Monitor
Making the health sector work for patients

2016/17 National Tariff Payment System: A consultation notice

Annex B5: Evidence on efficiency for the 2016/17 national tariff

11 February 2016

Monitor publication code: IRCP 05/16

Contents

1. Why we modelled efficiency

1.1. Rationale for setting an efficiency factor

1. The NHS is a highly valued service and significant element of public spending. It accounts for around 7% of GDP.[1] It is important that we maximise value for money within the NHS, to ensure the very best services for patients and that taxpayers' money is well spent.

2. In other parts of the economy, we get value for money by shopping around between sellers competing for our business. We then select the product or service that gets us the most gain for the least cost. This puts pressure on firms to adopt the best production processes and most efficient technologies, and to pass reductions in costs on through lower prices. Those that don't are driven out of business by those that do.

3. In the NHS prices aren't set by this process between buyers and sellers, so this mechanism is not available. Instead we set many prices centrally, so we need to find a way of driving value for money. The efficiency factor that we apply to the prices we set is how we do this.

4. However, it is difficult to get the efficiency factor right, and there are problems if we get it wrong. If the efficiency factor is set too high, then prices are too low. This can mean that the business of providing healthcare can become unsustainable. If it is set too low, then prices are too high. This can mean that we fail to provide as much healthcare as possible for patients, wasting taxpayers' money. To avoid these problems it is important to use the best evidence available to help us set the efficiency factor.

1.2. What we have done in recent years

5. There are three questions we must answer to set an appropriate efficiency factor:

    a. How much efficiency has the service as a whole achieved in recent years? We call this trend efficiency.

    b. How much extra efficiency might be achieved by less efficient trusts catching up with more efficient trusts? We call this variation in efficiency.

    c. What other information might suggest the future will be different to our prediction?

[1] OECD (2015) Focus on health spending available at http://www.oecd.org/health/health-systems/Focus-Health-Spending-2015.pdf

This report uses econometric analysis to provide evidence that addresses the first two questions.

6. As part of the evidence base for the 2015/16 national tariff, Deloitte produced analysis to inform Monitor's judgement on the level of the efficiency factor. This comprised an econometric model and a supporting case study model.

7. The econometric model used data from 165 acute trusts between 2008/9 and 2012/13 to estimate the scope for efficiency in 2015/16.[2] The case study model estimated the reduction in costs from a range of efficiency initiatives applied to a stylised 'average' trust. Deloitte concluded that the most efficient trusts could become 1.2% to 1.3% more efficient a year, but the averagely-efficient trust could do much better (up to an additional 5.6% if it caught up with the top decile performer).

8. Prior to the 2015/16 tariff, a number of third-party publications fed into the decision on the efficiency factor, alongside trusts' Cost Improvement Plans (CIPs) and commissioners' Quality, Innovation, Productivity and Prevention (QIPP) initiatives.[3]

2. What we did

2.1. Efficiency concepts

9. Before describing our analysis, we clarify our terminology around efficiency. We estimate two measures of efficiency called trend efficiency and variation in efficiency, as shown in Figure 1.

10. **Trend efficiency** is the average sector-wide efficiency gain we observe over time. This could arise from new technologies, improved hospital processes or less efficient trusts catching up with more efficient ones. We estimate trend efficiency as a percentage reduction in costs over time that does not vary by trust. Given the importance of achieving value for money in the NHS, we think it reasonable to set an efficiency ask at least at the level of historical trend efficiency.

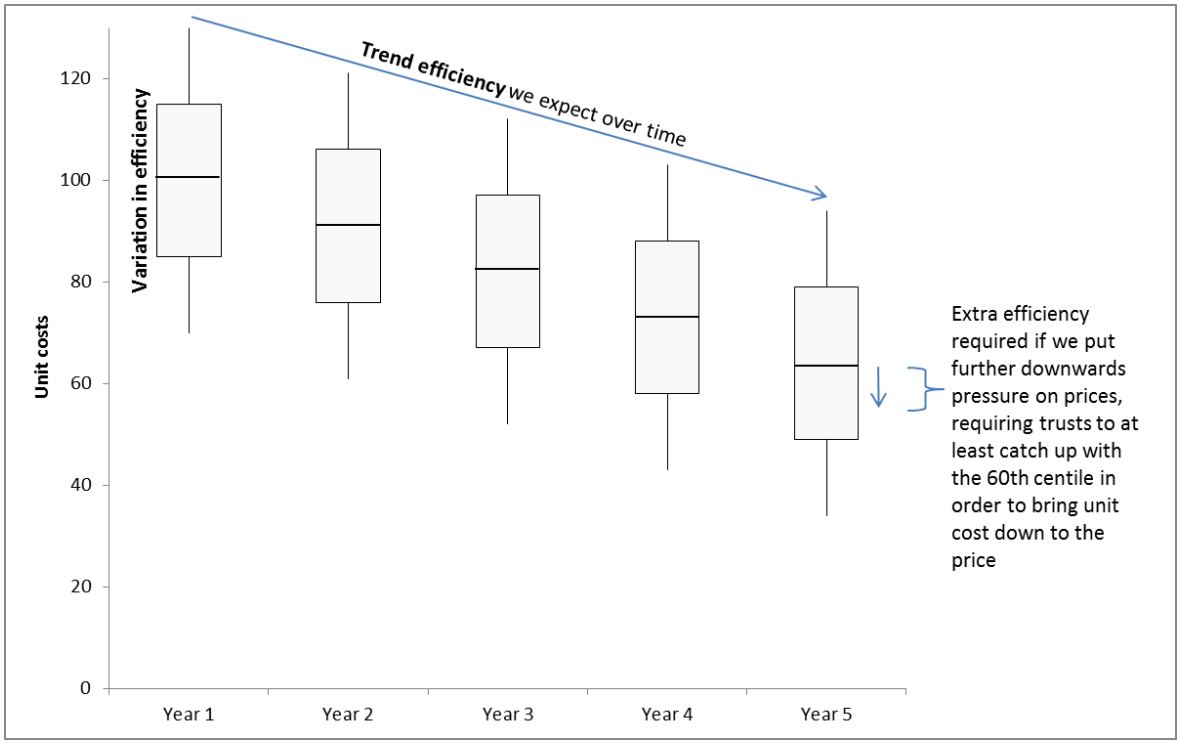
11. **Variation in efficiency** is the range of efficiency performance across trusts. This could arise from differences in take-up of technologies, or differences in hospital processes. We estimate variation in efficiency as a percentage

[2] The decision to use data from acute trusts, rather than non-acute trusts or private providers, was made on the basis of data availability and quality.

[3] McKinsey (2009) Achieving a World Class Productivity in the NHS 2009/10 – 2013/14: Detailing the Size of the Opportunity available at http://www.nhshistory.net/mckinsey%20report.pdf
Monitor (2013) Closing the NHS funding gap: how to get better value healthcare for patients https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/284044/ClosingTheGap091013.pdf

difference in costs from the average trust, which does not change over time but does vary by trust. We use this to inform our understanding of what a reasonable efficiency ask, over and above trend efficiency, would be based on the potential for less efficient trusts to catch up with more efficient trusts.

Figure 1: Trend efficiency and variation in efficiency



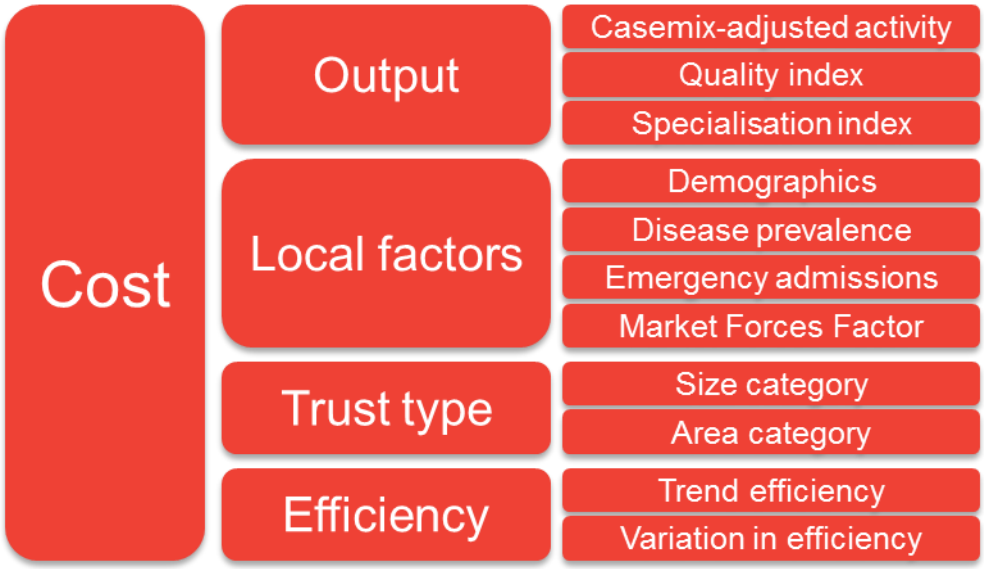
2.2. Our approach

12. The first step in estimating these efficiency concepts is to work out what drives trusts' costs. After controlling for a wide range of factors that drive trusts' costs, we interpret any left over changes in costs over time as trend efficiency and any left over differences in costs between trusts as variation in efficiency. We look at four types of factors that drive trusts' costs:

- **The healthcare that the trust provides.** This includes: casemix-adjusted hospital activity, the degree of specialisation in trusts, and the quality of service based on patient satisfaction.
- **Local drivers of costs that the trust cannot control.** This includes: local disease prevalence, demographics of the patient population, the proportion of emergency admissions, and the market forces factor.
- **Trust type.** This includes: the trust's categorisation (small, teaching, specialist etc) and former Strategic Health Authority (SHA) region.
- **Efficiency.** This includes: trend efficiency and variation in efficiency.

13. Figure 2 shows how these factors are related to costs.

Figure 2: Model specification



14. We use statistical methods to unpick the impact of each factor on trusts' costs.[4] This gives us an estimate of the effect of each driver on costs.[5] With the estimated impacts in hand, we remove their influence from trusts' costs and interpret the service-wide reduction in costs over time as trend efficiency, and the long-term differences between trusts' costs as variation in efficiency.

3. What we found

3.1. Efficiency results

Table 1: Efficiency estimates

	Estimate
Trend efficiency:	1.4%
Variation in efficiency:	
median to 60th centile	2.0%
median to 70th centile	3.6%
median to 80th centile	5.6%
median to 90th centile	7.6%

Source: Monitor analysis

15. Our analysis tells us that trusts become 1.4% more efficient each year on average. Around this trend we estimate that there is substantial variation in

[4] We adjusted trusts' costs for inflation in healthcare prices, to ensure we measured them in real terms. We used the inflation cost uplift in the national tariff as our measure of inflation.
[5] Our statistical results are in Table 4 on page 25.

efficiency. For example, in order for the average (median) provider to catch up to the 60th centile, it would need to become 2% more efficient on top of this 1.4% (see Table 1). In order for the average provider to catch up to the 90th centile, it would need to become 7.6% more efficient on top of this 1.4%.

3.2. Sensitivity checks

16. We checked how robust our results were by undertaking a number of sensitivity checks. Details can be found in section 4.3 below. Our results are robust to these checks.

4. What it means

4.1. Recommendation for the 2016/17 efficiency factor

17. Our estimate of trend efficiency tells us that it is reasonable for the efficiency factor to be at least 1.4%. At an efficiency factor of 1.4%, the sector would need to continue to achieve efficiencies at the rate it has managed over the last six years in order to maintain its financial position.

18. What our estimates say about how much larger the efficiency factor could be requires careful interpretation, however. This is because it is clear that catch-up opportunities exist, but it is less clear that these opportunities can be realised within the one-year timeframe of the 2016/17 national tariff. So to reach a recommendation on this, we considered four additional pieces of evidence:

- Previous analysis undertaken for Monitor by Deloitte, which suggested one year savings of between 1% and 1.4% are possible if the average trust were to catch up to better-performing trusts.
- Analysis undertaken by the Centre for Health Economics at the University of York of trust-level productivity, which highlights the consistency in trust productivity rankings between 2010/11 and 2012/13.[6] Consistent rankings suggest there is not much movement in the variation in efficiency between trusts.
- Health Foundation analysis of trust-level productivity,[7] which shows very little evidence of less productive trusts catching up with more productive trusts.

[6] Aragon, Castelli, Gaughan (2015) Hospital Trusts Productivity in the English NHS: Uncovering Possible Drivers of Productivity Variations *Centre for Health Economics Research Paper 117* https://www.york.ac.uk/media/che/documents/papers/researchpapers/CHERP117_hospital_trusts_productivity_English_NHS.pdf

[7] Lafond, Charlesworth, Roberts (2015) Hospital finances and productivity: in a critical condition? *The Health Foundation* http://www.health.org.uk/publication/hospital-finances-and-productivity-critical-condition

This suggests that the differences in efficiency between trusts tend not to lead to rapid improvement in less efficient trusts.

- Analysis by Monitor, based on the Centre for Health Economics and Health Foundation methods, which supports the idea that there has not been very much narrowing in the distribution of trust efficiency.

19. The results of our analysis support a range for the efficiency factor of 1.5% to 2.5%, namely trend efficiency of 1.4% plus catch-up of up to 1.1%.

20. Given the scale of financial challenge that we face in 2016/17, and the current state of provider finances, we recommend that an efficiency factor in the region of 2% is appropriate. This is towards the top end of what has been achieved in recent years and implies the sector needs to increase its efficiency gains by almost 50% above long term trend.

Annex 1: Technical details

1. Background

1.1. Purpose

1.1.1. Background to the efficiency factor

The NHS is a publicly funded system. If it fails to provide healthcare at maximum value for money, that means either more taxpayer funding is needed or less healthcare is provided. It is therefore important to maximise efficiency.[8]

Price-setting is one of the ways that we promote efficiency. This allows us to align the incentives of providers to provide efficient care, and commissioners to commission the right care. The efficiency factor is our lever for promoting efficiency through the price level. It acts as both an incentive and a signal.

- As an incentive it drives efficiency. In other markets we expect producers to pursue technological or process improvements that result in cost reductions so that they can lower prices, increase market share and extract profit. If producers do not keep up with the improvements of their competitors they lose market share and fail. This is less likely to occur in the NHS for two reasons. Firstly, NHS providers do not have the same profit-maximising incentives as other sectors. Secondly, even when costs are lowered, commissioners or patients cannot easily switch towards more efficient providers that pass on the benefits. This is for a number of reasons, including the need for providers to be located sufficiently close to their population and because a large number of services are priced nationally.

- As a signal it informs decisions about resource allocation. For commissioners it directly affects the price, which is the cost to them of procuring a service. With these prices, commissioners can decide how best to allocate resources locally. More generally, the efficiency factor represents our judgement of the improvement in efficiency the NHS can and should make. This can be used to aid planning at both the national and local levels.

1.1.2. Problems due to bad decisions on the efficiency factor

Setting the efficiency factor at the right level is challenging. Problems can arise from too high or low a factor:

- **If it is set too high** the business of providing healthcare can become increasingly unsustainable as prices are pushed further below costs. This

[8] In economic terms, the efficiency we mean here is productive efficiency, sometimes known as technical efficiency.

could mean we risk incentivising inappropriate cuts to costs that reduce safety, quality or access. Additionally, the more prices diverge from costs the more misleading the signal to commissioners. For example: this could lead commissioners to regard acute care as a relatively cheap way to provide care compared with community-based prevention schemes, and potentially lower overall system efficiency. Furthermore it is possible that the incentive for providers to reduce costs is actually weakened if the efficiency factor is considered to be infeasible.

- **If it is set too low** the commissioning of local services is needlessly restricted due to high prices, leading to less healthcare delivered to patients. The incentives for providers to realise potential efficiencies are blunted, and taxpayers' money may be wasted.

To avoid setting the efficiency factor too high or too low, we look at a wide range of information when deciding what the level of the efficiency factor should be. To inform this complex regulatory judgement we have updated and amended the econometric evidence Deloitte produced for us last year. Similar to last year, we do not believe data quality is good enough outside of the acute sector to reliably estimate efficiency through econometric means. We therefore restrict our analysis to secondary care.

1.2. What work has been done

1.2.1. Deloitte model

For the 2015/16 national tariff decision on the efficiency factor Deloitte conducted analysis on the scope for efficiency in the NHS.[9] This consisted of statistical analysis plus a case study.

- The statistical analysis estimated the scope for efficiency in 2015/16 using trust-level data from 2008/09 to 2012/13. By controlling for differences in activity, casemix, quality and other local cost drivers, it estimated that the level of efficiency of the most efficient trust increased at a rate of 1.2% to 1.3% a year. Additionally, the analysis estimated that the 90th centile trust was 5% to 5.6% more efficient than the median trust, while the 60th centile trust was approximately 1% more efficient than the median.

- The case study examined the effect that various efficiency levers could have on the costs of a notional 'average' hospital. These levers included increasing the day case rate, shortening length of stay and reducing use of agency staff. It suggested that an average trust could increase its efficiency by between 1% and 1.4% within a year through the efficiency levers identified.

[9] Deloitte (2014) Evidence for the 2015/16 national tariff efficiency factor https://www.gov.uk/government/consultations/nhs-national-tariff-payment-system-201516-engagement-documents

Our analysis builds on the Deloitte model. We use additional data for 2013/14 and make a small number of improvements to the method (explained further below).

1.2.2. Health Foundation model

In their report "Hospital finances and productivity: in a critical condition?" the Health Foundation presented an analysis of efficiency based on the Deloitte work.[10] Using a different set of factors driving cost to those used by Deloitte, the Health Foundation found that annual efficiency improvement averaged 0.4% between 2009/10 and 2013/14. Although this is substantially lower than the Deloitte estimate, the 95% confidence intervals of the two estimates overlap (that is, the intervals within which, with 95% confidence, we can say that the trust estimate lies).

As to catch-up efficiencies the Health Foundation's analysis suggests there is very little change in the distribution of productivity over time.[11]

1.2.3. Other NHS benchmarking analyses

The Centre for Health Economics[12] at the University of York and the Office of National Statistics (ONS) have both produced estimates of NHS productivity, based on the ratio of outputs to inputs. The CHE estimates suggest that productivity in the English NHS grew at 0.8% a year between 2004/05 and 2012/13.[13] This masked substantial volatility, however, with growth ranging from -2.3% to +4.4%. The ONS estimates suggest that productivity growth for the NHS across the UK averaged 1% a year in the same period. The ONS productivity growth estimates are less volatile than those from CHE, ranging from -1.3% to +3.5%.[14]

2. Method

2.1. Econometric models

2.1.1. Econometric benchmarking techniques

Econometrics is the application of statistics methods to economic data. In our application, it allows us to compare trusts' costs while controlling for a large number

---

[10] Lafond, Charlesworth, Roberts (2015) Hospital finances and productivity: in a critical condition? *The Health Foundation* http://www.health.org.uk/publication/hospital-finances-and-productivity-critical-condition

[11] Though productivity is a different concept to technical efficiency, the Health Foundation use real costs as the 'input' variable in their productivity index, so in this instance the two concepts are similar.

[12] Bojke, Castelli, Grašič, Street (2015) Productivity of the English NHS: 2012/13 Update *Centre for Health Economics Research Paper 110* https://www.york.ac.uk/media/che/documents/papers/researchpapers/CHERP110_NHS_productivity_update_2012-13.pdf

[13] Monitor calculations based on published CHE productivity figures, using cost-weighted activity

[14] ONS (2015) Public Service Productivity Estimates: Healthcare, 2012 http://www.ons.gov.uk/ons/dcp171766_393405.pdf

of cost drivers. Differences in costs that cannot be explained by these cost drivers using the econometric method are interpreted as efficiency.

Two econometric methods are typically used:

- Corrected Ordinary Least Squares (COLS)
- Stochastic Frontier Analysis (SFA)

Corrected Ordinary Least Squares

In COLS, total cost is regressed on a cost function. It is assumed that the specification accurately captures the underlying cost process, and so the residuals can be interpreted as differences in efficiency. The firm with the smallest residual is said to be operating at the efficient frontier, and all other firms' efficiency is measured relative to their distance from this frontier. For a standard cost function, the COLS regression is:

$$C_i = \alpha + \beta_1 w_i + \beta_2 y_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

with cost C_i, input price w_i, output y_i, ε_i is interpreted as firm *i*'s efficiency.

Stochastic Frontier Analysis

As with COLS, SFA assumes that the process being modelled is a cost function. But SFA does not interpret the whole residual as efficiency. Instead SFA models efficiency as an additional statistical error with a specific probability distribution. Maximum Likelihood estimation is then used to estimate firm-level efficiency. Typically, the efficiency term is restricted to be strictly positive, for instance by using the half-normal or truncated-normal distribution, and is interpreted as a distance from the efficient frontier. For a standard cost function, the SFA model is:

$$C_i = \alpha + \beta_1 w_i + \beta_2 y_i + u_i + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$u_i \sim N^+(0, \gamma^2)$$

with cost C_i, input price w_i, output y_i, and idiosyncratic error ε_i, u_i is interpreted as firm *i*'s efficiency.

2.1.2. Panel data techniques

Because our data consists of repeated observations of trusts over time, the statistical inference drawn from comparing two observations is likely to depend on whether the

two observations are from two trusts in the same year, from two years for the same trust, or from two different trusts in two different years. Techniques to isolate these effects are:

- Fixed Effects (FE)
- Random Effects (RE)
- the Mundlak transformation

Fixed Effects

Under FE, each trust is modelled as having its own intercept.

$$y_{it} = \alpha_i + x'_{it}\beta + \varepsilon_{it}$$

$$\varepsilon_{it} \sim N(0, \sigma^2)$$

Typically this is estimated by subtracting the within-trust mean from the dependent and each independent variable and using Ordinary Least Squares (OLS) on the transformed variables, in order to remove the trust-specific effect. For our purposes, FE has the drawback of subsuming any time-invariant effects in the data into the fixed effects (the α_i trust-level, time invariant effects) rendering them uninterpretable. As we need to be able to compare trusts to benchmark them, this is an insurmountable obstacle. So FE is not appropriate for our purposes.

Random Effects

In RE, each trust is modelled with an additional trust-specific error term. As with FE, this results in each trust having its own specific intercept. But instead of being 'fixed' to the sample of data, these intercepts are treated as 'random' draws from a normal distribution across the population.[15]

$$y_{it} = \alpha + x'_{it}\beta + u_i + \varepsilon_{it}$$

$$\varepsilon_{it} \sim N(0, \sigma^2)$$

$$u_i \sim N(0, \gamma^2)$$

This RE model can be estimated by Generalised Least Squares (GLS). Unlike the FE model, the RE estimates retain information in the data that varies between trusts but not over time. RE estimates are therefore amenable to econometric

[15] We note the difference between RE and SFA is the choice of distribution for the error term. Under RE this is a normal distribution and therefore symmetric. Under SFA this is a truncated normal distribution and asymmetric. Thus the modelling difference between RE and SFA can be considered an implicit belief about the underlying efficiency distribution: under SFA there is a long tail of increasingly inefficient trusts, whereas under RE the distribution of the most efficient trusts is mirrored in the distribution of the least efficient trusts.

benchmarking. However, RE estimates are biased if the error term *u* is not independent of the covariates *x*.

Where this is the case (as it is in our application), we can use the Mundlak transformation to return unbiased estimates:[16]

Mundlak transformation

The Mundlak transformation takes each independent variable and breaks the effect on the dependent variable in two: the effect related to a change within a single trust over time, and the effect related to a difference between trusts. In practice, each independent variable is split into its mean over time and its deviation from its mean over time. The number of independent variables therefore doubles, and the separate coefficients for the mean and the deviation-from-mean variables allow the effects of between- and within-variation to differ.[17]

$$y_{it} = \alpha + \bar{x}'_{it}\beta_{between} + \tilde{x}'_{it}\beta_{within} + u_i + \varepsilon_{it}$$

$$\bar{x}_{it} = \frac{\sum_{t=1}^{T} x_i}{n}$$

$$\tilde{x}_{it} = x_{it} - \frac{\sum_{t=1}^{T} x_i}{n}$$

$$\varepsilon_{it} \sim N(0, \sigma^2)$$

$$u_i \sim N(0, \gamma^2)$$

2.1.3. Efficiency concepts

The sector would be operating at maximum efficiency if all providers were providing a given set of services at the lowest possible cost. For a notional unit of healthcare, efficiency specifies how low the price of this unit is.

There are two efficiency metrics we are interested in. Firstly, we are interested to know the extent to which the provider sector is able to increase its efficiency over time. We call this trend efficiency. In our econometric models this is estimated by the coefficient on the trend term. We use this estimate to predict the trend of efficiency in the future, from which we can infer how much average costs are likely to fall by in the coming year due to efficiency improvements, all else equal.

[16] Applying the Mundlak transformation in an SFA context has previously been suggested by Faris, Filippini, Kuenzle (2003) Unobserved heterogeneity in stochastic cost frontier models: a comparative analysis *University of Lugano Department of Economics Working Paper 03-11*

[17] In the context of this paper, this could be because the effect on costs of one trust being twice the size of another trust differs from the effect of the first trust doubling in size.

Secondly, we are interested to know the extent of variation in efficiency across trusts in the sector. We call this variation in efficiency. Understanding this allows us to set an efficiency factor that is challenging for most trusts but still within the range of variation in efficiency that we see; put differently, this allows us to judge what is a stretching but achievable target.

These concepts of efficiency are related to frontier shift and catch-up efficiency, the terms that Deloitte used in their analysis last year. The efficiency frontier is the maximum possible efficiency a provider can attain. Frontier shift describes how this increases over time, as new technologies and processes make lower production costs possible. In practice, this efficient frontier and its change over time is inferred from those providers with the lowest unit costs. Catch-up efficiency describes the distance providers are from this frontier. Though this conceptual framework is appealing, we think the terms **trend efficiency** and **variation in efficiency** better describe what we are estimating:

- **Trend efficiency** (the trend term in our estimation) incorporates the sector-level reduction in unit costs over time, not just the reduction in unit costs of the most-efficient trusts, or the conceptual frontier itself. This means that the trend term also includes historical catch-up efficiency gains.
- **Variation in efficiency** (the trust-level efficiency effects in either the RE or SFA model) estimate the difference in unit costs on average across the whole time period. Some portion of these estimated opportunities for efficiency improvement may be eroded by trusts catching up within the time span. Without making much stronger assumptions in the model it is not possible to model this catch-up.

2.2. Models

2.2.1. Baseline models

Our baseline models are the RE and SFA specifications with the Mundlak transformation. The RE specification assumes that the efficiency effect follows a time-invariant normal distribution across trusts and is estimated by GLS with heteroscedastic-robust standard errors. The SFA specification assumes that the efficiency effect follows a time-invariant truncated normal distribution across trusts and is estimated by MLE. Both specifications use the same dependent and independent variables.

The dependent variable is the natural logarithm of costs deflated by the inflation cost uplift in each year. This is regressed on variables which we group into four categories:

- **The healthcare that the trust produces.** This includes: the natural log of casemix-adjusted hospital activity, the degree of specialisation in trusts and the quality of service based on patient satisfaction.
- **Local drivers of costs that the trust cannot control.** This includes: local disease prevalence, demographics of the patient population, the proportion of emergency admissions and the market forces factor (MFF).
- **Trust type.** This includes: the trust's categorisation (small, teaching, specialist etc) and former SHA region.
- **Efficiency.** This includes: trend efficiency and variation in efficiency.

These variables are described in Section 3 below.

2.2.2. Sensitivity check models

We check the sensitivity of our results to changes in modelling assumptions. These checks can be grouped into sampling changes and variable changes.

We test the effect of sampling changes by running our analysis with specific subsamples excluded. The changes we make are:

1. to exclude specialist, multi-service and teaching hospitals
2. to exclude trusts that may have less accurate coding. When an activity code is invalid, using as a proxy those trusts with a greater proportion of activity that is not recognised by grouping software and is therefore coded into the UZ01Z currency[18]
3. to restrict the sample to those trusts that we observe in every year.

We test the effect of changes to the variables in our models in the following ways:

4. Including the number of sites the trust operates across
5. Including the number of acute sites the trust operates across
6. Including the proportion of inpatients from a lower super output area designated as 'urban' by the ONS
7. Deflating costs by the MFF
8. Replacing the health-specific measure of inflation with the GDP deflator.

[18] Note, this does not capture activity that is incorrectly coded but falsely recognised under a valid currency.

2.3. Model interpretation

The key estimates of interest in our modelling are of trend efficiency, and variation in efficiency.

Trend efficiency is simply the estimated coefficient on the time trend term. Because the dependent variable is in natural logarithms, this coefficient is the estimated average annual change in real costs measured in percentage terms.

Variation in efficiency is extracted from the efficiency effects themselves. The differences in the efficiency effects are differences in real unit costs between trusts in percentage terms. This ensures that our estimates of the two concepts of efficiency are presented in the same units, namely percentage deviation in real unit costs. We present the variation in efficiency estimates in terms of percentage difference in real costs between the median (50th centile) provider and the 60th, 70th, 80th and 90th centile providers.

3. Data

3.1. Dataset description

We use an unbalanced panel dataset covering 170 trusts across the time period 2008/09 to 2013/14.[19] Although our panel of data is unbalanced (that is, has different numbers of trusts across time) entrants, exits and mergers should be related to the level of efficiency of a trust. If so, then excluding them from our sample (to balance it up) would give us a skewed perspective of efficiency in the whole system.[20]

3.2. Variables

3.2.1. Total costs

Total costs are taken from reference costs. We are modelling the efficiency of the acute sector, so we exclude costs for mental health and community services. We deflate costs by the national tariff's inflation cost uplift.

[19] We do not remove statistically influential observations. This is for three reasons: for dummy variables that relate to a small number of trusts, each trust is likely to exert a relatively large influence on the dummy coefficient; the Mundlak transformation increases the number of covariates substantially, and each additional covariate increases the chance of a type 1 error; those trusts that do exert a significant influence may be of particular interest to us in the stochastic frontier analysis, because their influence may be because they are close to the efficient frontier. For these reasons, we prefer to deal with issues of statistical influence by varying the sample in our sensitivity analyses.

[20] We test this decision by re-running our regressions with a reduced balanced panel (that is, using only trusts reporting data for the full time period) as a sensitivity test.

3.2.2. Hospital output

Output is the healthcare that hospitals produce, for which they incur costs. We use a number of output variables to control for the effects of scale and scope in trusts' production costs.

- We measure activity using the reference cost dataset, with casemix adjustment by cost weight. Each currency's cost weight is calculated each year as the currency's national average cost relative to the national average cost across all currencies.[21]

- We include an index for the quality of care that trusts produce to account for the additional costs this extra value may require. We extract the percentage of respondents that answered "strongly agree" (or "no" in the case of seeing errors or near misses in the last month) from the relevant questions in the NHS Staff Survey.[22] We then use principal component analysis to extract an underlying index of service quality. The questions, and their correlation with the first principal component, are reported in Table 2.[23]

[21] The only change we have made to this calculation is for the currencies that appear in both normal and excess bed days categories (both elective and non-elective). We regroup the costs and activity for these currencies, so that the costs reflect the full cost of the episode (inlier cost plus excess bed day cost) and the activity number reflects the numbers of episodes (number of inlier episodes, as every excess bed day requires a preceding inlier episode).

[22] The NHS Staff Survey is an annual questionnaire completed by staff in NHS providers.

[23] The first principal component captures 63% of the variation in the trusts' responses to the identified questions. This is slightly more than the corresponding principal component used in the Deloitte analysis previously (63% versus 61%).

Table 2: NHS Staff Survey questions and correlation with 1st principal component

Question:	Correlation with the 1st principal component
Training keeps me up to date with professional standards	0.72
Effective communication between senior management and staff	0.87
Senior managers where I work are committed to patient care	0.92
I am satisfied about the quality of patient care I give	0.87
Care of patients is my trust's top priority	0.93
I am able to deliver the patient care I aspire to	0.85
I cannot meet all the conflicting demands on my time at work	-0.51
I have adequate supplies, materials and equipment	0.87
There are enough staff at this trust for me to do my job properly	0.87
If a friend/relative needed treatment, I would be happy with the quality	0.79
Have you seen any errors, near misses or incidents that could have hurt a patient in the last month	0.23

Source: Monitor analysis

- Costs may be affected by the degree of specialisation in a trust. To control for this, we construct a specialisation index based on the method employed by Deloitte in previous work for us. We aggregate casemix adjusted activity to the HRG chapter level for each trust. The index measures the divergence of any individual trust HRG activity profile from the national aggregate profile.

$$ITI_i = \sum_{h=1}^{H} p_{ih} \ln\left(\frac{p_{ih}}{\varphi_h}\right)$$

where p_{ih} is the proportion of activity in chapter *h* for provider *i* and φ_h is the national average activity in chapter *h*.

3.2.3. Uncontrollable cost drivers

In addition to the cost of healthcare that trusts provide, our modelling needs to account for cost drivers outside of their control. This is necessary to avoid attributing changes in the environment within which trusts operate to our estimate of trend efficiency, or differences between trusts in the nature of uncontrollable cost drivers to our estimates of variation in efficiency.

- We control for the effect on costs of local demographics using Hospital Episode Statistics (HES) data. For the inpatient population we calculate the annual percentage who are:

- female
- under the age of 19
- over the age of 75
- from an ethnic minority background.

Our cut-offs for age categories align with the age split used in the national tariff. In addition to these variables, we calculate the average Index of Multiple Deprivation (IMD) overall score across all inpatients for each trust.[24]

- We control for the health of the local area using the disease prevalence reported by GP surgeries through the Quality and Outcome Framework (QOF). We take the disease prevalence across 20 disease categories for each GP surgery and map these proportions to acute trusts based on the proportion of inpatients in each trust coming from each surgery. We then estimate the principal components of disease prevalence, and follow Deloitte in using the first two components in our subsequent regression analysis. Due to changes in the scope of QOF collection we are unable to use disease prevalence for four conditions: the depression 1 indicator, osteoporosis, peripheral arterial disease and cardiovascular disease. The diseases that we consider, and their correlation with the first two principal components, are reported in Table 3.[25]
- We account for the effect of local market forces on costs using the underlying index of the MFF. This controls for unavoidable area-level variation in the cost of staff, land and buildings.

Table 3: Local disease prevalence and correlation with 1st and 2nd principal components

Disease	Correlation with first principal component	Correlation with second principal component
Coronary heart disease	0.93	-0.22
Stroke or Transient Ischaemic Attacks	0.94	-0.08
Hypertension	0.95	-0.01

[24] We prefer the IMD overall score to the alternative health domain because it uses a broader range of data related to deprivation, whereas the health domain is constructed solely from health measures including Years of Potential Life Lost, mood and anxiety disorders, morbidity measures and illness and disability ratios, and we prefer to account for local health through the local disease prevalence.

[25] The first principal component explains 67% of the variation in disease prevalence, the second principal component explains 8%. Together, they explain slightly more of the variation in disease prevalence than the corresponding principal components used in the Deloitte analysis previously (75% versus 70%).

Disease	Correlation with first principal component	Correlation with second principal component
Diabetes	0.80	0.20
Chronic Obstructive Pulmonary Disease	0.84	-0.10
Epilepsy	0.93	-0.11
Hypothyroidism	0.84	0.17
Cancer	0.77	0.47
Mental health	0.38	0.19
Asthma	0.90	-0.07
Heart Failure	0.91	-0.19
Heart failure due to LVD	0.70	-0.49
Palliative Care	0.58	0.58
Dementia	0.83	0.35
Depression	0.47	-0.67
Chronic Kidney Disease	0.82	-0.07
Atrial Fibrillation	0.88	0.14
Obesity	0.79	-0.21
Learning Disabilities	0.82	0.15
Smoking	0.97	-0.01

Source: Monitor analysis

3.2.4. Trust type

To ensure we are fully controlling for differences between trusts associated with the local area or other trust characteristics, we include a set of 1/0 indicator variables (dummies).

- The region in which the trust is located is represented by a geographical dummy. Each dummy corresponds to a former SHA area. The regions are:
    - East Midlands
    - East
    - London
    - South West
    - South East
    - North West
    - North East

- South
- South Central
- West Midlands
- Yorkshire and the Humber.

- We use the trust type dummies as defined in the Estates Return Information Collection.[26] Categories include:
  - Small acute trust
  - Medium acute trust
  - Large acute trust
  - Teaching acute trust
  - Multi-speciality acute trust
  - Specialist trust

4. Results

4.1. Regression results

The results from the headline RE and SFA models are presented in Table 4 below. Because the models use the Mundlak transformation, we report the estimates based on the within-trust variation and between-trust variation separately, in columns side by side.

Our primary aim is to control for as many relevant factors as possible, to obtain the best possible (unbiased) estimates of 'trend' and 'variation' (see Section 3.3). This can make interpretation of the other coefficients in our model difficult, but we prefer to over-control for cost drivers and face multi-collinearity than under-control and suffer omitted variable bias. Nevertheless, we make the following observations:

1. Many of our estimated coefficients are similar to the estimates produced for us by Deloitte last year. The significant coefficients on the underlying MFF index, the percentage of inpatients who are female, the percentage of inpatients who are over 75 years old, the first principal component of the disease index, and the dummies relating to multiservice, teaching and specialist (SFA only) trusts for the between-trust variable versions are all close to last year's coefficient estimates. A substantial number of the coefficients that are not statistically significant in our analysis were not significant last year either, with the exception

[26] Available at http://hefs.hscic.gov.uk/ERIC.asp

of the percentage of inpatients from ethnic minorities and a number of the trust type dummies. Two variables that are significant in our analysis were insignificant last year: the percentage of emergency admissions and the quality index. However, when the between-trust and within-trust variables are combined (undoing the Mundlak transformation) these fall back to their previous level of insignificance.

2. The estimated coefficients on the effect of activity on cost (0.95 between trusts and 0.68 within trusts) suggest that the result last year (overall elasticity around 0.8) masked a difference between the within-trust elasticity and the between-trust elasticity. Essentially, this implies the economies of scale we can see across different trusts are much more modest than the economies of scale an individual trust can observe over time.

3. The results of the RE and SFA models are very similar, despite assuming different distributions for the efficiency term and using different estimation techniques (GLS and MLE respectively). The Deloitte analysis last year found the same similarity between the two models. Two possible explanations are:

    a. The effect of the efficiency terms are small, so the specific distribution that we assume for them exerts little influence on the estimates of the coefficients.

    b. The profile of efficiency across providers is symmetric, so SFA selects a truncation point on the extreme end of the normal distribution, and there is little difference between distributions used under SFA (truncated normal) and RE (normal).

4. There are many more statistically significant coefficients for the variables based on the between-trust variation than those based on the within-trust variation. Our model specification explains more of the variance related to differences in trust costs than changes within a trust's costs over time.

5. The estimated coefficients in the within-trust columns are noticeably different from those in the between-trust columns. This supports our suspicion that the effect of a cost driver when we consider differences between trusts differs markedly from its effect when we consider changes for an individual trust over time. This is not surprising, as we have observations across many more trusts than time periods. Additionally, the Hausman test for the RE model supports our use of the Mundlak transformation.

Table 4: Regression results

	RE with Mundlak transformation				SFA with Mundlak transformation			
	Between		Within		Between		Within	
Variables	Coef	S.E	Coef	S.E	Coef	S.E	Coef	S.E
Ln(Activity)	0.95***	(0.033)	0.68***	(0.066)	0.95***	(0.019)	0.68***	(0.021)
MFF	1.1***	(0.297)	-1.1	(0.665)	1.1***	(0.298)	-1.1**	(0.500)
Emergency admissions %	-0.005**	(0.002)	-0.00027	(0.001)	-0.005***	(0.001)	-0.00027	(0.001)
Specialisation index	0.026	(0.019)	-0.0048	(0.012)	0.025	(0.019)	-0.0048	(0.010)
Under 18 %	-0.00012	(0.001)	0.0012	(0.002)	-0.00012	(0.001)	0.0012	(0.002)
Over 75 %	-0.0072**	(0.003)	0.0026	(0.003)	-0.0072***	(0.002)	0.0026	(0.002)
Female %	-0.0079***	(0.001)	0.004	(0.003)	-0.0079***	(0.001)	0.0040**	(0.002)
Ethnic minority %	0.00099	(0.001)	0.000056	(0.002)	0.00098	(0.001)	0.000055	(0.002)
Disease index1	0.0079*	(0.005)	0.025***	(0.005)	0.0079	(0.005)	0.025***	(0.004)
Disease index2	-0.00067	(0.012)	0.004	(0.004)	-0.00072	(0.013)	0.0040	(0.003)
IMD	-0.0023	(0.007)	0.00022	(0.011)	-0.0023	(0.006)	0.00022	(0.009)
Quality index	-0.0053*	(0.003)			-0.0054	(0.003)		
Medium acute	-0.031	(0.023)	0.0044	(0.017)	-0.031	(0.020)	0.0044	(0.014)
Multi-service acute	-0.1*	(0.053)	-0.059**	(0.026)	-0.1***	(0.038)	-0.059	(0.056)
Small acute	-0.034	(0.029)	-0.0073	(0.021)	-0.034	(0.024)	-0.0073	(0.016)
Specialist acute	-0.12	(0.093)			-0.12**	(0.051)		
Teaching acute	.041*	(0.021)	-0.013*	(0.008)	0.04*	(0.022)	-0.013	(0.025)
East Midlands	0.0092	(0.027)			0.0091	(0.031)		
East	-0.042*	(0.024)			-0.042	(0.027)		
London	0.017	(0.053)			0.017	(0.048)		
North East	0.015	(0.033)			0.015	(0.036)		
North West	0.0012	(0.020)			0.0011	(0.023)		
South Central	-0.0076	(0.031)			-0.0074	(0.034)		
South East	-0.02	(0.032)			-0.02	(0.034)		
South	0.19**	(0.074)			0.19*	(0.101)		
South West	-0.036	(0.022)			-0.036	(0.025)		
West Midlands	0.0095	(0.026)			0.0092	(0.025)		
Trend			-0.014***	(0.005)			-0.014***	(0.003)
Constant	4.9***	(0.686)			4.7***	(0.444)		
Obs (Groups)	981 (170)				981 (170)			
Tests of joint significance	$\chi^2(42)$	60431	P<0.01		$\chi^2(43)$	17330	P<0.01	
Hausman test	$\chi^2(17)$	0.000	P=1.000					

Source: Monitor analysis

4.2. Efficiency results

The efficiency results are the main focus of our analysis and are presented in Table 5. Our modelling estimates that efficiency has increased by 1.4% a year on average between 2008/09 and 2013/14. The predicted variations in efficiency suggest that the top decile provider is 7.6% more efficient than the median provider, while the 60th centile provider is 2% more efficient than the median provider.

The results are similar whether RE or SFA is used, despite the difference in estimation techniques. The RE variation in efficiency is slightly wider than that of SFA.

Compared with the analysis last year, we note that our estimate of trend efficiency is slightly higher. This difference is due to the combination of an additional year's data and the changes in method.

In both models the estimates of trend efficiency growth are highly statistically significant, although there has still been substantial variation over time. The 95% confidence interval of the trend efficiency estimates ranges from 0.5% to 2.4% in the RE model and 0.8% to 2.1% for the SFA model.

Table 5: SFA and RE efficiency results

	Random Effects	Stochastic Frontier Analysis
Trend efficiency:	1.4%	1.4%
95% confidence interval:	(0.5% to 2.4%)	(0.8% to 2.1%)
Variation in efficiency:		
50th to 60th centile	2.0%	2.0%
50th to 70th centile	3.6%	3.6%
50th to 80th centile	5.6%	5.5%
50th to 90th centile	7.6%	7.5%

Source: Monitor analysis

4.3. Sensitivity checks

The results of the sensitivity checks are presented in Table 6. Given the similarity between the RE and SFA headline models, we do not apply the sensitivity checks to the SFA model.

We find that the headline efficiency estimates are robust to changes in the trust population. Excluding specialist and/or teaching trusts increases the estimate of trend efficiency slightly, while excluding multi-service trusts and those that code a higher proportion of activity to the UZ01Z currency slightly lowers the estimate. However these changes tend not to substantially alter the 95% confidence interval, with the exception of the UZ01Z exclusion, that widens the confidence and reduces the statistical significance to the 10% level. Restricting the population to the fully

balanced panel does not alter the trend efficiency estimate, though it does narrow the estimate of variation in efficiency past the 80th centile.

The estimates of both trend efficiency and variation in efficiency are robust to the sensitivity checks on variables, other than those related to inflation. Although it is plausible that the number of sites and the proportion of inpatients from an urban area affect trusts costs, we do not identify any effect.

The largest change to our estimate of trend efficiency is made when we change the way we account for inflation. Replacing our health-specific measure of inflation with the GDP deflator pushes down our estimate to 0.4%, which we note is consistent with the analyses from both the Health Foundation and Deloitte (who also used the GDP deflator as a sensitivity check on their headline models). The choice of variable to control for inflation is important, because any increases in costs due to inflation do not subtract from the rate of efficiency growth. Our measure of health-specific inflation estimates that input prices were 17% higher by the end of a time period compared to the base year. The GDP deflator estimates that input prices were only 11% higher. The difference between these two estimates accounts for the difference in trend efficiency estimates. Our preference is to use the health-specific measure of inflation for two reasons:

1. The inflation trusts faced was higher than that of the economy in general, and represents an uncontrollable cost factor to trusts.

2. Trusts are compensated for inflation (through the inflation cost uplift), and this adjustment of the price level is distinct from the efficiency factor adjustment.

Table 6: Sensitivity analysis

	1) Exc specialist	2) Exc multi-service	3) Exc teaching	4) 1&2&3	5) Exc UZ01Z	6) Exc entrants and exiters	7) Inc number of sites	8) Inc number of acute sites	9) Inc urban %	10) Deflate by MFF	11) Deflating by GDP deflator
Trend efficiency:	1.6%***	1.3%***	1.7%***	1.7%***	1.1%*	1.4%***	1.4%***	1.4%***	1.4%***	1.4%***	0.4%
95% CI:	0.6% to 2.5%	0.4% to 2.3%	0.7% to 2.8%	0.6% to 2.8%	-0.04% to 2.1%	0.5% to 2.4%	0.4% to 2.3%	0.5% to 2.4%	0.4% to 2.3%	0.4% to 2.3%	-0.5% to 1.3%
Variation in efficiency:											
50th to 60th centile	1.4%	2%	1.5%	0.7%	1.6%	2%	1.8%	1.9%	1.8%	1.7%	2%
50th to 70th centile	2.2%	3.3%	3.7%	1.8%	3.4%	3.4%	3.5%	3.3%	3.3%	3.3%	3.7%
50th to 80th centile	4.3%	5%	5.6%	2.9%	5.3%	4.7%	6.1%	5.8%	5.6%	5.3%	5.6%
50th to 90th centile	5.8%	7.6%	7.1%	4.5%	7%	6.5%	7.8%	7.6%	7.5%	7.3%	7.6%
Observations	864	950	816	668	852	936	981	981	981	981	981
Groups	150	163	144	117	147	156	170	170	170	170	170
Coefficients											
Within							0.0004	0.003	0.021		
S.E							(0.0003)	(0.003)	(0.016)		
Between							-0.001	-0.006	-0.005		
S.E							(0.001)	(0.005)	(0.005)		

Source: Monitor analysis

5. Discussion and conclusions

5.1. Changes from last year

We have estimated a model of trust costs and derived from this estimates of the efficiency trend and variation in English NHS acute trusts. This model is based on a model developed by Deloitte for use as evidence to inform the 2015/16 national tariff. We have updated Deloitte's model from last year to include data from 2013/14, and made a number of refinements to the method.

We have made the following changes.

- Added an extra year's data (2013/14). This strengthens our confidence in our estimate of the efficiency improvement we see over time and ensures we are using the most up-to-date information (see 3.1).
- Changed the age threshold to calculate the proportion of inpatients that are young from under 15 to under 19 years (see 3.2.3).
- Included area-level effects (see 3.2.4).
- Based our expectation of the profile of activity across chapters (for the specialisation index) on what we actually observe at the national level, rather than using a uniform distribution (see 3.2.2).
- Adjusted inpatient cost to reflect the additional cost of excess bed days, but maintain the appropriate number of episodes (see footnote 21).
- Reduced the number of diseases covered by the local disease indices due to those that are recorded across all time periods (see 3.2.3).
- Moved from the health domain of the IMD to the overall score, in order to more precisely measure deprivation (see footnote 24).
- Adapted our specification of the model with the Mundlak transformation to account for the difference across the between-trust and within-trust effects (see 2.1.2).
- Retained the full set of observations available by not dispensing with trusts that influence our coefficient estimates (see footnote 19).

Overall, these changes have made relatively little difference to our estimates of efficiency.

5.2. Interpretation for 2016/17 efficiency factor

Our interpretation of the evidence presented here could support an efficiency factor between 1.5% and 2.5% as discussed in 4.1. In terms of the historical trend

efficiency, this range spans the upper half of the confidence interval of our estimate. The extent of variation in efficiency suggests that there is sufficient scope for additional catch-up to support efficiency factors greater than trend efficiency.