policy fit (Arkhangelsky et al., 2019), contributing to the literature that critically examines and extends the SC method for health policy evaluations (Kreif et al., 2016; O'Neill et al., 2020; Ryan et al., 2016). Finally, as a robustness check, we apply an ITS strategy, which does not use a control group but assumes the ability to construct the counterfactual from modeling the pre-policy trend of LOS in England only.

The setting for our study is the NHS in England which is a publicly funded healthcare system, within which hospital admitted care constitute approximately one quarter of total expenditure. As with any healthcare system there are a myriad of institutional details; we focus here only on the salient features for our study.

The foundations for establishing different purchasing mechanisms were put in place in 1989 when previously unified delivery organizations were separated into purchasers and providers, the latter usefully being regarded as hospitals for our study. Purchasers are public bodies that receive a budget through a government department and are charged with meeting the healthcare needs of the respective populations. Their precise composition, geographical responsibilities and titles have changed over time but these issues are not central to our focus. Hospitals are very predominantly in public ownership although in recent years there have been contracts with private sector (for-profit and not-for-profit) organizations that own hospitals and treat NHS patients. None of these are a part of our study.

From 1989 the intention was for purchasers to enter into contractual agreements with hospitals with discretion as to exactly what form those agreements took. However, the system of setting hospital level budgets – known in the purchasing terminology of the NHS as a Block Contract – tended to persist in spite of the intention that purchasing should move toward activity-related payments. Hence, starting in 2003 the DRG system we are studying began to be rolled out. In the NHS this was called PBR which is functionally a DRG system. Patients treated by a hospital are assigned to a category called a HRG which is equivalent both in purpose and definition to a DRG. The hospital is paid a fixed, nationally set, price for each patient in each HRG. Pertinent to our study the system differentiates between patients whose hospital treatment is planned in advance – termed elective treatments – and those who are admitted as an emergency (either through an emergency department or referred as an urgent case by their physician). A comprehensive description of the system is available in DHPRT (2012).

As described in more detail in Farrar et al. (2009) and in Chapter five of DHPRT (2012) the phased introduction of this system had a number of elements. Some hospitals – those which had gained a greater independent Foundation Trust status – and some kinds of elective treatments were brought into the system in a phased manner. This facilitates a standard before-and-after as well as a within England DiD design method for our evaluation. In addition, the neighboring NHS is Scotland did not adopt any such DRG financing reforms and can therefore serve as a possible control. In previous studies such as Farrar et al. (2009) and Chalkley et al. (2017), Scotland has been treated as a single entity as a counterfactual for England. One novel element of our study is to develop this further explore the formulation of a SC -- a weighted combination of localities in Scotland. Whilst the constituent populations of England and Scotland are different and therefore have different healthcare needs there are very few differences in population trends and hence methods which identify divergence in trends (omitting levels) following the move to DRG have intuitive credibility in respect of identifying the policy effect (Chalkley et al., 2017).

Whilst over the period we study there have been numerous other policy initiatives in respect to healthcare delivery in both England and Scotland none of these has been argued to have had a fundamental impact on the financing and delivery of healthcare that the DRG system has.

## 2 | DATA

This study concerns the analysis of aggregated, anonymized administrative healthcare data. No individuals can be identified from this study and it conforms with the ethical guidance and approval processes of our institution.

To investigate how the DRG system might have impacted on hospital stays we use detailed administrative records of each and every hospital treatment in England and Scotland over our study period: a total of more than 200 million treatments. Data are recording in both systems on the basis of hospital *episodes*. Thus, we use episode level data on in-hospital care from what is termed the Admitted Patient data set of Hospital Episode Statistics (HES) in England and the equivalent for Scotland which is called the Scottish Morbidity Record 01. These data are recorded in financial years running from 1 April to 31 March and so henceforth we use 2001/02 to denote the financial year April 2001 to March 2002, and so on. Our data cover the financial years 1997/98 to 2013/14. We exclude all patients recorded as maternity admissions and regular attenders (i.e., admissions that are part of a series of planned admissions, e.g., dialysis) in England as these are not present in the Scottish data set.

In the source data a patient's treatment may have several concurrent or sequential episodes of care in hospital and the nearest analogue of LOS is recorded as a Continuous Inpatient Spell (CIPS). To ensure the data at CIPS level is comparable for England and Scotland requires some processing because in the England definition gaps are allowed as patients move between hospitals whereas in Scotland the equivalent measure excludes such gaps (Health and Social Care

Information Centre, 2014; ISD Scotland, 2012). To make the measures comparable we modified the procedure to obtain CIPS in England so as not to allow for these gaps. We calculate the duration of the CIPS (which we will refer to as LOS henceforth) as the sum of the duration of the episodes that form the CIPS.

Each hospital treatment is classified according to the first episode of care so that the first episode in each CIPS was used to determine whether the treatment was an elective or an emergency and which DRG (HRG) it was. The HRG classification system changes over time and so we unified classification on the HRG version used in the financial year 2009/10. Whilst Scotland does not use the DRG system the Scottish treatments can be classified on the same basis. Both datasets have high quality reporting, with over 88% of records with main diagnosis and procedures being recorded correctly (ISD Scotland, 2012, NHS Digital); the percentage of valid codes is higher in England, this could be due to the use of HRGs for payment, which provides an extra incentive to record data correctly. Thus we create the same classification of patients across the two countries. This provides us with a unique long run and comparable data set spanning two jurisdictions over 15 years.

The source data is at the level of each individual treatment in hospital, this comprised 183 million observations from England and 18 million from Scotland. To compare changes in LOS other things equal we wish to take account of as many differentiating factors concerning these treatments whilst making the analysis tractable. For each treatment we observe age, sex and a deprivation measure based on the patient's location. We therefore aggregate the data into 78,000 HRG-country-year combinations, for each of which we calculate the average LOS and the proportions of patients falling in each age group, sex and deprivation decile. The HRG-country-year data has more observations for England (40,000) than for Scotland (38,000) but observations are equally distributed between Elective and Emergency activity. For context it is useful to note that in 2013, England's population was around 10 times that of Scotland (53.87 vs. 5.33 m) and that Scotland has a slightly older population, with a smaller proportion of under 20 s (21.9% vs. 23.8%) and a larger proportion of over 60 s (22.5% vs. 21.6%) than England (ONS, 2013). National Health Service expenditure per capita was slightly larger in Scotland (£2148 vs. £2000 in 2011/12; Hawe & Cockcroft, 2013). A detailed comparison of the two health systems is contained in Chapter four of Bevan et al. (2014) which establishes that in terms of most published indicators of health system configuration the two countries are very similar.

Figure 1 shows the average LOS in both countries, separately for Elective and Emergency activity. Note that the *y*-axis scales are not the same in both plots, Elective CIPS are shorter than Emergency ones. The vertical line corresponds to 2003/04, the year when the DRG system was introduced.

Length of stay has decreased over time in many countries, this reduction has coincided with the introduction of prospective payment in several countries, but it is not limited to them (OECD, 2013). One key driver in these reductions, which have been occurring since the 1960 s in, for example, the US, is changing medical technology and practise style (Kalra et al., 2010). The question therefore becomes one of to what extent policy interventions, such as payment reform have either led to or accelerated that trend. From the Figure it can be seen that Scotland has exhibited a decline in lengths of stay over the study period but that it appears to be a slower decline than in England. Whereas there have been policy targets in terms of reducing lengths of stay in Scotland – these formed one element of efficiency targets between
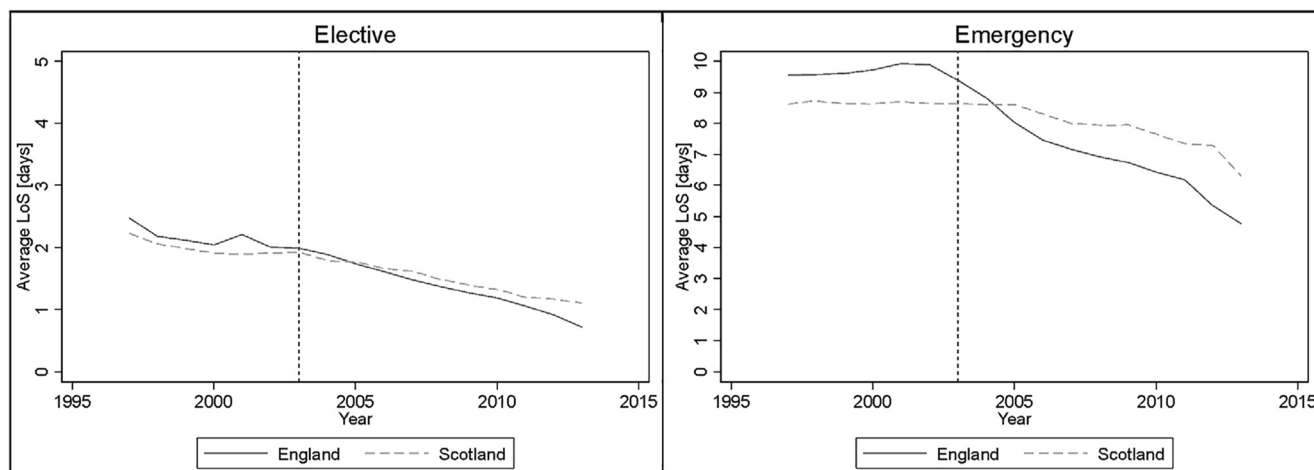


**FIGURE 1** Average length of stay (LOS)

2008/09 and 2010/11 (NHS Scotland, 2014) – there has been no financial incentive or penalty of the kind embodied in the DRG system in England. Thus, our research question concerns to what extent the decline in LOS in England was a consequence of the adoption of the DRG system. We investigate this research question by constructing a counterfactual: what would have been the LOS in England in the absence of introducing the DRG system, and our empirical strategies use England's pre-existing trend as well as that for Scotland to construct this counterfactual.

It is apparent from the figure that there are potential differences in trend between England and Scotland pre-policy. As is often the case in time-series data there are idiosyncratic deviations from a perceived average trend, and this is particularly the case for England which does display some small upward deviations prior to the policy intervention. It is for this reason that we consider a variety of methods in addition to conventional difference-in-differences which relies on a parallel trends assumption.

## 3 | METHODS

Throughout, we use the potential outcomes framework (Rubin, 1974). Suppose there are $i = 1,...,n$ units, and $T$ time periods, where $t = 1,...,t'$ are pre-treatment, and $t' + 1,...,T$ are post-treatment. The potential outcomes (LOS) for HRG $i$ in period $t$ in the presence and absence of the policy are denoted by $Y_{it}^1$ and $Y_{it}^0$ respectively. Let $D_{it}$ be an indicator equal to one if unit $i$ is treated (exposed to the policy) in period $t$ and zero otherwise. In this setting $D_{it}$ will take the value zero for admissions up to 2002, and one from 2003 in England, and zero for Scotland for the entire time period. Hence the observed outcome can be written as:

$$Y_{it} = D_{it}Y_{it}^1 + (1 - D_{it})Y_{it}^0$$

We assume the following linear model for the potential outcome in the absence of treatment:

$$Y_{it}^0 = X_{it}'\beta + \lambda_t\mu_i + \varepsilon_{it}$$

where $X_{it}$ is a $(1 \times k)$ vector of observed time-varying covariates, $\beta$ is the $(k \times 1)$ vector of their coefficients which is assumed to be the same for both groups, $\mu_i$ represents an unobserved time-invariant variable with $\lambda_t$ capturing the effect of that unobserved variable in period $t$ and $\varepsilon_{it}$ represents exogenous unobserved idiosyncratic shocks. Allowing for an additive treatment effect that may differ by HRG and period ($\tau_{it}$), the observed outcome can be written as:

$$Y_{it} = X_{it}'\beta + \lambda_t\mu_i + D_{it}\tau_{it} + \varepsilon_{it} \tag{1}$$

The estimand of interest throughout is the average treatment effect for the treated (ATT), ($\tau_{it} \mid D_{it} = 1$), separately over time periods, and also aggregated over the post-treatment period, $t > t'$.

In the following sections, we briefly outline the three main methodological approaches we use in the paper: (3.1) differences in differences, with extension to non-parallel trends, (3.2) the SC approach and its extension, synthetic differences in differences and finally, (3.3) the ITS approach, for comparison. We summarize the main assumptions and advantages and limitations of these methods in Table 1.

### 3.1 | Difference-in-differences

Equation [1] can be estimated using a difference-in-differences (DiD) model, if the effect of the time-varying confounder can be assumed to be constant $\lambda_t = \lambda$. In this case, the parallel trends assumption holds (Angrist & Pischke, 2009; Jones & Rice, 2011):

$$E(D_{it} = 1, X_{it}) = E(D_{it} = 0, X_{it}) \quad \forall t > t' \text{ (A1: Parallel trends)}. \tag{2}$$

In [2] $t'$ represents the final pre-treatment period, and the ATT can be consistently estimated by DiD using a two-way fixed effects regression (Bertrand et al., 2004; Carpenter & Stehr, 2008; Fletcher et al., 2015; Wen et al., 2015)

$$Y_{it} = X_{it}'\beta + \mu_i + \delta_t + D_{it}\tau + \varepsilon_{it}, \tag{3}$$

**TABLE 1** Comparison of the econometric approaches applied

|  | Assumptions | Pro | Con |
|---|---|---|---|
| Differences in differences | Parallel trends | Well understood method with limited data requirement | Reliance on parallel trends which is implausible in study setting |
| Differences in differences with unequal trends | Potentially non-parallel, but linear trends | Well understood method, extends differences in differences to allow for non-parallel trends | Reliance on correct specification of trends, requires several pre-policy time periods for fitting pre-policy trends |
| Synthetic control | No parallel trends assumption. | Doesn't require parallel trends. | Requires several pre-policy time periods. |
|  | Weighted combination of control units reproduces treated outcome trends | Intuitive approach with easy visual expansion of success in re-creating pre-policy trends. | In its original version requires the treated outcomes to be a convex combination of control outcomes (doesn't allow for large pre-policy differences) |
| Synthetic differences in differences | Parallel trends in reweighted outcome | Relaxes both synthetic control and DiD assumptions. Uses re-weighting to create parallel in pre-policy trends, adjusts for remaining pre-policy difference | Data adaptive method, can be "black box" |
| Interrupted time series | Linear trend with change in slope | Well-understood method | Reliance on correct modeling of counterfactual with treatment group's own trend |

In [3] $Y_{it}$ is the average LOS in HRG-country $i$ in year $t$, the explanatory variable $D_{it}$ indicates policy implementation (a dummy variable indicating the onset of PbR in England), $X'_{it}$ are control variables of HRG overall patient characteristics (age, sex, deprivation proportions), $\mu_i$ are the HRG fixed effects, $\delta_t$ are the time fixed effects, and $\tau$ is the ATT (alternatively the treatment effect $\tau_t$ can be estimated for each post-treatment period).

In order to capture a potential heterogeneity in the effect of the policy over time, we extend [3] by interacting the post-policy dummy and indicators for the year:

$$Y_{it} = X'_{it}\beta + \mu_i + \delta_t + D_{it}\,\delta_t\tau_t + \varepsilon_{it}. \tag{4}$$

Next, we use a DiD specification that relaxes the parallel trends assumption by fitting separate trends for the treated and control groups (Bell et al., 1999, Moreno-Serra and Wagstaff, 2009). The common trend is $t$, the differential trend for England is $t * E$ and the differential trend in the post treatment period is $t * D_{it}$:

$$Y_{it} = X'_{it}\beta + \mu_i + \alpha_1 t + \alpha_2 t * E + \alpha_3 t * D_{it} + \varepsilon_{it} \tag{5}$$

We test the parallel trends assumption required in Equations [3] and [4] by comparing the trends in LOS for England and Scotland before the intervention, that is, up to 2002. For each country and type of admission we regress average LOS on a time trend and indicators for age groups, sex, deprivation deciles, and HRGs.

## 3.2 | Synthetic control and synthetic differences in differences

The SC method provides and alternative estimator of the ATT (Abadie et al., 2010) when the effects of unobserved confounders, $\lambda_t$, cannot be assumed to be constant over time.

In summary, the SC method finds a weighted average of the control units $(1,...,J)$ – the SC – with similar outcomes and observed covariates to the treated unit $(j = 1)$, over the pre-intervention period:

$$\sum_{j \in Control} w_j Y_{jt} = Y_{1t}, \quad \forall t \leq T_0 \quad \text{and} \quad \sum_{j \in Control} w_j Z_i = Z_1, \quad \forall t \leq T_0 \tag{6}$$

In [6] $w_j$ is an element of $W$ representing the weight for control $j$, with $0 \leq w_j \leq 1$. If the SC and treated unit appear similar in terms of outcomes over an extended period, it is plausible that they are similar in terms of both observed and unobserved determinants of the outcome variable (Abadie et al., 2015). Hence, the post-intervention outcome for the SC represents the counterfactual, treatment-free potential outcome for the treated unit.

The SC approach was originally proposed for contexts with a single treated unit but can be readily extended to contexts with multiple treated units by applying the method to each treated unit in turn, or by averaging the treated units to obtain a single 'treated unit' (Nonnemaker and Farrelly (2011), Dube and Zipperer (2013), Kreif et al. (2016), inter alia). We follow this approach and use the control pool of 14 SHB to create a synthetic (aggregate) England.

A potential limitation of the SC approach is that since the weights are restricted to be between 0 and 1, the treated unit must lie in the 'convex hull' of the control units to avoid bias (Abadie et al., 2010). If the levels of LOS has been very different in Scotland to England, this assumption may not be plausible. To address this, we also employ an extension of the SC approach, the Synthetic Difference-in-differences (Arkhangelsky et al., 2019), which relaxes this assumption, by first constructing an imperfect SC group, then finding an interval in the pre-policy period where parallel trends can be assumed, and adjusting the remaining bias in the effect estimates using DiD.

For both the SC approaches, we use the data in a different way because in place of HRG-country totals, we need totals for the treated and control units: England and the 14 SHB. As for the HRG-country totals we calculate the proportion of population by age group, sex and deprivation decile.

Hence, the dependent variable is the average LOS in unit $i$ (where the unit can be England or one of the SHB) in year $t$. The years 1997/98–2002/03 are the pre-intervention period, that is, the years used to calculate the weights for each SHB in order to recreate the observed LOS in England. The predictors are the patient characteristics (proportion of patients in age groups, sex, deprivation deciles) and indicators for each HRG.

## 3.3 | Interrupted time series

It may be possible that both the DiD and the SC approaches fail in their assumptions: by finding evidence against the parallel trends assumption, and also by failing to provide an adequate (synthetic) control group of the untreated Scotland that matches the pre-policy trends of England. In this case, given the relatively long pre-policy data available, a simple but potentially useful approach is available that may provide a good estimate of the counterfactual. In an ITS analysis, we use the pre-policy trends of England to estimate what would have happened without the policy. A straightforward specification allows for a change in the slope of the trend:

$$Y_{it} = X'_{it}\beta + \alpha_1 t + \alpha_2 t D_{it} + \varepsilon_{it}, \tag{7}$$

where, as before, $D_{it}$ is 0 before the start of the policy, and 1 after the start, $\alpha_1$ is the slope the underlying linear trend, and $\alpha_2$ is the change in the slope of the trend, after the start of the policy, but with using data from England only. In period $t$, the effect of the policy can be expressed as $\alpha_2 t$.

## 4 | RESULTS

All analyses are conducted separately for elective and emergency CIPS. Unless otherwise indicated in the Table, the results correspond to the period 1997/98–2013/14. All analyses are conducted using Stata, except the synthetic difference-in-difference that uses $R$.

## 4.1 | Difference-in-differences

The results of the standard DiD analysis (Table 2) show that the implementation of PbR in England led to reductions in LOS by 0.6 days in Elective and 1.2 days for Emergency treatment, which, compared to the averages before PbR started

**TABLE 2** "Classic" difference-in-differences regression results

| | Elective | Emergency |
|---|---|---|
| PbR indicator | −0.6149*** | −1.1922*** |
| | (0.2380) | (0.1544) |
| Year dummies | YES | YES |
| Country | YES | YES |
| Sex | YES | YES |
| Age groups | YES | YES |
| Deprivation deciles | YES | YES |
| HRGs | YES | YES |
| N | 39,435 | 38,980 |
| R-squared | 0.4293 | 0.7029 |

*** indicates significance at 1%.

**TABLE 3** Comparison of results year by year

| | Classic DiD with year effects | | DiD with time trend | | ITS | | SDID | |
|---|---|---|---|---|---|---|---|---|
| | Elective | Emergency | Elective | Emergency | Elective | Emergency | Elective | Emergency |
| 2003 | −0.3017 | 0.4226 | −0.2551 | −0.5206** | −0.3641* | −0.5760*** | −0.014 | −0.377 |
| | (0.5080) | (0.3288) | (0.3520) | (0.2283) | (0.2022) | (0.1506) | | |
| 2004 | −0.0781 | 0.1387 | −0.2976 | −0.6074** | −0.4248* | −0.6719*** | 0.000 | −0.848 |
| | (0.5076) | (0.3290) | (0.4106) | (0.2664) | (0.2359) | (0.1757) | | |
| 2005 | −0.1936 | −0.4398 | −0.3401 | −0.6941** | −0.4855* | −0.7679*** | −0.076 | −1.730 |
| | (0.5058) | (0.3275) | (0.4693) | (0.3044) | (0.2696) | (0.2009) | | |
| 2006 | −0.2061 | −0.5068 | −0.3826 | −0.7809** | −0.5462* | −0.8639*** | −0.106 | −1.825 |
| | (0.4982) | (0.3224) | (0.5280) | (0.3425) | (0.3032) | (0.2260) | | |
| 2007 | −1.3917*** | −0.5673*** | −0.4251 | −0.8677** | −0.6069* | −0.9599*** | −0.152 | −1.790 |
| | (0.4966) | (0.3217) | (0.5866) | (0.3806) | (0.3369) | (0.2511) | | |
| 2008 | −0.7701 | −1.0970*** | −0.4676 | −0.9544** | −0.6676* | −1.0559*** | −0.126 | −2.020 |
| | (0.4959) | (0.3219) | (0.6453) | (0.4186) | (0.3706) | (0.2762) | | |
| 2009 | −0.2932 | −1.1448*** | −0.5101 | −1.0412** | −0.7283* | −1.1519*** | −0.142 | −2.130 |
| | (0.4963) | (0.3218) | (0.7039) | (0.4567) | (0.4043) | (0.3013) | | |
| 2010 | −0.2604 | −0.9485*** | −0.5526 | −1.1280** | −0.7889* | −1.2479*** | −0.173 | −2.118 |
| | (0.4959) | (0.3216) | (0.7626) | (0.4947) | (0.4380) | (0.3264) | | |
| 2011 | −0.4913 | −2.1277*** | −0.5951 | −1.2148** | −0.8496* | −1.3439c | −0.200 | −1.973 |
| | (0.4970) | (0.3223) | (0.8213) | (0.5328) | (0.4717) | (0.3515) | | |
| 2012 | −1.3544*** | −3.7521*** | −0.6374 | −1.3015** | −0.9103* | −1.4399*** | −0.281 | −2.545 |
| | (0.4974) | (0.3228) | (0.8799) | (0.5708) | (0.5054) | (0.3766) | | |
| 2013 | −1.3816*** | −3.0236*** | −0.6801 | −1.3883** | −0.9710* | −1.5359*** | −0.378 | −2.466 |
| | (0.4975) | (0.3239) | (0.9386) | (0.6089) | (0.5391) | (0.4017) | | |

Abbreviations: ITS, interrupted time series; SDID, synthetic difference-in-differences.

*, **, *** indicate significance at 10%, 5% and 1%, respectively. Placebo tests for SDID results are similar to those for the overall SDID results (Figure 4).

(average LOS in England in 2002 was 2.01 days for elective and 9.89 days for emergency) represent reductions of 30% for Elective and 12% for Emergency treatments.

To account for potentially dynamic impacts of the policy, we estimated an extension of the DiD model, by adding year specific effects (see Equation [4]). Year-by-year results are reported in Table 3 (estimated regression coefficients can be found in the Appendix, Table A1). For elective activities, we find that all year effects have a negative sign, but except for 2007, 2013 and 2015, they have 95% confidence intervals that include zero. For emergency activities, we observe

significant yearly impacts, starting at a reduction of 0.6 days in 2007 and gradually increasing to a reduction of three days by 2013.

We test the parallel trends assumption by comparing the trends in LOS for England and Scotland before the intervention. For each country and type of admission we regress average LOS on a time trend and the same controls we have used before (indicators for age groups, sex, deprivation deciles, and HRGs). The estimated trends are reported in Table 4. The results of a Wald test (with null hypothesis that the coefficients in the two countries are the same), indicate that the parallel trends assumption does not hold in the pre-treatment period: the null hypothesis is rejected at 10% level for Elective activity, and at 5% for emergency activities.

Hence we implement a further extension of the DiD approach, and estimate Equation [5] that allows for an interaction between trends and treatment status. Year-by-year results are reported in Table 3 and estimated regression coefficients can be found in the Appendix, Table A2. Here, we find significant decreases only for emergency conditions, ranging from 0.5 days in 2003 to 1.4 days in 2013.

## 4.2 | Synthetic control approaches

Using the traditional SC approach to create the 'best' weighted combination of Scotland Health Boards, we did not manage to achieve a satisfactory pre-policy fit of the LOS data of England in the pre-intervention period (Figure 2).

In order to address the unsatisfactory pre-treatment fit, we employ the Synthetic Difference-in-differences method, using again the SHB as control units for England. As a first step we estimate the best possible synthetic England (see dotted blue lines on Figure 3), aiming to match well the pre-intervention data of England (solid blue line on Figure 3). Then, we adjust the remaining pre-policy differences in length of stay (LOS) by applying DiD on the treated and SC data. The straight blue line represents the change in LOS from a pre-treatment period through the post treatment period for the control group, while the solid red line represents the change for the treated group and the dashed red shows the counterfactual change that would have happened, had the treated group moved in parallel to the control group. The effect of policy was then estimated as the difference between the two red lines in the post-policy period.

|  | Elective | Emergency |
|---|---|---|
| England | −0.0486 | 0.1620*** |
|  | (0.0417) | (0.0301) |
| Scotland | −0.2482*** | 0.0191 |
|  | (0.0920) | (0.0581) |
| Wald test Prob > Chi2 | 0.0759 | 0.0258 |

**TABLE 4** Estimated trend coefficients. Pre-treatment period (1997/98–2002/03)

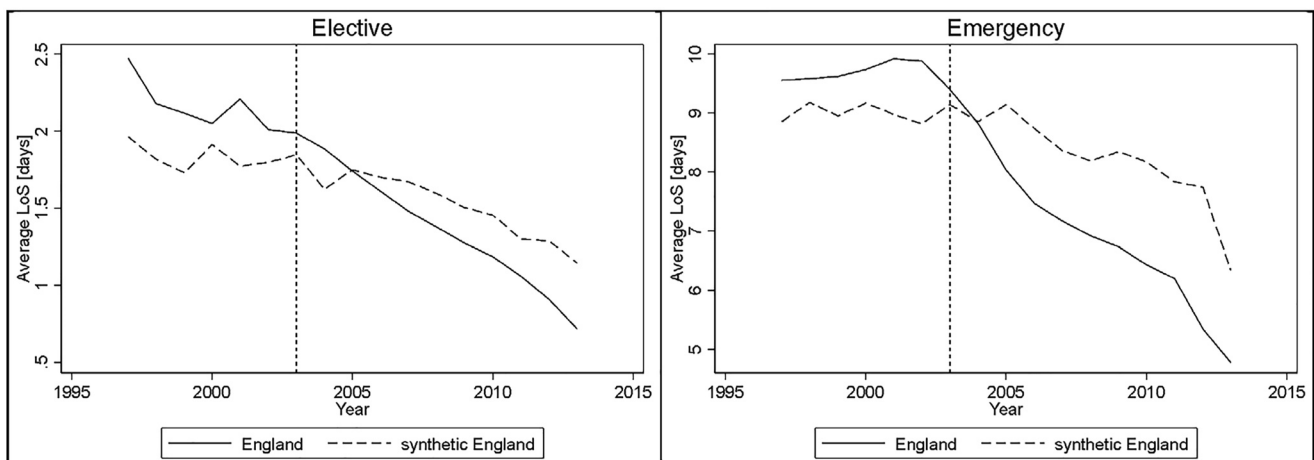*** indicates significance at 1%.



**FIGURE 2** Synthetic control (SC). Treated Unit: England. Control Units: Scottish Health Boards (SHB)
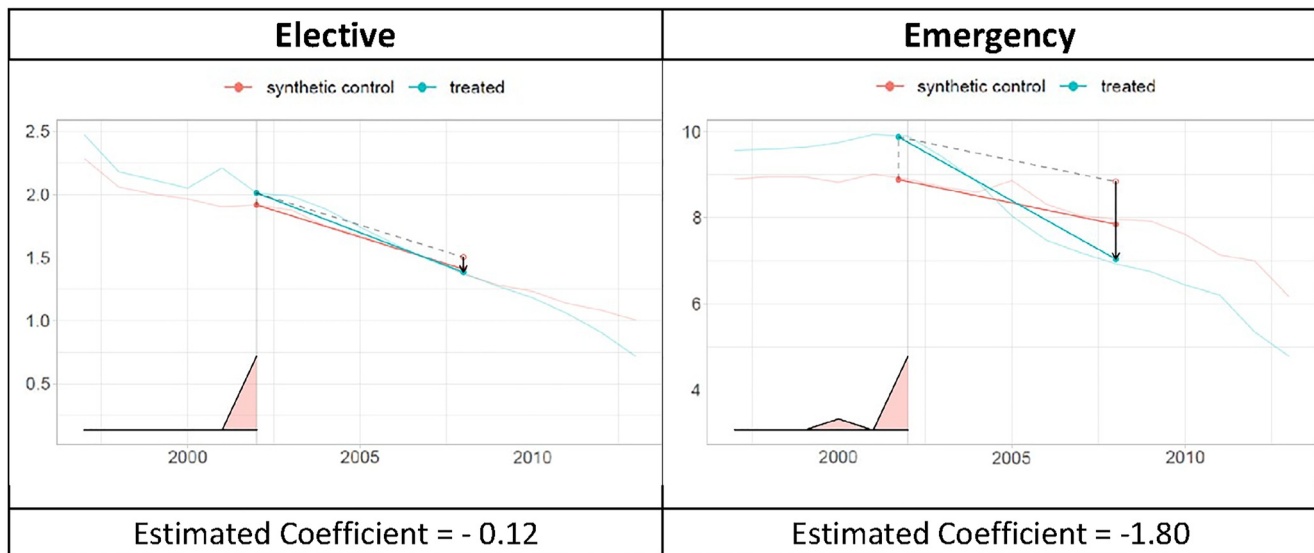
**FIGURE 3** Synthetic difference-in-difference. Treated Unit: England. Control Units: Scottish Health Boards (SHB)
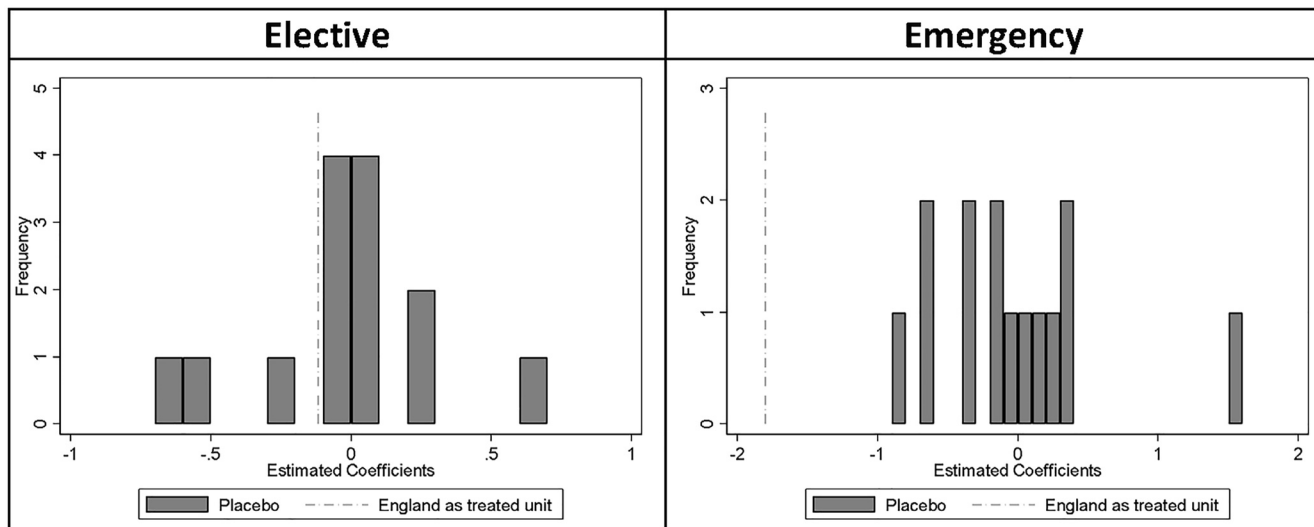


**FIGURE 4** Synthetic difference-in-difference. Placebo Tests

To determine whether the results shown in Figure 3 are significant, we use a placebo test; that is, we remove England from the data and make each Scottish Health Board the treated unit. Figure 4 shows the histograms with the distribution of the estimated coefficients from using different Health Boards as the treated unit, the vertical line shows the estimated coefficient when England was the treated unit. The estimated coefficient for elective activities lies within the placebo estimates, while the one for emergency is outside that distribution.

We estimate the ATT overall on the post-policy time period, but also year-by-year, applying the methodology considering each year of the post-treatment period one their own, see Table 3 for the estimates, we do not report the placebo tests as they are similar to those for the overall results (Figure 4), that is, within the placebo estimates for Elective and outside the placebo range for emergency, in most years.

We find that the point estimates from the synthetic difference-in-differences (SDID) broadly correspond to the findings from the DiD approaches: there are close to zero effects found (generally inside of the CIs of the year-specific DiD) for electives, while relatively large, increasingly negative effects (corresponding to the effects found with the DiD with time trends, for the end of the observation period) are found for the emergency conditions.
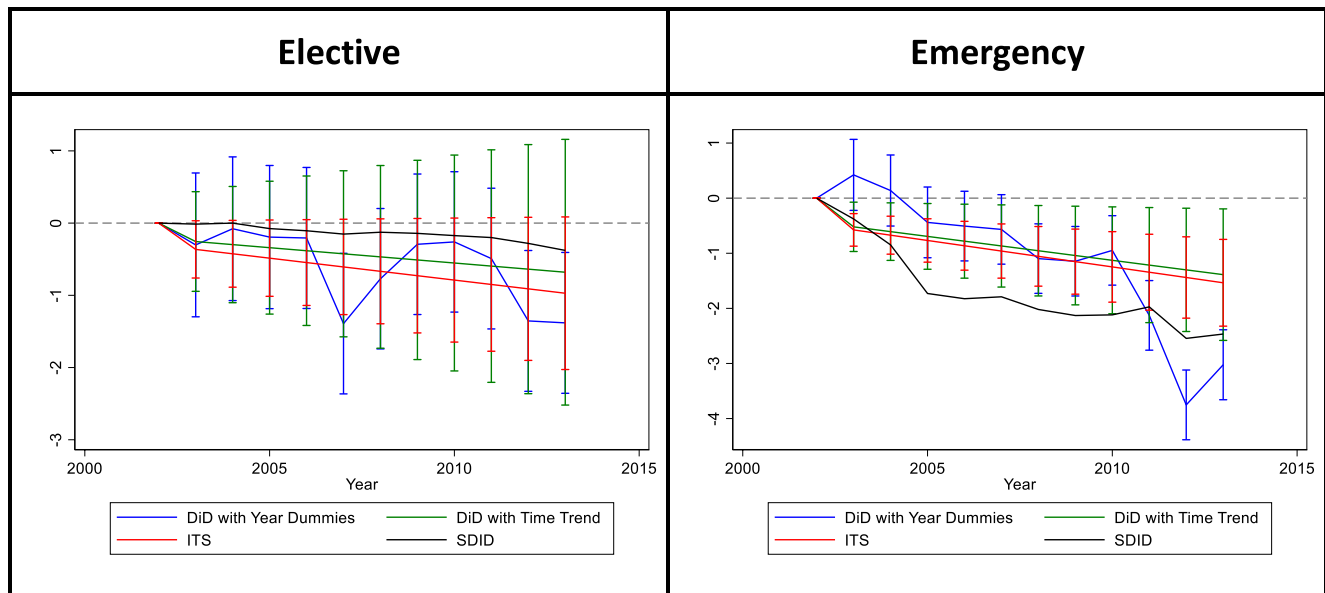
**FIGURE 5** Summary results (see Table 3)

## 4.3 | Interrupted time series

Finally, we explore the long series of data available and use pre-PbR England as a 'control' for itself once the policy is in place.

The ITS results are very similar to those obtained with the Difference-in-differences regressions (see Table A2 and Table A3 in the Appendix). However, comparing the ITS results with the DiD results year by year, Table 3, we see that the ITS year results have higher absolute value than those for DiD with Trend, and they do not always have the same sign.

Figure 5 shows the year-by-year effects reported in Table 3 graphically.

## 4.4 | Summary of results

The key estimates of interest – the impact of the DRG system on LOS are summarized in Table 3 and Figure 5. Across all of the empirical methods used there is evidence of a substantial and growing impact of the DRG system in terms of reducing LOS. The results under the different methods differ in terms of point estimates and precision, the latter largely being driven by the greater demands made by some of the methods on the data in regard to identification. Nevertheless, there is a strong degree of agreement regarding the magnitude and, save for the DiD with year effects, the timing of the effects.

For elective care we estimate a long run effect (measured in 2013) of between −1.4 and −0.7 days using DiD and ITS methods. The SDID gives an estimate of −0.4. These figures correspond between 35% and 70% reductions relative to the 2002 average LOS. For emergency care the full range of results is from −1.4 to −3 days, corresponding to between 14% and 30% reductions relative to 2002 average LOS. In comparison the initial effects over the period 2003–2005 are smaller and of borderline significance. The results for emergency treatment in particular are not well-determined over this earlier period.

## 5 | DISCUSSION

A key concern regarding the impact of DRG systems is their impact on both the quality and duration of treatment. In the context of the system studied in this paper – the NHS in England – one hoped-for benefit of payment reform was to encourage a greater throughput of patients. This was of particular relevance to a system characterized by waiting lists (Harrison & Appleby, 2005).

Previous studies, such as Farrar et al. (2009) have established that the English DRG system reduced hospital LOS but in common with most studies across many jurisdictions for which DRG systems have been introduced there are substantial limitations in the data that have been available and hence the robustness of the methods used and the subsequent findings. Against this background we assembled the longest run data set to date, covering 15 years of hospital treatment across two jurisdictions, one of which did not introduce a DRG system and which serves as a useful control for the policy enacted in England. We have subjected those data to an extensive range of econometric methods that have been developed to establish the impact of policy intervention from routinely observed data. Across all of those methods the results concur – there was a substantial, long run and increasing over time impact of the DRG in reducing lengths of stay. The overall reductions in LOS over this long run have been greater than previously estimated.

There are of course a number of caveats. Our study has concerned just one specific healthcare system and has often been noted health systems are idiosyncratic in regard to many features. The most relevant features of the NHS in England that need to be considered before drawing any analogy between its experience with DRGs and the likely impact in other domains are: it is publicly funded, the majority of its hospitals were (and continue to be) publicly owned, it moved to DRGs from a system of approximately fixed budgets. All of these features likely impacted on the way in which the payment reform played out. It is also important to note that whereas our results are similar across a broad range of empirical methods they are not identical and all of those methods have limitations.

The 15-year time horizon we study is both a strength and a limitation of our study. We have been able to consider substantial before- and after-implementation experiences of the policy intervention, but all methods rely on an important contextual assumption; that no major other health policies take place either in England or Scotland around the time of the DRG payment reform. Detailed comparisons of the English and Scottish health systems are presented by Bevan et al. (2014) and National Audit Office (2012). In respect of hospital services, both studies highlight both the common experiences of these countries prior to the DRG reform we study. Both also highlight this reform, and the accompanying changes in institutional structures, as the main point of departure between the two systems. There have, of course, been other policy initiatives in both jurisdictions but these do not coincide with DRG reform and in any case the general approach has been common across the jurisdictions, specifically an emphasis on efficiency savings and a focus on quality of care. This does however raise the question of whether changes can be attributed to the DRG policy or other initiatives.

All our methods also rely on the assumption that the counterfactual outcomes of England without the policy have been adequately constructed. With the exception of the ITS approach, which exploited the deviation from England's own pre-policy trend, all approaches used information from Scotland to construct the counterfactual. While the standard DiD approach assumed that any unobserved factors that can affect LOS either have a common trend between Scotland and England (e.g., overall efficiency gains in healthcare technology), or the differences between such factors don't change over time (e.g., country specific healthcare practices, healthcare needs) and those components that change, have been adequately modeled with the available covariates (age, deprivation and sex composition of patients). The DiD with time trend aims to relax this assumption by modeling the pre-policy trends separately, however relies on the assumption that these linear trends are correctly specified. The synthetic difference in differences relaxes the parallel trends assumption in a more flexible way, by finding a combination of SHB which matches as close as possible the pre-policy trends of England, with the hope that this weighted combination will also recreate the unobserved components that may have a changing impact over time. The results of our tests of the parallel trends assumptions give an indication of these latter two approaches being more reliable in this setting. This also is supported by a visual inspection of trends in these data where, as discussed under the Data section, there are idiosyncratic deviations from trend prior to the policy intervention in England. For these reasons the synthetic differences in differences findings are probably to be preferred. We are constrained by examining a system reform that was introduced without variation in the degree of price incentives and hence our results relate to the overall adoption of the system, not its intensity.

The broad consensus in our findings across different methods provides some reassurance that the attribution of changes to the DRG policy is reasonable but we cannot preclude that the effect we observe is exclusively due to the use of DRGs rather than their use in conjunction with other measures that have been common across the two jurisdictions. For example, it may be that a DRG system makes other commonly pursued cost control measures more effective – thus leading to reductions in LOS.

Overall, our results support the view that DRG payment reform gives rise to real effects on the delivery of healthcare and reduces the duration of hospital stays. That is consistent with reducing the resources used in hospital care. We have added to the existing body of knowledge by providing evidence that these effects persist and grow over time. From a policy perspective these results indicate that DRG payment is an effective tool in establishing control over rising healthcare costs. Given the growth of now ubiquity of DRG financing this is of considerable policy relevance.

## CONFLICT OF INTEREST

None.

## DATA AVAILABILITY STATEMENT

This study concerns the analysis of aggregated, anonymized administrative health care data. No individuals can be identified from this study and it conforms with University of York ethical guidance and approval processes.

## ORCID

*María José Aragón* ![ORCID] https://orcid.org/0000-0002-3787-6220
*Martin Chalkley* ![ORCID] https://orcid.org/0000-0002-1091-8259
*Noémi Kreif* ![ORCID] https://orcid.org/0000-0001-9008-5690

## REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, *105*, 493–505.

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*, 495–510.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics*. Princeton University Press.

Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., & Wager, S. (2019). *Synthetic difference in differences*. National Bureau of Economic Research Working Paper Series.

Bell, B., Blundell, R., & Reenen, J. V. (1999). Getting the unemployed back to work: The role of targeted wage subsidies. *International Tax and Public Finance*, *6*, 339–360.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we Trust differences-in-differences estimates? *Quarterly Journal of Economics*, *119*, 249–275.

Bevan, G., Karanikolos, M., Exley, J., Nolte, E., Connolly, S., & Mays, N. (2014). *The four health systems of the UK: How do they compare?* Nuffield Trust and Health Foundation.

Braithwaite, J. (2018). Changing how we think about healthcare improvement. *BMJ*, *361*, k2014.

Busse, R., Geissler, A., Quentin, W., & Wiley, M. (2011). *Diagnosis-related groups in Europe: Moving towards Transparency, efficiency and quality in hospitals*. McGraw-Hill Education.

Carey, K. (2015). Measuring the hospital length of stay/readmission cost trade-off under a bundled payment mechanism. *Health Economics*, *24*, 790–802.

Carpenter, C. S., & Stehr, M. (2008). The effects of mandatory seatbelt laws on seatbelt use, motor vehicle fatalities, and crash-related injuries among youths. *Journal of Health Economics*, *27*, 642–662.

Cerna, L. (2013). *The nature of policy change and implementation: A review of different theoretical approaches*. OECD.

Chalkley, M., & Malcomson, J. M. (1998). Contracting for health services with unmonitored quality. *The Economic Journal*, *108*, 1093–1110.

Chalkley, M., Mccormick, B., Anderson, R., Aragon, M. J., Nessa, N., Nicodemo, C., Redding, S., & Wittenberg, R. (2017). *Elective hospital admissions: Secondary data analysis and modelling with an emphasis on policies to moderate growth* (Vol. 5). Health Services and Delivery Research.

Department Of Health Payment By Results Team [DHPRT]. (2012). *A simple guide to payment by results*.

Desharnais, S., Kobrinski, E., Chesney, J., Long, M., Ament, R., & Fleming, S. (1987). The early effects of the prospective payment system on inpatient utilization and the quality of care. *Inquiry*, *24*, 7–16.

Dube, A., & Zipperer, B. (2013). Pooled synthetic control estimates for recurring treatment: An application to minimum wage studies. In U. O. Massachusetts (Ed.), *Amherst working paper*.

Farrar, S., Yi, D., Sutton, M., Chalkley, M., Sussex, J., & Scott, A. (2009). Has payment by results affected the way that English hospitals provide care? Difference-in-differences analysis. *BMJ*, *339*.

Fletcher, J. M., Frisvold, D. E., & Tefft, N. (2015). Non-linear effects of soda taxes on consumption and weight outcomes. *Health Economics*, *24*, 566–582.

Forgione, D. A., Vermeer, T. E., Surysekar, K., Wrieden, J. A., & Plante, C. C. (2005). DRGs, costs and quality of CARE: An agency theory perspective. *Financial Accountability and Management*, *21*, 291–308.

Harrison, A., & Appleby, J. (2005). *The war on waiting for hospital treatment: What has labour achieved and what challenges remain?* (pp. 1–87). King's Fund. https://www.kingsfund.org.uk/sites/files/kf/field/field_publication_file/war-on-waiting-hospital-treatment-labour-full-report-john-appleby-anthony-harrison-kings-fund-4-august-2005.pdf

Hawe, E., & Cockcroft, L. (2013). *OHE guide to UK health and health care Statistics*. Office of Health Economics.

Health and Social Care Information Centre. (2014). *Methodology to create provider and CIP spells from HES APC data*. Retrieved from https://files.digital.nhs.uk/8C/0A1F5F/Compendium%20User%20Guide%202015%20Feb%20Annex%2013%20V1.pdf

ISD Scotland. (2012). *Assessment of SMR01 data 2010 – 2011*.

Jones, A. M., & Rice, N. (2011). Econometric evaluation of health policies. In P. C. Smith (Ed.), *Oxford Handbook of health economics*. Oxford University Press.

Kahn, K. L., Keeler, E. B., Sherwood, M. J., Rogers, W. H., Draper, D., Bentow, S. S., Reinisch, E. J., Rubenstein, L. V., Kosecoff, J., & Brook, R. H. (1990). Comparing outcomes of care before and after implementation of the DRG-based prospective payment system. *JAMA*, *264*, 1984–1988.

Kalra, A. D., Fisher, R. S., & Axelrod, P. (2010). Decreased length of stay and cumulative Hospitalized days despite increased patient Admissions and readmissions in an area of urban poverty. *Journal of General Internal Medicine*, *25*, 930–935.

Kreif, N., Grieve, R., Hangartner, D., Turner, A. J., Nikolova, S., & Sutton, M. (2016). Examination of the synthetic control method for evaluating health policies with multiple treated units. *Health Economics*, *25*, 1514–1528.

Lægreid, P., & Neby, S. (2016). Gaming, accountability and Trust: DRGs and activity-based funding in Norway. *Financial Accountability and Management*, *32*, 57–79.

Ma, C.-T. A. (1994). Health care payment systems: Cost and quality incentives. *Journal of Economics and Management Strategy*, *3*, 93–112.

Meng, Z., Hui, W., Cai, Y., Liu, J., & Wu, H. (2020). The effects of DRGs-based payment compared with cost-based payment on inpatient healthcare utilization: A systematic review and meta-analysis. *Health Policy*, *124*, 359–367.

Moreno-Serra, R., & Wagstaff, A. (2009). System-wide impacts of provider-payment reforms: Evidence from the health sectors of central and Eastern Europe and central Asia. *BMC Health Services Research*, *9*, A2.

National Audit Office. (2012). *Healthcare across the UK: A comparison of the NHS in England, Scotland, wales and Northern Ireland*.

NHS Digital The quality of nationally submitted health and social care data.

NHS Scotland. (2014). *HEAT Targets due for delivery 2006-2015*.

Nonnemaker, J. M., & Farrelly, M. C. (2011). Smoking initiation among youth: The role of cigarette excise taxes and prices by race/ethnicity and gender. *Journal of Health Economics*, *30*, 560–567.

OECD. (2013). *Health at a Glance 2013*.

O'Neill, S., Kreif, N., Sutton, M., & Grieve, R. (2020). A comparison of methods for health policy evaluation with controlled pre-post designs. *Health Services Research*, *55*, 328–338.

ONS. (2013). *Population estimates for UK, England and wales, Scotland and Northern Ireland: mid-2012 to mid-2016*. Office for National Statistics.

Or, Z., & Hakkinen, U. (2011). DRGs and quality: For better or worse? In R. Busse, A. Geissler, & W. Quentin (Eds.), *Diagnosis-Related Groups in Europe. Moving towards transparency, efficiency and quality in hospitals*. Open University Press.

Palmer, K. S., Agoritsas, T., Martin, D., Scott, T., Mulla, S. M., Miller, A. P., Agarwal, A., Bresnahan, A., Hazzan, A. A., Jeffery, R. A., Merglen, A., Negm, A., Siemieniuk, R. A., Bhatnagar, N., Dhalla, I. A., Lavis, J. N., You, J. J., Duckett, S. J., & Guyatt, G. H. (2014). Activity-based funding of hospitals and its impact on mortality, readmission, discharge destination, severity of illness, and volume of care: A systematic review and meta-analysis. *PLoS ONE*, *9*, e109975.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688–701.

Ryan, A. M., Krinsky, S., Kontopantelis, E., & Doran, T. (2016). Long-term evidence for the effect of pay-for-performance in primary care on mortality in the UK: A population study. *Lancet*, *388*, 268–274.

Street, A. D., O'Reilly, J., Ward, P., & Mason, A. (2011). DRG-based hospital payment and efficiency: Theory, evidence, and challenges. In R. Busse, A. Geissler, & W. Quentin (Eds.), *Diagnosis-Related Groups in Europe. Moving towards transparency, efficiency and quality in hospitals*. Open University Press.

Wen, H., Hockenberry, J. M., & Cummings, J. R. (2015). The effect of medical marijuana laws on adolescent and adult use of marijuana, alcohol, and other substances. *Journal of Health Economics*, *42*, 64–80.

## APPENDIX A: REGRESSION RESULTS

**TABLE A1** "Classic" difference-in-differences with year specific effects regression results

|  | Elective | Emergency |
|---|---|---|
| PbR*2003 | −0.3017 | 0.4226 |
|  | (0.5080) | (0.3288) |
| PbR*2004 | −0.0781 | 0.1387 |
|  | (0.5076) | (0.3290) |
| PbR*2005 | −0.1936 | −0.4398 |
|  | (0.5058) | (0.3275) |
| PbR*2006 | −0.2061 | −0.5068 |
|  | (0.4982) | (0.3224) |
| PbR*2007 | −1.3917*** | −0.5673* |
|  | (0.4966) | (0.3217) |
| PbR*2008 | −0.7701 | −1.0970*** |
|  | (0.4959) | (0.3219) |
| PbR*2009 | −0.2932 | −1.1448*** |
|  | (0.4963) | (0.3218) |
| PbR*2010 | −0.2604 | −0.9485*** |
|  | (0.4959) | (0.3216) |
| PbR*2011 | −0.4913 | −2.1277*** |
|  | (0.4970) | (0.3223) |
| PbR*2012 | −1.3544*** | −3.7521*** |
|  | (0.4974) | (0.3228) |
| PbR*2013 | −1.3816*** | −3.0236*** |
|  | (0.4975) | (0.3239) |
| Year dummies | YES | YES |
| Country | YES | YES |
| Sex | YES | YES |
| Age groups | YES | YES |
| Deprivation deciles | YES | YES |
| HRGs | YES | YES |
| N | 39,435 | 38,980 |
| R-squared | 0.4294 | 0.7044 |

*** and * indicate significance at 1% and 10%, respectively.

**TABLE A2** Difference-in-differences with time trend regression results

|  | Elective | Emergency |
|---|---|---|
| Time trend | −0.1852*** | −0.1864*** |
|  | (0.0168) | (0.0109) |
| Trend * England | −0.0159 | −0.0882* |
|  | (0.0718) | (0.0466) |
| Trend * PbR indicator | −0.0425 | −0.0868** |
|  | (0.0587) | (0.0381) |
| Country | YES | YES |
| Sex | YES | YES |

**TABLE A2** (Continued)

|  | Elective | Emergency |
| --- | --- | --- |
| Age groups | YES | YES |
| Deprivation deciles | YES | YES |
| HRGs | YES | YES |
| N | 39,435 | 38,980 |
| R-squared | 0.4285 | 0.7001 |

***, ** and * indicate significance at 1%, 5% and 10%, respectively.

**TABLE A3** ITS regression results

|  | Elective | Emergency |
| --- | --- | --- |
| Time trend | −0.1972*** | −0.2657*** |
|  | (0.0347) | (0.0317) |
| PbR indicator * trend | −0.0607* | −0.0960*** |
|  | (0.0337) | (0.0251) |
| Country | YES | YES |
| Sex | YES | YES |
| Age groups | YES | YES |
| Deprivation deciles | YES | YES |
| HRGs | YES | YES |
| N | 20,280 | 20,146 |
| R-squared | 0.7778 | 0.8263 |

*** and * indicate significance at 1% and 10%, respectively.