Deloitte.

Methodology for efficiency factor estimation.

Final report

23 April 2014

This final report has been prepared on the basis of the limitations set out in in the Important Notice From Deloitte on page 1.

Contents

Important Notice from Deloitte

This Final report (the "Final Report") has been prepared by Deloitte LLP ("Deloitte") for Monitor in accordance with the contract with them 26 November 2014 ("the Contract") and on the basis of the scope and limitations set out below.

The Final Report has been prepared solely for the purposes of developing a framework for considering efficiency factor determination, as set out in the Contract. It should not be used for any other purpose or in any other context, and Deloitte accepts no responsibility for its use in either regard.

The Final Report is provided exclusively for Monitor's use under the terms of the Contract. No party other than Monitor is entitled to rely on the Final Report for any purpose whatsoever and Deloitte accepts no responsibility or liability or duty of care to any party other than Monitor in respect of the Final Report or any of its contents.

The information contained in the Final Report has been obtained from Monitor and third party sources that are clearly referenced in the appropriate sections of the Final Report. Deloitte has neither sought to corroborate this information nor to review its overall reasonableness. Further, any results from the analysis contained in the Final Report are reliant on the information available at the time of writing the Final Report and should not be relied upon in subsequent periods.

Accordingly, no representation or warranty, express or implied, is given and no responsibility or liability is or will be accepted by or on behalf of Deloitte or by any of its partners, employees or agents or any other person as to the accuracy, completeness or correctness of the information contained in this document or any oral information made available and any such liability is expressly disclaimed.

This Final Report and its contents do not constitute financial or other professional advice, and specific advice should be sought about your specific circumstances. In particular, the Final Report does not constitute a recommendation or endorsement by Deloitte to invest or participate in, exit, or otherwise use any of the markets or companies referred to in it. To the fullest extent possible, both Deloitte and Monitor disclaim any liability arising out of the use (or non-use) of the Final Report and its contents, including any action or decision taken as a result of such use (or non-use).

Executive summary

Monitor's main duties under the Health and Social Care Act ("HSCA") 2012 are to protect and promote the interests of people who use health care services, by promoting the provision of health care services which are economic, efficient and effective whilst maintaining or improving the quality of services.[1]

The efficiency factor is a key lever for Monitor to establish positive incentives around achieving greater efficiency. Currently it is applied directly to all services with a nationally determined price, primarily in secondary acute care. However, it also has wider implications across the sector including for locally priced services, such as the majority of mental health and community services. For these services, it is typically used as an anchor point for negotiations between providers and commissioners.

If developed in a robust manner, the factor could lead to efficiency gains across the sector allowing resources to be appropriately allocated and potentially reinvested to improve services, ultimately benefiting patients.

Deloitte has been commissioned by Monitor to recommend an enduring framework for setting the efficiency factor for 2015/16 and beyond. The framework has been developed considering health care, regulatory and international precedent. It has also been tested with health care experts, sector stakeholders and other regulators through a series of structured workshops. As well as describing the framework, this report sets out a longer term vision and a proposed approach to estimating the factor for 2015/16.

Current state

Adjusting prices for expected efficiency gains was introduced within the NHS to national tariff in 2005/06.[2] The Department of Health ("DH") determined the efficiency requirement primarily on the basis of the funding gap; the difference between commissioner allocation and projected commissioner expenditure.

For 2014/15, Monitor has developed this approach by setting a factor which encompasses a range of evidence around achievable efficiency gains.[3] The process that Monitor followed sought to be more transparent and allowed for greater sector consultation. For example, Monitor published an

[1] HSCA 2012 (Section 62).

[2] Department of Health (2011), A simple guide to payment by results

[3] Monitor and NHS England (2013) 2014/15 National Tariff Payment System: An engagement document

engagement document in June 2013[4] which established a range for the efficiency gains expected for 2014/15. This engagement was prior to the full consultation, published in the autumn.[5]

Efficiency factor precedent

There is limited international precedent on the use of explicit efficiency factors in health care systems. Where prices are regulated, the efficiency requirement is typically determined implicitly, as the difference between the provider's actual cost and the regulated price. This price is often set on the basis of average costs across the sector.

However, efficiency adjustments are a common feature of incentive based pricing in most regulated settings. For example, efficiency factors form a part of price controls applied in telecommunications, railways, energy and water. Typically, given the challenges associated with estimation, factors are set based on triangulating a number of efficiency estimates, whilst exercising a degree of discretion. Some of these challenges are potentially abated with health care, given the more rich data which is available. As such, Monitor can benefit from the application of more sophisticated techniques.

Framework for setting efficiency

Drawing upon this precedent, a systematic framework for efficiency factor setting has been proposed. The proposed framework incorporates five key steps to support the development of the efficiency factor.

1. **Definition of efficiency.** A precise definition of efficiency is important to help targeted estimation and ensure stakeholders understand the types of efficiency savings that are included or excluded. This will support the design of appropriate incentives and allow for increased transparency and clarity of messaging to the sector.

2. **Level of efficiency.** The efficiency factor is currently set at a sector wide level. In order to ensure that the policy lever establishes optimal incentives and sends out the right price signals, it will be important to consider the different levels at which the factor can be applied, for example, by provider type or by service.

3. **Estimation methods.** The efficiency factor could be estimated using a variety of approaches, ranging from quantitative and econometric approaches to purely qualitative reviews of international best practices. It will be important that the estimation approach chosen is robust, transparent and feasible.

4. **Estimation challenges.** There are a number of estimation challenges that may arise in efficiency factor determination, particularly around measuring outcomes or accounting for case-mix differences across providers. As such, it will be important that Monitor is aware of these challenges in selecting the appropriate estimation methodology, and make adjustments as required.

[4] Monitor (2013), National Tariff 2014/15: An Engagement Document.

[5] Monitor (2013), 2014/15 National Tariff Payment System

5. **Discretion in efficiency estimation.** Finally, setting the efficiency factor will involve a degree of judgment and discretion. This is important given the challenges related to estimation and Monitor's wider duties as the sector regulator.

Long term recommendations

Core to Monitor's approach to future efficiency factor setting is that it is applied at the appropriate level of disaggregation. For example, the factor could be disaggregated around providers, service groups or provider types. Disaggregating the uniform target could help establish a range of more powerful incentives. However, it is also likely to be difficult to implement, given the context of a complex provider and service landscape. Further, the current evidence to support such disaggregation is limited.

This report has appraised the incentives, challenges and signals created by various options. Based on this, the future approach could be designed to capture differences in the scope for future efficiency gains, create more targeted incentives and help to support provider sustainability. In order to achieve this, it is likely that a split on both a service and provider level will be appropriate.

- **Service groups.** Efficiency factors could be set by service groups to help account for differential market and supply side conditions services. This could, for example, initially involve some disaggregation across secondary acute care, where data is more readily available. These groups could also eventually align to Monitor's and NHS England's joint long-term pricing strategy.

- **Banding.** Within these service groups, providers could also be placed into bands with different efficiency factors. This could reflect differences in the scope for future efficiency gains, allowing for more specific incentives to be set.

Given the lack of data and existing evidence, the preferred option has been developed on the basis of theoretical considerations only. As such, it is recommended that Monitor develops tests of the underlying hypotheses supporting service grouping and provider banding. In addition to these formal tests, Monitor should also seek to engage with the sector around proposed changes.

2015/16 efficiency factor

The potential benefits associated with a more disaggregated efficiency factor for 2015/16 are restricted due to a number of overarching limitations.

- **Data and estimation challenges.** The data available to underpin the estimation of efficiency varies significantly. For example, in secondary care there is significant data available; including voluntary patient-level costing information collected for 2012/13. In community care, however, service definitions are often unclear, reducing the ability to measure the scale and acuity of services being delivered in any estimation of efficiency.

- **Evidence base surrounding disaggregation.** There is limited evidence to examine if there are material differences in scope for efficiency savings across different services. As part of the estimation of the 2015/16 efficiency factor, some initial testing around the presence of

different supply and efficiency conditions might be possible. However, any evidence is likely to require further refinement before being sufficiently accurate to underpin price setting.

These limitations suggest that in the short term, maintaining a sector wide efficiency approach is likely to be more appropriate. However, this approach will need to carefully consider how to account for those providers who are furthest away from the efficient frontier. Moreover, an efficiency target which seeks to establish pricing at the frontier could risk the sustainability of some providers. To mitigate this, Monitor could exercise a degree of discretion and conduct an appropriate impact assessment.

To estimate the uniform efficiency factor, more sophisticated econometric top-down methods could be used. However, acknowledging that data limitations still exist, triangulation of the results should be undertaken using other available evidence, such as bottom-up provider models, achieved cost improvement plans and relevant literature.

Next steps

It is envisaged that Monitor will continue to develop the underlying evidence base and seek feedback from the sector on the recommendations set out in this report. As a first step, this could involve incorporating key learnings from the 2015/16 efficiency estimation process.

1 Introduction

Monitor's main duties under the Health and Social Care Act ("HSCA") 2012 are to protect and promote the interests of people who use health care services, by promoting the provision of health care services which are economic, efficient and effective whilst maintaining or improving the quality of services.[6] The pricing regime needs to embody these duties and create appropriate incentives for efficiency and quality gains.

Where prices are regulated, an 'efficiency factor' is often adopted as part of, or alongside, the price cap in order to incentivise future efficiency improvements. An efficiency factor can be a powerful instrument which, if developed in a robust manner, can support efficiency improvements across the sector in order to allow resources to be appropriately allocated and potentially reinvested to improve services. However, efficiency factors also have the potential to create distortions if not deployed carefully. As such, it is critical that the framework for, and estimation of, any efficiency factor is carefully considered.

1.1 Current state

The efficiency factor is currently applied directly to all services with a nationally determined price, primarily in secondary care. For locally determined services, such as the majority of mental health and community services, it is used as an anchor point for negotiations between providers and commissioners. As stated in the 2014/15 National Tariff Document ("NTD")[7]:

"Commissioners and providers should have regard to the national tariff efficiency and cost uplift factors for 2014/15 when setting local prices for services without a national price for 2014/15, if those services had locally agreed prices in 2013/14."

For 2014/15, Monitor has proposed to apply a single efficiency factor uniformly across all services, as published in the NTD.[8] In previous years the Department of Health ("DH") primarily determined the efficiency factor on the basis of the expected funding gap; the difference between commissioner allocation and projected commissioner expenditure. Monitor has developed this approach by setting an adjustments for estimated efficiency gains, provider cost inflation and other factors such as those for service development.

[6] HSCA 2012 (Section 62).

[7] Monitor and NHS England (2013), 2014/15 National Tariff Payment System, Section 7.4.1

[8] Ibid

The 2014/15 efficiency factor is based on an estimate of achievable efficiency gains that has been informed by a range of evidence including:

- Existing evidence on achievable efficiency savings and relevant best practice, including a more recent study commissioned by Monitor; [9]
- Past productivity gains implied by Cost Improvement Programme ("CIP") and Quality, Innovation, Productivity and Prevention ("QIPP") initiatives to calibrate the potential for savings; and
- Financial returns provided by NHS foundation trusts.

The process that Monitor has followed in setting the tariff for 2014/15 has sought to be more open, transparent and allow for greater feedback from the sector. For example, Monitor published an engagement document in June 2013 [10] which established a range for the efficiency gains expected for 2014/15. This engagement was prior to the full consultation published in the autumn.[11]

1.2 This report

Deloitte has been commissioned by Monitor to develop an enduring framework for setting the efficiency factor, and subsequently, apply this framework to define and estimate the efficiency factor for 2015/16. The framework described in this report has been developed with consideration to health care, regulatory and international precedent described in Section 2. The framework has also been tested on industry experts, regulatory economists and sector stakeholders through a series of structured workshops.[12]

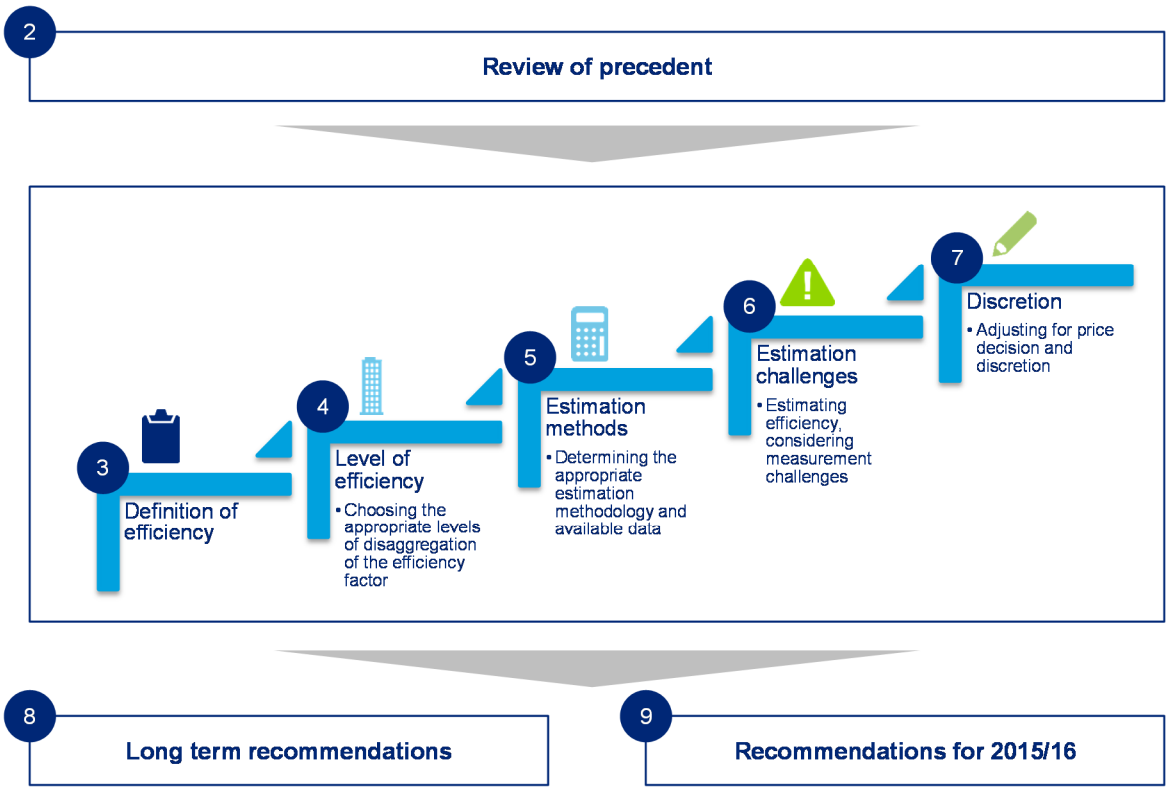
[9] Monitor (2013), Improvement opportunities in the NHS: Quantification and Evidence.

[10] Monitor and NHS England (2013), National Tariff 2014/15: An Engagement Document.

[11] Monitor and NHS England (2013), 2014/15 National Tariff Payment System

[12] Appendix C provides further details on stakeholder engagement.

Figure 1: Efficiency factor framework*



**numbers represent relevant sections in this report.*

The framework incorporates five key steps to support the appropriate development of the efficiency factor:

1. **Definition of efficiency.** Efficiency is often confused with a number of interrelated concepts. A precise definition of efficiency is therefore important to help targeted estimation and ensure stakeholders understand the types of efficiency savings that are included or excluded. This will support the design of appropriate incentives and allow for increased transparency and clarity of messaging to the sector. Section 3 establishes a working definition of efficiency.

2. **Level of efficiency.** The efficiency factor is currently set at a sector wide level. In order to ensure that the policy lever establishes optimal incentives and sends out the right price signals, it will be important to consider the different levels at which the factor can be applied, for example, by provider type or by service. The different potential levels of disaggregation, and their relevance for different pricing decisions, are described in Section 4.

3. **Estimation methods.** The efficiency factor could be estimated using a variety of approaches, ranging from quantitative and econometric approaches to purely qualitative reviews of international best practices. Section 5 provides an overview of the various possible approaches and establishes their relevance.

4. **Estimation challenges.** Estimating the efficiency factor is particularly challenging in the context of health care services. As such, it will be important that Monitor is aware of these challenges in selecting the appropriate estimation methodology, and make adjustments as

required. For instance, the framework should take into account the challenges associated with measuring outcomes or accounting for case-mix differences across providers. Potential challenges that may arise in estimating the efficiency factor are discussed in Section 6.

5. **Discretion in efficiency estimation.** Finally, the efficiency factor will need to be considered in conjunction with the overall pricing decision. Some discretionary factors that may be relevant to the efficiency factor setting are outlined in Section 7.

This framework will be applied in Section 8 to arrive at the long term recommendations for efficiency factor determination. The application of the framework in context of the relevant limitations in 2015/16 is discussed in Section 9. References and further technical details have been included in supporting annexes.

1.3 Terminology clarification

Throughout this report the term efficiency factor refers to the efficiency gains that could be made by providers whilst maintaining their current quality of service provision. It is distinct from other efficiency concepts and terms that are often considered in the context of price setting.

- **Efficiency requirement.** This is the term used in previous national tariffs to describe the efficiency savings that need to be made by the sector as a whole in context of the overall affordability challenge.

- **Efficiency target.** This is a measure of the efficiency savings that a provider must make such that the cost of service provision does not exceed the price set. Where the efficiency factor is set on a uniform basis, it is not necessarily the case that the efficiency target is also the same. For example, a provider who is much further from the chosen efficient frontier will have a greater efficiency requirement, despite a uniform efficiency factor.[13]

This framework will be concerned only with the efficiency factor and, as such, will not consider factors such as affordability or tariff leakage. Whilst these factors are important considerations in overall price setting, it is recommended that they are considered independently from the efficiency factor. [14]

Currently, the efficiency factor directly applies only to services with nationally determined prices and is typically used as an anchor point for local price negotiations. This report will not seek to evaluate the reasonableness of applying the efficiency factor within different payment systems that Monitor and NHS England develops as part of its joint pricing strategy.[15]

---

[13] Under the current tariff design, where prices are based on reference costs, the efficiency target is equal to the sum of: the difference between a provider's costs and the average reference cost, and, the efficiency factor (or requirement, in previous tariffs).

[14] The related factors could be identified and built into separate adjustments in the tariff equation.

[15] Monitor's and NHS England's joint vision for development of the payment system is outlined here: Monitor (2014), *How Monitor and NHS England are working to make the payment system do more for patients from 2015/16.*

2 Review of precedent

A range of evidence and precedent across a number of regulatory settings and international health care systems has been reviewed.[16] These have provided important insights which have informed the framework developed. This section sets out the core findings from the three areas of precedent which have been considered:

1. Precedent in other regulated sectors within the UK;
2. International precedent within the health care sector; and
3. Precedent within the NHS.

A summary of key findings and potential implications for the framework is provided in Section 2.4.

2.1 Precedent in other regulated sectors

A common feature of incentive regulation schemes is the application of an efficiency or productivity adjustment. [17] This has been applied across several regulated sectors in the UK, where prices are generally allowed to change by the rate of inflation minus, amongst other factors, an X-factor or efficiency factor.

The efficiency factor is an index number and reflects efficiency gains that can be achieved by the regulated provider over the price cap or tariff period. Whilst there is no commonly used explicit definition, the efficiency factor in regulated sectors in the UK typically reflects the difference between a provider's actual cost performance and the performance of a fully efficient provider; the minimum cost that is required to provide a specific level of output for a given quality.

A number of recent price controls have been reviewed to understand common techniques and challenges encountered by regulators in estimating the efficiency factor. The regulated sectors considered include:

- Communications, regulated by Ofcom;
- Water and sewerage, regulated by Ofwat;
- Electricity and gas networks, under Ofgem regulation;

[16] Appendix A provides a full list of the precedents reviewed.

[17] For a useful introduction see Makholm (2007), Elusive Efficiency and the X-Factor in Incentive Regulation: the Törnqvist v. DEA/Malmquist Dispute, NERA publications.

- Post and mail, previously overseen by Postcomm and currently Ofcom regulation;[18] and
- Railways, regulated by the Office of Rail Regulation (ORR).

Table 1 provides an overview of the approaches used across the regulated sectors.

[18] Ofcom took over regulation of UK's postal services from Postcomm from October 2011.

Table 1: Overview of regulatory frameworks in the UK

	Regulator	Sector	Review Year	Market	Regulatory framework	Methodology	Quality of service	Efficiency factor	Review Period
	Ofgem	Power	2009	Electricity Distribution	RPI-X	**Top-down**: Pooled Corrected Ordinary Least Squares ("COLS") **Internal benchmarking on key outputs** (use of benchmark indicators) **Additional Methods**: • Value of Lost Load ("VoLL") based on survey analysis; • Forward-looking cash-out model	Several proxies on performance, not always included in the model: • Customer interruptions • Customer minutes lost • Unplanned interruptions • Pre-arranged interruptions • HV benchmark • Interruptions over 12 hours	0.7%-1% (depending on the services) per annum	5 years
	Ofgem	Gas	2012	Gas Distribution	RIIO	**Top-down:** Pooled COLS on capital expenditure ("CAPEX"), replacement expenditure ("REPEX") and total expenditure ("TOTEX")	Not implemented in the model: • Customer satisfaction • Reliability and availability • Safe network services • Connection terms • Environmental impact • Social obligations	Differential catch-up for each company based on submitted business plans Frontier shift includes 1% operating expenditure ("OPEX") efficiency and 0.7% CAPEX and REPEX efficiency	8 years
	Ofwat	Water	2009	Tariffs for Water	RPI ± K + U	**Top-down:** Pooled COLS on CAPEX; **Additional Methods:** CAPEX incentive scheme ("CIS")	No	Continuing OPEX efficiency (X factor) 0.25% per annum; Industry annual average K of +0.5% (-0.2% at draft) per annum.	5 years
	Ofcom	Telecom	2009	BT Openreach	RPI - X	**Top-down**: Panel Stochastic Frontier Analysis ("SFA") on TOTEX **Bottom-up modelling**: Long run incremental cost ("LRIC")	• Percentage of orders completed with the committed time • Average number of fault reports per 1000 switched lines • Average number of working hours required to repair faults	-2.5% to +5% per annum (depending on the services)	5 years
	Postcomm	Mail	2005/ 2006	Services not open to competition	RPI-X	**Top-down**: • Internal benchmarking using SFA and Data Envelopment Analysis ("DEA") • International postal operators productivity comparison **Total factor productivity ("TFP")** **Bottom-up review**: expert review and assessment of efficiency savings from introducing initiatives across different stages of Royal Mail's operations (delivery, collection, sortation, etc.)	No	2.75% to 3.25% per annum	4 years
	ORR	Railways	2013	Network access	RPI-X	**Top-down (on TOTEX):** • SFA • COLS • DEA	Safety indicators not applied in the main model	17% efficiency gains related to OPEX and 15.8% related to maintenance and renewals over 5 years.	5 years (or large scale interim review)

Source: Deloitte analysis and references in Appendix A.

2.1.1 Key themes

The following key themes have emerged from the evidence reviewed.

Separation of the efficiency factor from other incentives

RPI-X is the most common form of pricing regulation employed in the UK, allowing prices to rise by inflation less (often alongside other adjustments) the efficiency factor. A number of recent price controls in telecoms, railways and postal sectors have employed the RPI-X approach. There are also a number of variants to the standard RPI-X formula. For example:

- Ofwat consider a modified formula for RPI-X regulation in the water sector: RPI ± K + U. The K factor takes into account the efficiency factor but also other factors such as the capital investment required meeting the statuary obligations, returning on capital and tax requirements. 'U' represents any unused price limit that the company has carried forward and which may be used in future years. This latter factor applies in cases in which the company has not increased prices as much as they could have in the past.

- Ofgem has recently developed a new approach to pricing regulation, the so called 'Revenue = Incentives + Innovation + Outputs' ("RIIO") scheme. The pricing formula tries to promote efficiency cost savings, also providing sufficient incentives for capital investment and environmental considerations.

Sample sizes and comparators

From the literature surveyed, [19] all regulators have employed some form of top-down econometric analysis to estimate efficiency factors. One of the main challenges that regulators have faced is the availability of reasonable sample sizes. Some of the potential ways in which regulators have approached this challenge include:

- **Time-series data.** Most regulators have utilised historic time-series data to increase the observations used as the basis of estimation for the catch-up and frontier shift components.

- **International data.** Ofcom, for example, has historically benchmarked BT against the US local exchange carriers ("LECs"). The ORR has also collected data on international comparators.

- **Internal benchmarking.** Internal benchmarking compares the cost performance of similar units within the same organisation. It was the approach used by Postcomm to identify potential inefficiencies in Royal Mail's mail distribution units.

[19] A full list of literature surveyed can be found in Appendix A.

Despite these efforts, the quantity and quality of data utilised often opens up the estimation of the efficiency factor to challenge. Further, data availability may limit the econometric methods which can be deployed.

Triangulation of approaches

Typically regulators employ more than one type of approach to support the determination of efficiency factors. For instance, Postcomm used a combination of various top-down econometric approaches and a bottom-up review of Royal Mail's operations to estimate the efficiency factor. Further, Ofgem considers a 'Value of Lost Load' model when determining the efficiency factor for the electricity distribution regulatory framework, together with more traditional top-down methods.

Multiple estimation approaches are used given that estimating the efficiency factor is inherently challenging. Triangulating the results of alternative methodologies, effectively, reduces the uncertainty around the point estimate derived from a single approach. By introducing a range of estimates, regulators are required to act with a degree of discretion in setting the final efficiency factor. For instance, LECG (2005), commissioned by Postcomm for the Royal Mail 2005/06 price cap review points out*: "We have considered the scope for efficiency savings using a variety of methods. None of these methods by itself provides a precise picture of the scope for savings during the forthcoming price control, and each requires us to exercise a degree of judgement when determining the implications for Royal Mail. However, by approaching the efficiency assessment from a number of different directions we avoid placing undue weight on any one piece of analysis. Instead, we look at a broad range of evidence and set cost allowances based on overall picture that emerges. This helps to minimise the extent to which our overall conclusions might otherwise be subject to error".*

Accounting for quality

Quality is typically difficult to control for in the estimation of efficiency, primarily due to measurement issues. The lack of a robust approach to account for quality could lead to efficiency estimates being biased upwards or downwards, depending on the relationship between cost and quality.

Given these measurement issues, regulators often resort to proxy measures of quality or very narrow indicators. For example, Ofcom considers the average number of faults per line as a proxy of the quality of the signal transmission. Ofgem keeps in high consideration the interruptions in the gas and electricity services by using, for example, customer interruptions, customer minutes lost and unplanned interruptions. Additionally, Ofgem also utilises customer satisfaction (measured through surveys), service reliability and availability, and safety of network service. These quality indicators are sometimes difficult to implement in efficiency estimation methods and are therefore used separately in the price review in order to provide additional evidence.

Variation of the efficiency factor by regulated service

Efficiency factors can potentially vary by the services which are regulated. For example:

- Ofcom uses different efficiency factors for leased lines prices versus wholesale broadband access.[20]

- In setting its price controls for 2015-20, Ofwat has indicated an intention to apply two separate efficiency factors for water and wastewater wholesale services.[21]

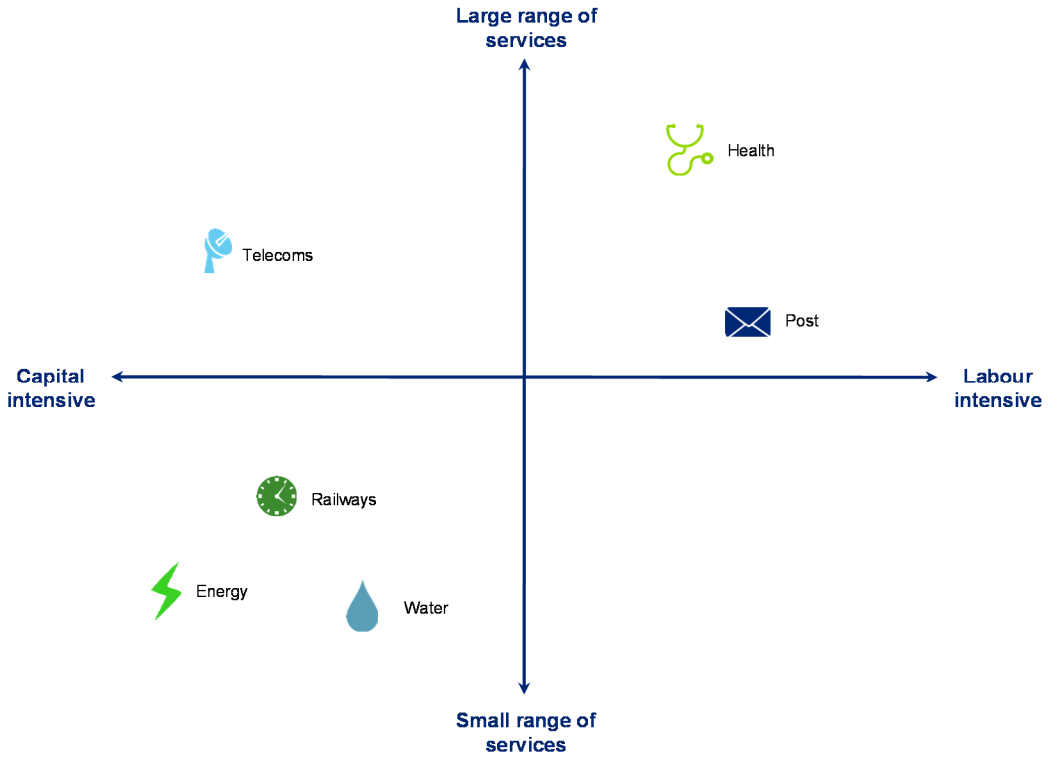
The differences in efficiency factors should relate to the underlying supply conditions, for example whether different services are characterised by significantly different delivery models and hence have different scope for efficiency savings.

Figure 2 illustrates the differences between sectors according to two dimensions: the range of services provided and factor intensity. Network industries such as telecoms are typically capital intensive and provide a small range of services. The health care industry instead is a more labour intensive sector, more comparable to the postal sector; and involves a far greater number of services.

[20] This can be seen in Ofcom (2013) WBA Charge Control, and Ofcom (2013) Leased Lines Charge Control.

[21] Ofwat (2013), Setting price controls for 2015-20: Final methodology and expectations for companies' business plans

Figure 2: Illustrative sector key differences



Source: Deloitte analysis

Cross-subsidisation between services

An important factor to consider when regulating multi-service providers is how to mitigate the risk of cross-subsidies between unregulated and regulated services. Other sector regulators have typically addressed the problem by applying, for instance:

- Detailed regulatory reporting requirements, as those imposed by Ofcom on BT; and
- Legal or functional separation between regulated and unregulated segments of the business, as is the case in the energy sector.

2.2 International health care precedent

This sub-section considers international evidence in efficiency setting and tariff design within health care. The health care systems reviewed included:

- European average cost based tariffs;
- National efficient price in Australia; and
- Efficiency through regulated competition in the Netherlands.

The European perspective: Average cost based tariff[22]

Activity based funding has become the most common mechanism for reimbursing hospitals in Europe. In particular, since the mid-1980s, several European countries have introduced diagnosis-related group ("DRG") based payment systems with the objective of improving efficiency and enhancing financial transparency.

National tariffs are determined at the DRG level on the basis of the average cost of services. However, prices are often adjusted above the national average to take into account additional provider specific factors. For instance, in Germany and France, tariffs are adjusted to reflect regional differences in costs, similar to the market forces factors ("MFF") used in England. In Germany a further adjustment is made to account for quality of service. In addition, several countries have developed mechanisms to identify outlier cases and to pay hospitals separately for the extra costs of treating such patients.

The tariff determination is implemented without an explicit estimation of the efficiency factor, which is implicitly calculated as the difference between actual cost performance and the sector average performance. The rationale for this methodology is that service providers identified as inefficient will be pushed to improve their efficiency towards the efficiency of the average provider. However, the analysis is very simplistic and does not take into consideration scale effects or quality factors that drive cost differentials and, consequently, may impose an unfair burden on smaller or better quality providers. In addition, this approach does not provide any incentives to improve efficiency above the sector average.

The Australian example: National Efficient Price

The Australian regime does not incorporate within its pricing framework an efficiency adjustment factor *per se*. Similar to the European regime described in the previous section, Australia determines a benchmark price, called the "national efficient price" ("NEP"), which reflects the national average cost of health care provision. The efficiency factor is determined for each provider as the difference between actual cost and the sector average performance after accounting for provider specific factors such as location.

The Independent Hospital Pricing Authority ("IHPA") uses the term National Weighted Activity Unit ("NWAU"), which is a measure of health care activity expressed as a common unit. Conceptually, the average hospital service is worth one NWAU. The most intensive and expensive activities are worth multiple NWAUs, the simplest and least expensive are worth fractions of an NWAU. The NEP is essentially the price of one NWAU.

---

[22] This section draws from: O'Reilly, J. et al. (2012), Paying for hospital care: the experience with implementing activity-based funding in five European countries. Health Economics, Policy and Law, 7:73-101; and European Observatory on Health Systems and Policies, Diagnosis-Related Groups in Europe. Open University Press.

The Netherlands: Efficiency through regulated competition

The Dutch health care system went through a radical reform in 2006, when a dual system of mandatory public insurance and voluntary private insurance moved to mandatory private insurance. This reform marked a departure from a supply and price regulated regime to a regulated competitive regime. Essentially, the reform introduced a mandatory private insurance for the whole population.

Under this regime, insurance providers purchase health services directly from health providers, and in turn sell these services to 'consumers'. Insurers can set up their network of providers and selectively contract for discounted services whereas consumers can generally switch between insurers once a year.[23] This system resembles some characteristics of a competitive market and reduces the need for explicit price-focused regulation.

2.3 Precedent within the NHS

For 2014/15, Monitor has proposed a formula to calculate national tariffs similar to RPI-X, where RPI is actually replaced by the cost uplift factor. The cost uplift factor is an inflation factor reflecting more specific health care related costs, for example inflation in regulated drugs prices.[24] The cost uplift factor differs across some services to reflect differences in inflationary pressures. Despite some variance in the cost inflation by service, the efficiency factor is currently estimated and applied uniformly across all services, as published in the NTD.[25] The efficiency factor currently applies to national prices and is also an anchoring point for negotiations between providers and commissioners on local prices.

In previous years the DH primarily determined the efficiency factor on the basis of the expected funding gap; the difference between commissioner allocation and projected commissioner expenditure. Monitor has developed this approach by setting an adjustments for estimated efficiency gains, provider cost inflation and other factors such as those for service development.

Figure 3 shows the evolution of the annual net change in tariff, decomposed between the cost uplift and the efficiency factor over time. Between 2004/05 and 2009/10, the tariff adjustment was always positive reflecting relatively modest efficiency factor adjustments and significant cost uplift factors. However, since 2011/12, year-on-year changes in tariff have been negative.

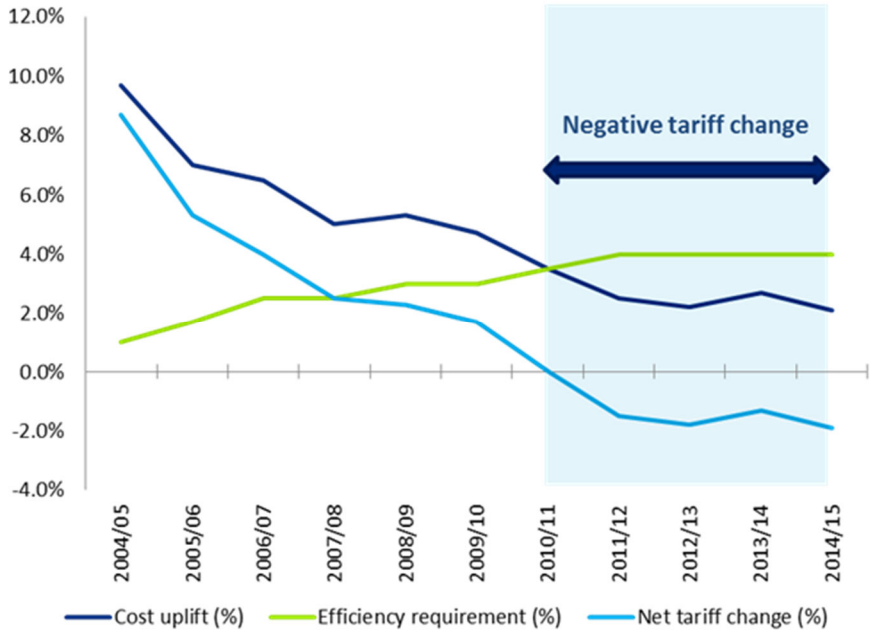
---

[23] Mosca, Ilaria (2013), Evaluating reforms in the Netherlands' competitive health insurance system. In Eurohealth, Incorporating Euro Observer, Vol 18, No.3, 2012.

[24] In particular, the components of the cost uplift factor are: pay, drugs and other operating inflation; service development; capital cost; and HRG specific Clinical Negligence Scheme for Trust ("CNST").

[25] Monitor and NHS England (2013) 2014/15 National Tariff Payment System: An engagement document

Figure 3: Net tariff change, cost uplift and efficiency factor



Source: Deloitte analysis based on DH Payment by Results Guidance 2005/06 to 2013/14[26] and Monitor[27])

To help foster stability, Monitor has set national prices for 2014/15 by adjusting the 2013/14 prices by a cost uplift and an efficiency factor. For 2014/15, the underlying cost base on which prices are set has not been updated. The efficiency factor has been estimated based on:

- Research papers by McKinsey (2009), King's Fund (2010) and Nuffield Trust (2012), as well as a more recent study by Monitor (2013);[28]
- Past productivity gains implied by CIP and QIPP initiatives which were used to calibrate the potential for savings; and
- Financial returns provided by NHS foundation trusts.

The process that Monitor has followed in setting the tariff for 2014/15 has sought to be more open, transparent and allow for greater feedback from the sector. For example, Monitor published an

[26] Reports and spreadsheets available at: Department of Health's National Archives: http://webarchive.nationalarchives.gov.uk/+/https://www.gov.uk/government/organisations/department-of-health

[27] Monitor and NHS England (2013) 2014/15 National Tariff Payment System: An engagement document

[28] Monitor (2013), Improvement opportunities in the NHS: Quantification and Evidence.

engagement document in June 2013, [29] which established a range for the efficiency gains expected for 2014/15. This engagement was prior to the full consultation published in the autumn.

2.4 Implications

The body of evidence considered across a range of regulated settings and international health care systems has provided a number of insights which have been useful in informing the development of the framework. The key findings and the relevant implications for the efficiency framework are discussed below.

1. **There are typically two core components encompassing the efficiency factor.** In previous price cap reviews in other UK regulated sectors, the efficiency factor encompasses both cost savings that can be achieved by becoming as efficient as the most efficient comparable provider (the catch-up component) as well as expected gains that can be realised in the future due to productivity improvement opportunities affecting the whole sector (the frontier shifts).

2. **A number of rigorous approaches have been applied in the literature and previous regulatory reviews to determine the efficiency factor.** These include top-down cost benchmarking, bottom-up cost modelling, bottom-up cost reviews and total factor productivity ("TFP") techniques. Monitor can benefit from the application of these more rigorous approaches available in the literature.

3. **Monitor has an opportunity to exploit greater data availability.** Often regulators are materially constrained in the estimation of efficiency due to a lack of comparable providers. Monitor has access to a much richer pool of comparators in addition to detailed cost, volume and other information.

4. **Estimates of efficiency are often triangulated.** Efficiency is difficult to measure given that it is largely unobservable. Often regulators seek to triangulate estimates using a variety of approaches and evidence. Monitor would benefit from relying on an approach which draws on a range of evidence.

5. **Regulators act with some discretion.** Given efficiency can only be imperfectly measured, and that alternative methodologies and sources of information are often used to make inference, regulators typically exercise some degree of discretion. It is recommended that any discretion by Monitor is exercised with reference to a clear framework linking directly to achieving Monitor's duties.

6. **Regulators find it challenging to account for quality in estimating efficiency.** The quality of service delivered is a key determinant of the costs of provision. The lack of a robust approach to account for quality could lead to efficiency estimates being biased upwards or downwards, depending on the relationship between cost and quality. Measuring quality is often challenging, possibly leading regulators to exclude quality from efficiency

[29] Monitor and NHS England (2013), National Tariff 2014/15: An Engagement Document

estimation. The enduring framework should consider alternative methods to account for quality of service.

7. **Demand and supply conditions in health care are different than in other sectors and this should be considered in the estimation of the efficiency factor.** The approach to the efficiency factor estimation will need to account for the fact that the health care sector, compared to other regulated industries, is characterised by a much larger set of services, a much greater heterogeneity within services and a more significant workforce component of total cost.

3 Definition of efficiency

Efficiency is defined in a number of different ways across the economic literature. In order to ensure that the efficiency factor sets the appropriate incentives, it will be important that a precise definition of the type of efficiency being considered is clearly established. This will support robust estimation and allow for increased transparency and clarity of messaging to the sector. This section sets out the different types and sources of efficiency and examines their relevance to the efficiency factor.

3.1 Types of efficiency

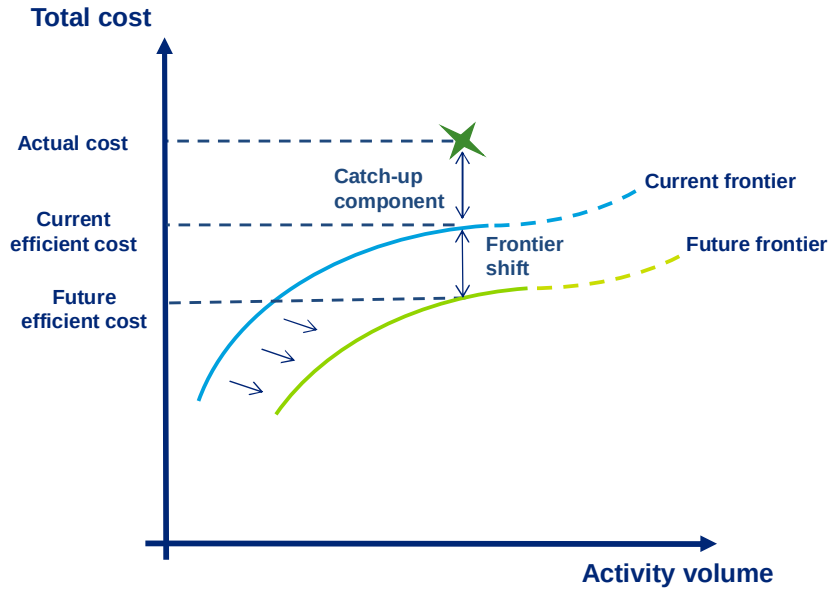
The efficiency factor can be described as comprising two distinct components.

1. **Catch-up.** This captures the efficiency savings associated with the provider becoming as efficient as the most efficient comparable provider in the sector. The catch-up component is typically provider specific and allows for benchmarking and the assessment of current provider efficiencies.

2. **Frontier shift.** This captures efficiency savings from the potential future sector wide productivity gains due to technological advances or service delivery optimisations. This is the forward-looking component of the efficiency factor and aims to capture the dynamic nature of productivity change within the sector.

The catch-up and frontier shift are illustrated in Figure 4.

Figure 4: Efficient frontier



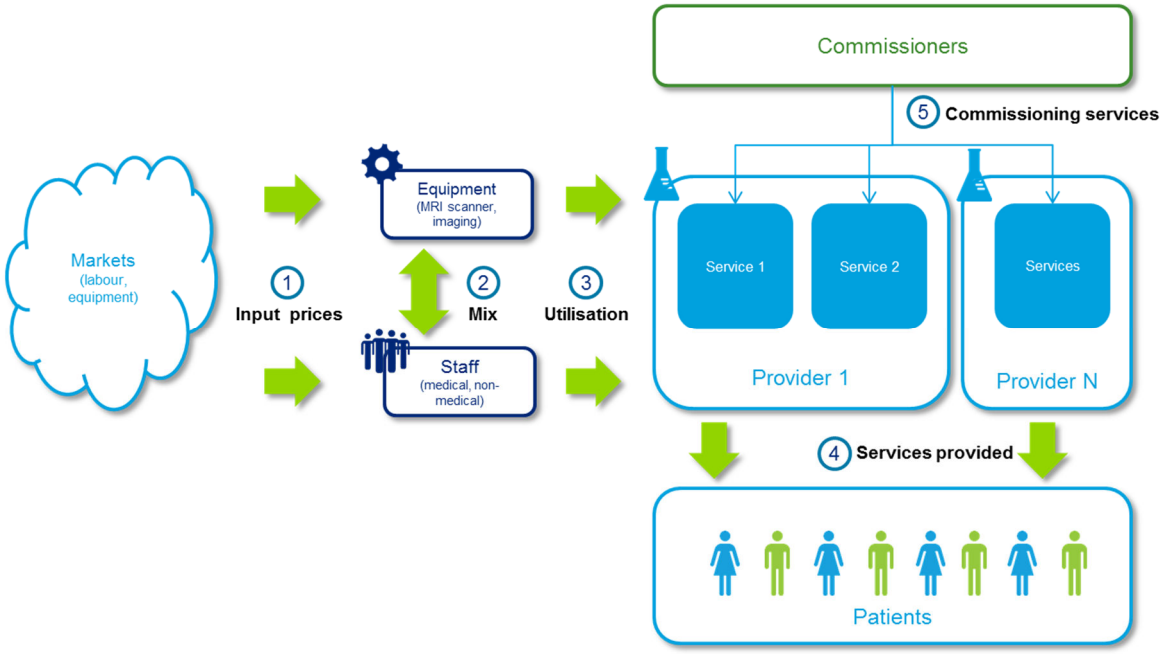
Source: Deloitte analysis

The blue cost curve represents the current efficient frontier, or the lowest cost required to produce a given level of output. All providers will be located above this curve, and therefore have scope to reduce their costs to the efficient level. The light green curve represents the future efficient frontier. The distance between the two is the frontier shift. The current catch-up efficiency is the distance between the current cost and the current efficiency frontier.

3.2 Sources of efficiency

Potential sources of efficiency within the health care sector can broadly be grouped into five categories as illustrated in Figure 5.

Figure 5: Sources of efficiency



Of these five sources of efficiency, the first four are considered to be relevant to the efficiency factor.

1. **Input prices.** Providers purchase and procure various inputs to deliver health care services. Inputs are likely to include medical staff, non-medical staff, equipment, drugs and consumables. The ability for providers to purchase and procure inputs at lower prices will impact their overall efficiency.

2. **Mix of inputs.** There are likely to be different ways in which services can be delivered, whilst maintaining quality. For example, pathology services can be delivered through various levels of automation, changing the equipment and staffing mix required. Varying the mix of inputs can change the level of efficiency of providers.

3. **Utilisation.** Staff and equipment can be utilised to different degrees. For example, in acute general wards bed occupancy rates of around 85% is widely acknowledged as the maximum occupancy that could be reached without impacting on patient safety and quality.[30,31,32] As

[30] Jones (2011), "Hospital bed occupancy demystified and why hospitals of different size and complexity must run at a different average occupancy", British Journal of Healthcare Management.

[31] McKee (2004), "Reducing hospital beds. What are the lessons to be learned?", European Observatory on Health Systems and Policies

[32] Department of Health (2004), "Shaping the Future NHS: Long Term Planning for Hospitals and Related Services"

such, occupancy rates below 80% could be indicative of potential efficiency gains to be made, as long as appropriate staffing levels are in place.

4. **Services provided.** Providers can exercise a degree of discretion relating to the services delivered to patients, whilst maintaining the level of quality. For example, some patients could be treated as either a day case or as an inpatient. Where appropriate, providers could seek to reduce costs through increasing the proportion of day cases.

Figure 5 also identifies a fifth source of inefficiency, the commissioning of services. Commissioners are allocated funds centrally for the commissioning of health care services for their local population. The use of these funds can lead to different health outcomes, contingent on the commissioning decisions made. However, efficiency differences driven by such decisions are beyond the control of providers, and as such, are not considered to be appropriate for inclusion within the efficiency factor.

4 Level of efficiency

The efficiency factor reflects the efficiency gains that the provider may be expected to make, whilst maintaining the quality of service delivered. The factor has been applied uniformly to all nationally determined prices. However, there may be reasons supporting the application at a more disaggregated level, for example, by provider type or service group. Significant differences in the underlying supply conditions may exist, leading to differential scope for efficiency savings at the given level of disaggregation.

Monitor needs to assess the alternative levels at which the efficiency factor could be disaggregated. This section considers this question based on a number of factors, notably, the incentives established and the differences in underlying market conditions.

4.1 Selection of appropriate level

When determining the appropriate level for the efficiency factor, Monitor will need to assess whether there are material differences in scope for efficiency savings across the different levels. Four considerations are likely to be relevant; market condition, incentives, feasibility and simplicity. The remainder of this section sets out these factors in greater detail.

Market conditions

The scope for efficiency savings will be determined, to a large extent, by the underlying market conditions:

- Demand conditions could relate to local health needs, commissioners and reimbursement mechanisms; and
- Supply conditions include underlying differences in delivery models and starting and future levels of efficiency.

Incentives

The level of disaggregation of the efficiency factor may also provide scope to establish different incentives. Disaggregated factors could incentivise a number of behaviours.

- **Innovative models of service delivery.** For example, a targeted factor around a particularly pathway could help to incentivise integration of care. This could be achieved by setting the factor such that integrated delivery models form the basis of the efficient frontier.

- **Greater flexibility.** A highly disaggregated efficiency factor may also limit the scope for providers to exercise discretion with regards to the efficiency opportunities they prioritise.

- **Positive organisational behaviours**. For example, an efficiency factor that is specific to a provider may incentivise positive organisational behaviours to achieve or outperform the efficiency factor.

Simplicity

Simplicity will be important to ensure that the efficiency factor sends out clear price signals. Commissioners and providers use the efficiency factor more widely as an anchor for local contracting and pricing for services off the national tariff. A highly disaggregated efficiency factor could create more complexity in local negotiations.

Feasibility

Finally, the appropriate level should be feasible to estimate. The lack of robust underlying data to support the estimation methodology may lead to inaccurate estimates, distorting price signals and potentially creating perverse incentives.

4.2 Potential levels of efficiency

This section discusses the different levels of efficiency that could be considered:

- Uniform efficiency factor;

- Provider specific efficiency factor;

- Efficiency factor by service group; and

- Efficiency factor by provider type.

Uniform

A single efficiency factor could be applied across the sector. This would have a number of advantages:

- **Simplicity.** A sector wide efficiency factor allows for a consistent and simple signal of the efficiency gains that could be achieved. This would be particularly important given that the efficiency factor is used as an anchor for negotiations between providers and commissioners for other services that don't have nationally determined prices.

- **Feasibility.** A uniform factor could be relatively simple to estimate, and as such, in the absence of robust underlying data, would minimise the scope for potential distortions.

However, a uniform factor does not provide the scope to capture underlying differences in market conditions, existing and historic reimbursement systems, service delivery models and other factors

which may affect the scope for efficiency savings. Further, a uniform factor does not allow for the setting of targeted incentives through the efficiency factor.

Provider specific

The efficiency factor could be estimated and applied separately for each provider. This allows for a consideration of the relative inefficiencies of individual providers and could have a number of advantages.

- **Supporting provider sustainability.** If some providers are significantly further from the efficient frontier they could be allowed to catch-up over a longer period. This could support more sustainable efficiency gains.

- **Allowing for specific factors.** Provider specific efficiency factors could allow for differences in economies of scale, case-mix and a variety of uncontrollable factors driving efficiency.[33]

However, a provider specific factor may potentially create distortions within the system. For example, differential pricing across providers resulting from specific efficiency factors may incentivise commissioners to choose providers where a lower price is observed. Where tariffs support competition based on quality, rather than price[34], this could be undermined.[35]

Further, setting provider specific efficiency factors will require confidence that the estimation of each provider's overall efficiency is relatively accurate. This could be challenging given the currently available data.

Provider level efficiency factors also need to be considered in the context of other pricing instruments. Table 2 provides a summary of relevant provider specific adjustments available to Monitor.

[33] It is noted that some of these factors could be argued to be within the control of the providers.

[34] The King's Fund (2012), Payment by Results: How can payment systems help to deliver better care

[35] This could be mitigated, through offsetting allocation payments to commissioners. However, such payments are complicated to estimate and include within the allocation formula.

Table 2: Other pricing policy instruments[36]

Adjustment	What it reflects	Description
Market Forces Factor	Regional cost differences	Accounts for geographical differences in input prices that providers face, relating mainly to staff, land and building costs.
Top up payments	Patient complexity	Accounts for cost differences arising because some providers systematically serve more complex patients with specialised services.
Local modifications	Unavoidable and structural costs	A provider can apply for local modifications to tariffs for specific services if the provider can reasonably demonstrate that the costs are higher due to uncontrollable structural factors beyond those captured by the MFF, which accounts for unavoidable costs. These applications are to be made after national prices have been set.
Local variations	Innovative service delivery	Allows for innovation in service delivery through bundling or unbundling existing national currencies or creating new integrated currency.

For example, local modifications allow providers to apply for an adjustment to the national tariff given diseconomies of scale or a complex case-mix, which they are unable to control. Provider specific targets may therefore overlap with such adjustments. When considering this level of disaggregation, Monitor should examine the potential interactions with its other policy instruments.

Service group

Determining the efficiency factor by service group could be appropriate where there exist large differences in current efficiency or anticipated differences in future supply or demand conditions. Whilst health care costs tend to be predominantly labour driven[37], some services could still have greater potential to achieve efficiencies. For example, costs of routine chemistry and haematology tests are highly dependent on the extent of automation incorporated in process design.[38]

Efficiency opportunities could also vary by point of delivery. For example, in outpatient settings, there could be significant efficiency savings through reducing the number of did-not-attends ("DNAs"). In inpatient settings, sources of efficiency savings could include reducing the length of stay, improving utilisation of theatres and wards and optimal use of agency or locum staff.

Service specific factors could be important in the context of various payment systems for NHS funded care. For example, as national payment mechanisms are potentially developed for mental health and community health services, differential factors could apply to these services.

---

[36] Monitor (2013), 2014/15 National Tariff Payment System

[37] The King's Fund (2013), "NHS and social care workforce: meeting our needs now and in the future?"

[38] Department of Health (2006), Report of the Review of NHS Pathology Services in England, Chaired by Lord Carter of Coles, an independent review for the Department of Health

Appendix B provides a list of potential service groupings that could be considered in context of efficiency setting.

Provider type

An alternative approach to disaggregate efficiency targets could be to group providers based on certain characteristics. This could reflect systematic differences in the provider's ability to achieve efficiency savings, for example, through differential cost bases, incentives to invest in cost saving technology or even the mix of services offered. The factors that could be considered when grouping the providers are summarised in Appendix B.

5 Estimation methods

There are a number of approaches that could be undertaken to estimate the efficiency factor. These could range from simple benchmarking exercises, which rank providers based on available cost data, to sophisticated econometric approaches that control for factors such as quality and case-mix. This section sets out the different approaches that could be undertaken and the framework that could be applied to select the appropriate approach.

5.1 Selection of appropriate approach

In selecting the appropriate approach, Monitor could assess each against a number of criteria:

- **Level of efficiency factor.** The preferred approach should be able to estimate the efficiency factor at the selected level of disaggregation. In reality, there may need to be a feedback mechanism to adjust the level of the efficiency factor, should estimation not prove to be feasible.

- **Feasibility.** The choice of the preferred approach will be reliant on the quality and reliability of the underlying data required to support estimation.

- **Accuracy.** Accepting that efficiency is difficult to estimate, the preferred approach adopted should be as accurate as possible. For example, given provider differences, efficiency measures based purely on current best practice or simple benchmarking that does not account for case-mix should be given lower priority.

5.2 Potential estimation methods

This section will discuss the different approaches that could be undertaken. These are:

- **Simple benchmarking.** Benchmarking and ranking of providers based on available average cost data, for example, the reference cost index (“RCI”).

- **Productivity indexing.** These include TFP approaches that may be used to examine sector wide trends and productivity shifts.

- **Top-down approaches.** These include methods such as data envelopment analysis (“DEA”) and econometric techniques.

- **Bottom-up modelling.** These estimate efficient costs by building detailed supply side capacity models for defined activity levels.

- **Other approaches.** These include methods such as bottom-up reviews and qualitative reviews of best practice and available literature.

5.2.1 Simple benchmarking

Simple benchmarking uses available cost, quality or activity data to measure a provider's performance in comparison to a set of comparator providers. Benchmarking is often used to measure performance using a specific indicator resulting in a metric of performance that is then compared to others. Such methods are widely used within the sector as they are easy to understand and relatively simple to construct.

5.2.2 Productivity indexing

Productivity is often defined as a ratio comparing the total amount of output produced to the volume of input utilised in the production process. TFP growth is used to assess the change in productivity over time, and is computed by dividing an index of output growth by an index of input growth.[39] TFP approaches have been used in the health care sector to assess productivity, and can be used to measure the frontier shift of efficiency. The indices used to measure input and output growth tend to differ, and controls can be included to account for external factors.

Box 1: TFP precedent in the health care sector[40, 41]

The TFP approach has been used on a number of occasions to estimate productivity trends in the health care sector. Two important precedents in this regard are the studies conducted by the Office of National Statistics ("ONS") and the DH (DH 2014, ONS 2012). These studies have highlighted some important limitations of such approaches.

TFP measures are particularly sensitive to the specification of output and inputs. This has been highlighted in the DH study, where using a revised output index significantly altered the productivity estimates from -0.2% to 0.4%.

Adjustment for quality is also a significant challenge. A number of different approaches have been used to measure quality. For example, the DH study creates an NHS output index, which captures activity type by point of delivery and includes specific outcome measures. In the ONS study the output index captures activity by point of delivery before aggregating, and is cost-weighted and a quality-adjustment factor recommended in the Atkinson Review (2005) is applied. The quality measures employed include survival rates, waiting times, and National Patient Survey results.

However, the ability to control for quality is limited compared to other top-down approaches because the quality adjustment is sensitive to the weighting given to the quality indicators.

---

[39] Castelli et al (2009). "Getting out what we put in: how productive is the NHS in England?"

[40] ONS (2012), "Public Service Productivity Estimates: Healthcare, 2010".

[41] Department of Health funded the Centre for Health Economics, University of York, CHE Research Paper 94 (2014), "Productivity of the English National Health Service from 2004/5: Updated to 2011/12".

5.2.3 Top-down approaches

Data Envelopment Analysis

DEA has been widely used to measure efficiency in the health care sector. This estimates the efficient frontier based on a consideration of different capacity combinations that could be used to meet a particular level of activity or demand. DEA normalises the output and case-mix of all comparators and determines the most efficient resource utilisation amongst the providers considered. Similar to some econometric methods such as Stochastic Frontier Analysis ("SFA"), this technique estimates both catch-up and frontier shift components and can also account for uncontrollable factors which may impact costs. However, unlike SFA, it does not require a production function specification, which is advantageous given that the production processes and intensity of resource utilisation vary across providers.

DEA is a highly data driven approach and can exploit the rich information sets available within the health care system. However, the method is highly sensitive to outliers and any measurement errors can inflict significant bias on the estimates. This could potentially be a limitation given the numerous cost distortions and differences in allocation methodologies across providers.

Box 2: DEA precedent in the health care sector[42]

A cross-sectional DEA has previously been estimated in the academic literature for 171 acute hospitals for the year 1995/96.The analysis focused on the total cost, and technical efficiency scores were calculated using a range of outputs, including inpatient episodes, outpatient attendances, A&E attendances, teaching, and research.

The study considered five alternative model specifications and found that the greater the model complexity, the smaller the variation in efficiency scores across hospitals. Across the five models, the average efficiency score ranged from 75 to 99%.

Econometric techniques

A number of econometric techniques can be used to estimate efficiency. These approaches are typically robust to estimation uncertainty and could be advantageous when there are cost distortions within the sector. These can also account for uncontrollable factors that may impact costs. However, these techniques usually require the production function to be specified, which may be limiting when comparing providers with different production processes or intensity of resource utilisation.

[42] Jacobs, R., Smith, P., and Street, A. (2006), "Measuring Efficiency in Health Care: Analytic Techniques and Health Policy".

Box 3: SFA precedent in the health care sector[43]

An academic study used an SFA model to estimate efficiency, using a specification and dataset developed by the DH based on 226 acute hospitals for the year 1995/96.

This involved regressing total hospital cost on several variables including the number of inpatient admissions, outpatient attendances, patient demographic information such as average patient age and gender, and MFF. They also accounted for case-mix and specialisation.

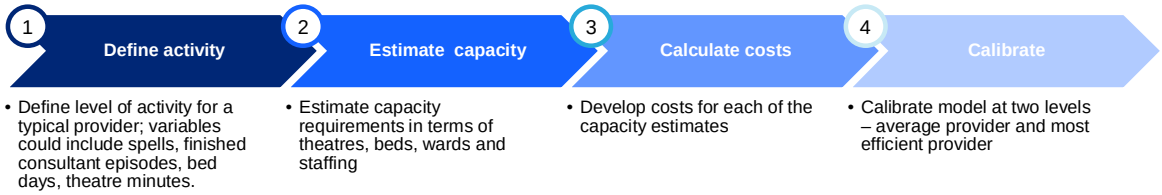
The study investigated the sensitivity of the SFA results across alternative specifications. The average efficiency score was consistent across different models, ranging from 85 to 90%. Relative provider-specific efficiency scores were also consistent across the alternative models

In common with simple benchmarking and productivity indexing, these approaches allow for the estimation of the historic shift in the frontier. It is noted that where a new frontier may need to be defined, around a new model of care, these approach could be less applicable.

5.2.4 Bottom-up modelling

Bottom-up modelling involves developing a detailed supply side model of an efficient provider through a bottom-up estimation of capacity required with regard to certain operational, clinical and quality metrics. Whilst top-down econometric approaches provide important insight on the potential levers of efficiency, a bottom-up framework may be useful in understanding how these levers could be used to generate the efficiency savings identified by top-down methods. Typically, a bottom-up model will involve four overarching steps, outlined below in Figure 6 below.

Figure 6: Bottom-up modelling steps



5.2.5 Other approaches

Bottom-up reviews

Bottom-up reviews involve a detailed examination into the expenditure and production processes of the provider in order to identify potential cost saving initiatives that could be introduced and the efficiency gains that can be achieved. Providers currently conduct bottom-up reviews when

[43] Jacobs, R., Smith, P., and Street, A. (2006), "Measuring Efficiency in Health Care: Analytic Techniques and Health Policy".

developing their CIPs. The granularity of these reviews can vary significantly, from full transformation plans with specific workforce and capital implications though to summary descriptions of high level benchmarking identified opportunities.

These studies can provide some understanding of the achievable efficiencies within the sector and as such could support the development of the appropriate efficiency target.[44] However, setting the efficiency factor on the basis of these plans could potentially create perverse incentives around the savings described, as providers acknowledge that their plans will determine the efficiency factor.

Best practice and precedent

International best practice and precedent can also provide insight on achievable efficiencies within the sector. This could highlight innovative delivery models that have been successfully used to generate cost savings. These approaches are less data driven and could be used when the efficiency factor is being estimated for services for which limited evidence is available.

[44] CIPs were considered as part of the evidence base underpinning the 2014/15 efficiency factor

6 Estimation challenges

This section examines typical efficiency estimation challenges in health care and discusses some of the potential mitigations. It is important that Monitor is transparent in its communications to the sector regarding the limitations associated with any efficiency analysis.

6.1 Accounting for quality

Quality is challenging to measure in health care given that it is largely unobserved and hard to verify. [45] There are a number of indicators that can be used to proxy for quality, but these cannot always be uniformly applied and compared across providers and service groups. Further, there are often challenges with quality indicators also indicating underlying population need. For instance, high hospital standardised mortality rates may be indicative of poor quality of care, or due to a higher elderly population in the local catchment area. [46]

These measurement issues may make it difficult to control for quality in the efficiency estimation process.

Suggested mitigations

- **Controlling for local population needs.** The estimation process could control for, where possible, underlying population needs by including relevant demographic indicators, for example, indices of multiple deprivation, [47] percentage of people over the age of sixty-five or ethnic composition.
- **Excluding extremes.** Estimation could be conducted on a truncated sample, excluding providers who are outliers. In terms of quality providers performing very poorly on different quality metrics, for example trusts that have been flagged as major concern and are under Care Quality Commission ("CQC") enforcement could be excluded from the estimation sample.

[45] As in Halonen-Akatwijuka and Propper (2013), "Competition, Equity and Quality in Healthcare".

[46] For example in Goodacre et al. (2013), "What do hospital mortality rates tell us about quality of care?"

[47] Indices of Multiple Deprivation (IMD) provide a relative measure of deprivation along seven different dimensions including income, employment, health, disability, education, crime, barriers to housing and services and living environment.

- **Using a range of different quality proxies.** The estimation process could incorporate multiple proxies covering the different quality dimensions of patient safety, experience and outcomes.

- **Including specific services.** Some services have better data around outcomes and experience, for example, providers are required to collect Patient Recorded Outcome Measures (PROMs)[48] for certain orthopaedic procedures such as hip replacements. Efficiency estimation could focus around services with the greater data availability.

6.2 Controlling for case-mix

The complexity of case-mix can impact the costs of service provision. [49] Providers with a more complex case-mix may incur higher treatment costs, which may not be due to inefficiencies in service provision. For example, for acute providers, there may be large variation in resource consumption even within HRGs due to factors such as age, existence of comorbidities or other complexities. Similarly, under adult and elderly mental health services, service users are classified into clusters according to the perceived level of need. However, the needs of the service users are not just dependent on clinical diagnosis but could also depend on complexity factors such as carer status, English as a first language and substance abuse. For example, the provider may need to have translation services in place for a patient who does not speak English. This may lead to significantly higher treatment costs and skew costs within the cluster.

Suggested mitigations

- **Demographic indicators.** Using demographic indicators partially captures the complexity of case-mix. For example, there is evidence to suggest that people from Black, Asian and minority ethnic ("BAME") groups have specific health requirements. For example, Black Caribbean people are reported to have high rates of hypertension and a higher probability to contract sickle cell disease; all ethnic minority groups are reported to have high rates of diabetes, and Black Caribbean people, particularly young men, have high rates of admission to hospital with severe mental disorders.[50] This could be partially accounted for by controlling for the proportion of ethnic minorities in the catchment area.

- **Stratified sample.** Providers could be stratified to account for different case-mix. For example, specialist teaching hospitals may have a more complex case-mix than district general hospitals.

---

[48] DH (2009), "Guidance on the routine collection of Patient Reported Outcome Measures (PROMs)".

[49] As in Street and Daidone (2011), "Estimating the costs of specialised care".

[50] House of Commons Health Committee (2009), Health inequalities, third report of session 2008-09, Volume 1

- **Complexity index.** Indices could be constructed to proxy for complexity. For example, by using HRG specific costs to capture case-mix complexity. [51]

6.3 Cost distortions

Distortions in efficiency estimation may also arise from underlying differences in costing. Previous audits of reference costs have identified inappropriate inclusions or exclusions of services, poor arrangements in cost allocation and material overstatement of activity to be some of the potential reasons for material errors in submissions.[52] This could lead to systematic differences in costs and therefore bias efficiency estimates.

Further, in the absence of prescriptive rules on allocation, there may be wide variation in reported costs due to differences in allocation. In the review of reference costs undertaken, 75% of the trusts were recommended to undertake a review of their cost allocations. Material differences in allocation were particularly observed in non-admitted patient care.[53]

Distortions have also been identified in the reporting of costs across elective and non-elective services.[54] In particular, there have been concerns regarding the over-reporting of costs for elective services and under-reporting for non-elective, potentially resulting in distortions in reported costs across providers. [55]

Suggested mitigations

- **Payment by results ("PbR") income.** A potential control could be the proportion of PbR to non-PbR income within the trust. This could control for the potential magnitude of cross-subsidisation.

- **Treatment of shared costs.** Shared costs could be excluded from the efficiency estimation process.

- **Differentiating providers.** Controls could be introduced where systematically higher costs are identified. This could be by including dummies or stratifying the sample to base estimates on different provider types.

---

[51] As in Jacobs, R., Smith, P., and Street, A. (2006), "Measuring Efficiency in Health Care: Analytic Techniques and Health Policy". In particular, they created a weighted inpatient activity index where the weights reflected HRG national average costs

[52] Audit Commission (2011), PbR Annual Report 2011

[53] Ibid

[54] Monitor (2013), A fair playing field for the benefit of NHS patients, Monitor's independent review for the Secretary of State for Health

[55] Ibid

6.4 Dealing with causality issues

Provider costs may be driven by a number of factors, including activity, case-mix, local demographics, input price inflation and wages. It will be important to identify which of these factors are within the providers' control and thus potentially driving inefficiencies.

The causality of the relationship between costs and its drivers is sometimes ambiguous. For example, complexity of case-mix is a driver of higher costs, but providers may also exercise discretion in their case-mix by treating only low complexity patients. Similarly, providers may choose their preferred activity levels by controlling the available capacity.

Suggested mitigations

- **Short and long-run.** A distinction between the factors which are within the providers' control in the short-run can be made. Where a short-run argument is plausible, estimation issues could be significantly reduced.

- **Using advanced econometrics.** A number of econometric techniques are available to deal with such estimation issues, for example, instrumental variable analysis.

7 Role of discretion

Setting the efficiency factor will involve a degree of judgement and discretion. This is important given the challenges related to estimation and Monitor's wider duties as the sector regulator. In exercising such discretion, however, Monitor should establish a clear rationale for its decisions. This section sets out a method through which discretion can be more systematically employed, including establishing a set of criteria which could be used to evaluate decisions.

7.1 Areas of discretion

Efficiency estimation typically results in a range of estimates for the catch-up and frontier shift. This range is likely to be the result of imperfect estimation and different methods measuring slightly different aspects of efficiency. Given this range, Monitor must determine a point estimate to use within price setting. Applying regulatory discretion is particularly important given the known shortcomings of individual estimation techniques and uncertainties around the data.

Discretion is further required in the determination of the appropriate glide path. The glide path is the proportion of the catch-up to the efficient frontier required within the tariff cycle. This would require judgment around the potential efficiency savings achievable in the context of the tariff life-cycle.

Decisions around the efficiency factor and glide path are ultimately related; they both determine the overall efficiency requirement in any given year. Given they are related, it is important that both their impacts are considered against the criteria established in this section.

7.2 Criteria to inform judgement and discretion

In exercising discretion Monitor should consider three factors:

1. Impact on the sector;
2. Link to Monitor's overall strategy; and
3. Sector views.

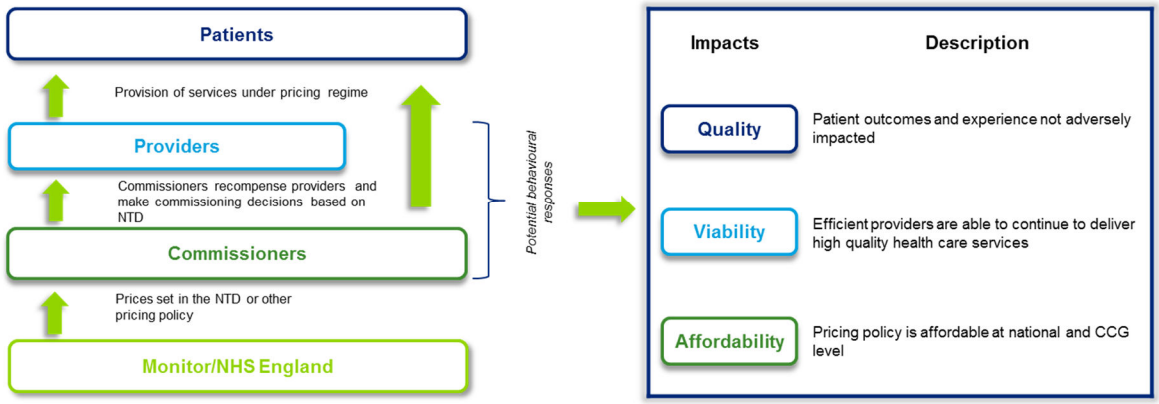
Impact on the sector

When determining the appropriate efficiency factor and glide path, Monitor will need to consider its main duty to protect and promote the interests of people who use health care services. The impact of different efficiency factors will therefore need to be determined in terms of their transmission from commissioners to providers and ultimately patients. For example, selecting an efficiency factor closer to the top-end of the range could lead to provider sustainability challenges, should the

estimation of the range be inaccurate. However, if the efficiency requirement can be met it could free greater resources for the purchase of health care services.

Monitor's pricing impact assessment ("IA") framework and tool provides an appropriate basis on which to consider such decisions. The IA framework defines three overarching impact areas to evaluate when setting pricing policy; affordability, viability and quality. These areas align to commissioners, providers and patients respectively. This is depicted in Figure 7. The degree to which formal impact assessment should be undertaken will depend on the level of uncertainty around the efficiency range estimated.

Figure 7: Impact framework for efficiency decisions



As part of determining the impact on the sector, the achievability of the efficiency factor will need to be evaluated, particularly in the context of the glide path determination. For example, should significant service reconfiguration be implied by the choice of efficiency factor, then sufficient lead time to achieve such efficiencies should be incorporated through a longer glide path. The IA framework could be used to identify the appropriate balance between commissioner affordability and provider viability.

Link to Monitor's overall strategy

It is important that decisions are also related to Monitor's and NHS England's joint pricing strategy. For example, setting a higher efficiency factor could incentivise different models of service delivery, such as integrated care.

Sector views

Monitor should critically consider the views of stakeholders as a further source of evidence regarding the achievability of efficiency. The views of the sector should be considered at different points in the determination of the appropriate efficiency factor. For example, Monitor should ideally engage more informally with the sector pre-formal consultation and draw on previous stakeholder consultations. Engaging the sector will also help to increase the transparency associated with efficiency setting.

Long term recommendations

8 Long term recommendations

The efficiency factor provides an opportunity for Monitor to establish positive incentives to ensure NHS services are efficiently and effectively delivered, ultimately benefiting patients. The current approach to efficiency, founded on a uniform factor, represents a compromise. Whilst having the advantage of simplicity and therefore being more easily estimated, the uniform factor does not capture the wide variation across the provider landscape.

This section sets out the long term recommendations for efficiency factor setting, developed using the framework described in the previous sections. [56]

- Section 8.1 assesses the alternative options that could be adopted for efficiency estimation leading to a recommended preferred approach for the long term summarised in Section 8.4;
- Section 8.2 considers how this data and preferred approach could be operationalised in terms of measuring relative efficiency;
- Section 8.3 considers the developments in data that would be required to allow the implementation of the preferred long term approach; and
- Section 9 details the recommended approach for the 15/16 national tariff in the absence of such data.

The recommendations established are provisional and should be further refined, incorporating key learnings from the 2015/16 efficiency estimation process and stakeholder feedback.

8.1 Options appraisal – level of efficiency

In the long-run Monitor should consider different levels of disaggregation of the efficiency factor. The appropriate level in the long term should suitably reflect the scope for efficiency savings, whilst being consistent with the incentives that Monitor wants to set.

[56] This section sets out the arguments around the appropriate level of efficiency and suitable estimation methods. For the remaining steps (definition of efficiency, estimation challenges and discretion), it is recommended that Monitor give due regard to the criteria within the framework

Option 1: Current state – uniform factor

Currently, a single efficiency factor is set, which applies uniformly across all nationally determined prices.[57] However, a uniform factor does not provide the scope to capture underlying differences in market conditions, existing and historic reimbursement systems, service delivery models and other factors which may affect the scope for efficiency savings. Further, a uniform factor does not allow for the setting of targeted incentives through the efficiency factor. As such, this approach is likely to be less appropriate for the longer term.

Option 2: Provider specific

The efficiency factor could be estimated and applied separately for each provider. This level of disaggregation would allow for a more gradual and sustainable shift to an efficient cost base for providers currently a significant distance from the frontier. Further, those closer to the frontier could be given more powerful incentives to continue to seek further efficiency opportunities.

However, setting the efficiency factor at a provider level could lead to some unintended consequences. For example, providers could have the opportunity to influence their future efficiency factors on the basis of their current outturn. Differential prices through provider specific factors could also incentivise commissioners to choose providers where a lower price is observed. In the context of current national prices, which aims to create competition on quality not price, this could be undesirable.[58]

Further, setting a provider specific factor requires significant confidence in the underlying estimation techniques. Moreover, such estimation approaches would need to accurately account for a wide variety of relevant uncontrollable cost drivers by provider, which could be challenging. These could include case-mix, quality and economies of scale which may vary significantly across providers.

Option 3: Provider type

Efficiency factors could also be disaggregated by provider type. This would involve setting differential factors reflecting systematic differences in the provider's scope for efficiency savings. Providers could be grouped according to various characteristics, including:

- Foundation trust and non-foundation trusts;
- Specialist/teaching and district general hospitals; and
- NHS, private and third sector providers.

---

[57] Monitor (2013), 2014/15 National Tariff Payment System

[58] This could also be somewhat abated by commissioners receiving offsetting allocation increases, achieving neutrality, where price differences are observed.

Further, this approach will rely on provider level data and therefore could be less prone to distortions created through potential misallocation of costs.

Despite these advantages, there are a number limitations associated with this approach.

- **Organisational neutrality.** The factor seeks to incentivise efficiency in service delivery and as such should be neutral to organisational structure. There may be instances where such structures impose uncontrollable restrictions on the ability to achieve cost savings. However, using high level groups is unlikely to capture these effects.

- **Variation across services.** Systematic differences in scope for efficiency savings could apply only to a subset of services offered by the organisation. For example, specialist teaching hospitals may incur systematically higher costs[59] for routine services whilst providing specialist services at efficient costs.

Option 4: Service group

An efficiency factor at a service level could be appropriate where there exists significant differences in current efficiency or anticipated differences in future demand and supply conditions. For example, some services maybe more conducive to the uptake of new technology and process, leading to a faster shift in the efficiency frontier. Further, past reimbursement systems may have led to differential incentives to achieve efficiency savings, potentially creating differences in scope for future efficiency gains.[60]

Service level efficiency factors could also support wider objectives around incentivising innovative models of service delivery, for example, support a move to greater integrated care.[61]

Estimating efficiency factors by service will require cost data at a granular service level. This may pose significant estimation challenges, given that service level data often encompasses significant allocation of shared and common costs.[62]

As a starting point, for example, the efficiency factor could be disaggregated at a higher level, by acute, community, mental health and ambulance services. This could have a number of advantages.

---

[59] These higher costs incurred are due to the use of specialist and teaching inputs in the delivery process. As such, these costs generate positive externalities for society and should not be attributed to inefficiencies.

[60] As future reimbursement mechanisms are developed, Monitor may choose to structure these in order to achieve desired incentives, and these may require their own efficiency adjustments.

[61] For example, service groups could be carved out across different care settings to set differential efficiency factors. To carve out differential targets, however, Monitor will need to understand the relative weightings of costs of provision of different care providers within integrated settings.

[62] This could be abated by aggregating services which have common costs and the use of PLICS data, as this becomes more robust and readily available. However, typically efficiency estimation approaches work more effectively where there are only limited cost allocations.

- **Feasibility.** Typically, provisioning of these four service groups are distinct and overlap with provider boundaries. Estimation at this level therefore, would mostly involve costs at total provider level and could potentially reduce allocation related distortions.

- **Delivery models across such services are likely to be different.** Mental health and community service provision for example, involve a higher proportion of direct staff costs, compared to most acute services. This could influence the current and future efficiency levels achieved.

- **Different reimbursement systems.** These services are in different stages of currency development and are typically reimbursed in different ways. For example, the majority of community services are still reimbursed through block contract arrangements, [63] which essentially make payments based on capacity rather than activity. Most acute services, on the other hand, have nationally determined prices.

- **Market conditions.** The demand and supply conditions across such services have historically been quite different. For example, the market for community services underwent significant transformation in 2011, with the separation of provider arms from Primary Care Trusts.[64]

Despite these advantages, groupings around these four service types could become less relevant as the market develops, whilst pricing becomes more pathway and patient centric. Further, one of Monitor's core responsibilities is to enable better integration of care, so services are less fragmented and easier to access.[65] As such, it is important that the level of disaggregation employed is not inconsistent with integrated care.

Option 5: Banding

Providers could be grouped into bands with different efficiency factors. These bandings would be based on the relative inefficiencies of the provider, as measured by their distance from the efficient frontier. A banded approach to efficiency setting could safeguard sustainability whilst keeping incentives in place for those who can achieve greater efficiencies.

However, such approaches rely significantly on the underlying estimation techniques and ultimately the robustness of key cost and activity information. Moreover, the estimation would need to accurately account for factors such as case-mix or scale in order to estimate relative inefficiencies and determine appropriate banding thresholds.

---

[63] Monitor (2012). Evaluation of the reimbursement system for NHS funded care

[64] DH Transforming Community Services programme; 2010/11 NHS Operating Framework and the Health White Paper

[65] HSCA (2012) Section 62

Further, there may be inherent risks of weaker incentives to achieve efficiency gains, if providers determine that their future factor could be influenced on the basis of their outturn.

Finally, in common with provider specific efficiency factors, and in the context of national prices, banding could lead to commissioners observing different prices by provider for the same service. This could be contrary to quality based competition unless commissioner allocations are also adjusted.

Preferred option

The current approach to efficiency, founded on a uniform factor, represents a compromise. Whilst simple to estimate, such an approach does not capture the wide variation across providers and services.

In the long term, the preferred level of efficiency should seek to capture differences in the scope for future efficiency gains, allowing for the setting of more targeted incentives whilst supporting sustainability. In order to achieve this, it is likely that the appropriate level of disaggregation comprises the following.

1. **Service groups.** Efficiency factors could be set differentially across different service groups. These groupings would be based on underlying market conditions and processes underpinning the provisioning of services. As outlined above, groupings could be based around acute, mental health, community and ambulance services as a starting point.

2. **Banding.** Within these service groups, providers could also be placed into bands with different efficiency factors. This could reflect differences in the scope for future efficiency gains, allowing for the setting of more targeted incentives whilst supporting sustainability.

8.2 Options appraisal - estimation methods

A number of different estimation methods could be used to estimate the efficiency factor. The appropriate method employed in the long term should be based on an assessment of the benefits and associated limitations.[66]

Option 1: Simple benchmarking

Simple benchmarking uses available cost, quality or activity data to measure a provider's performance in comparison to a set of comparator providers. Such methods are widely used within the sector as they are easy to understand and relatively simple to construct. However, there are a number of limitations associated with simple benchmarking which make its use undesirable. Specifically, these approaches struggle to control for provider differences which may be driving costing differences, for example, case-mix, quality differences and economies of scale. Although some adjustments can be applied to control for these factors, these are often based on simplified assumptions.

[66] These methods have been described in greater detail within framework chapter on estimation methods.

Option 2: Productivity indexing

Productivity is typically defined as a ratio comparing the total amount of output produced to the volume of input utilised in the production process. Productivity indexing could provide some insight into historical gains and trend rates in productivity growth. However, these approaches are typically sensitive to the choice of indexation method used to combine outputs and inputs.

Option 3: Data Envelopment Analysis

DEA is has been widely used to measure efficiency in the health care sector. This is a top-down approach which estimates the efficient frontier based on a consideration of different capacity combinations that could be used to meet a particular level of activity or demand. Although DEA is used relatively extensively in the literature, as previously noted, it can be highly sensitive to outliers and the selection of inputs and outputs. This could potentially be a limitation given the numerous cost distortions and differences in allocation methodologies across providers.

Option 4: Econometric techniques

Econometric techniques allow for the identification of both frontier shift and catch-up, whilst controlling for a range of different factors. These approaches could allow for the estimation of relative inefficiencies of individual providers, supporting banded efficiency factors. Further, models could be constructed for particular service groupings facilitating the setting of specific factors. However, such techniques could be data-intensive and relatively complex.

Option 5: Bottom-up modelling

Bottom-up modelling involves developing a detailed supply side model of an efficient provider through a bottom-up estimation of capacity required with regard to certain operational, clinical and quality metrics. These models are complementary to top-down approaches and could be used to identify potential efficiency opportunities within an organisation. However, these models provide information only around the catch-up efficiency and are typically constructed around a single provider scenario. The latter limitation means it may be difficult to use them for very broad conclusions.[67] Additionally, developing bottom-up models is time consuming and highly resource intensive.

[67] For example, a single bottom-up model would not be able to inform a banded approach.

Preferred option

A number of provider specific top-down econometric models to estimate the efficiency factor could be developed. These could be supported by bottom-up models and further literature review. Since this approach relies on complex modelling, Monitor will need to be transparent about the data it uses, any assumptions or manipulations and the precise techniques used.

It is recognised that the overall challenges in efficiency estimation, even in the longer term, are likely to be significant. To account for this, Monitor should continue to triangulate the findings and engage with the sector.

Across the techniques described, Monitor could seek to include private and third sector providers. Where sufficient data allows, international comparators could also be incorporated.

8.3 Data dependencies

There are a number of key data requirements which need to be fulfilled in order for the preferred options to be realised. The availability and quality of information varies significantly across different services. In order to realise an efficiency factor varying by service group and provider band, a wider coverage of different service groups will be required, in particular for mental health and community services.

- **Cost information.** For robust estimation of efficiency, cost information will need to be consistently reported and available at significant levels of disaggregation. Currently information which is available primarily comprises reference cost submissions. Although this data has been collected for a number of years, there are some known limitations highlighted in previous studies.[68] As such, patient-level information and costing systems ("PLICS") data should underpin efficiency factor estimation. The use of PLICS data provides the opportunity to be more targeted in terms of cost inclusions, as well as significantly increasing the sample size. However, this information will be required over a number of years[69], once a refined collection has been established.[70]

- **Activity information.** Volume information is required to ensure efficiency estimation accounts for the scale and mix of services provided. Currently, activity information can be obtained through various sources including, reference cost submissions and Hospital Episode Statistics ("HES"). The granularity of this information is relatively aligned to the preferred option in acute secondary care. However, significant improvement in the reliability and detail is required in mental health and community services, in particular with regard to comparability of information. As such, this may require the development of a more standardised service definition for community activity data and consistency in clustering across mental health providers.

---

[68] Monitor (2012), Evaluation of the reimbursement system for NHS-funded care

[69] At least two to three years of information may be required in order to capture potential time trends.

[70] Monitor is currently undertaking an exercise which will define the future direction of travel for cost information for the service groups identified. It is recommended that this piece of work acknowledges the specific requirements outlined.

- **Quality information.** The quality of the service provided needs to be controlled for when estimating efficiency. At present, there are a broad range of different indicators to proxy for quality. In common with activity and cost information, there are greater gaps in available quality metrics for community and mental health services. A pre-requisite to estimating efficiency for these services will be to have a number of relevant measures, established on the basis of a common framework.

- **Other information.** Information will also be required capturing local health needs and supply side conditions, including demographics, deprivation and prevalence of different conditions.

8.4 Summary recommendation

The choice of the appropriate level of efficiency has been based on theoretical considerations only. As such, it is recommended that Monitor tests the underlying hypotheses supporting service grouping and provider banding.

A number of example tests are described as follows. In addition to these formal tests, Monitor should also seek to engage with the sector around these issues.

Testing service grouping

- **Top-down tests.** On the fulfilling of the key data dependencies, econometric investigations can consider the range of efficiencies prevalent across different service groupings. Where material variation is observed across groups, this could be supportive of differential groupings.

- **Bottom-up tests.** Before data dependencies are realised, some initial testing could be undertaken through the development of bottom-up models encompassing different service groups. However, such models are unlikely to provide fully conclusive evidence due to the wide variation of operating models, case-mix and scale observed across providers.

Testing provider bands

- **Provider efficiency tests.** In order to indicate that banding maybe appropriate, the stability of estimates over time could be considered. If this is observed, it could indicate that a degree of confidence can be placed on the underlying estimation approach.

- **Provider ranking tests.** There are various top-down methods which can be employed to estimate efficiency. To be confident in a banded approach, consistency in the findings of these methods should be sought.

The evidence base is currently more developed for certain services, particularly within acute secondary care.[71] This may support a phased approach to testing and evidence gathering and Monitor could start to undertake testing in areas where greater information is available.

[71] For example, Monitor recently completed the first round of voluntary PLICS collections covering 66 acute providers.

Recommended approach for 2015/16

9 Recommended approach for 2015/16

Section 8 makes a number of recommendations that Monitor could use to inform efficiency factor estimation in the medium and long-run. However, in the short-run, there may be a number of limitations impacting the feasibility of the recommended approach. This section sets out the proposed approach for the 2015/16 efficiency factor, in light of the long-run recommendations and relevant constraints.

9.1 Level of efficiency

9.1.1 Limitations relevant to 2015/16

In the long term, it is recommended that Monitor consider an efficiency factor disaggregated by service groups, if there is sufficient and robust evidence to suggest material differences in scope for efficiency savings. However, the potential benefits associated with a more disaggregated efficiency factor for 2015/16 are restricted due to a number of overarching limitations.

- **Data and estimation challenges.** The data available to underpin the estimation of efficiency varies across service groups. For example, in secondary care there is a significant amount of data available; including voluntary patient-level costing information collected for 2012/13. In community care, however, service definitions are often unclear, reducing the ability to measure the scale and acuity of services being delivered in any estimation of efficiency.

- **Evidence base surrounding disaggregation.** There is limited evidence to examine if there are material differences in scope for efficiency savings across the service groups, as recommended. As part of the estimation of the 2015/16 efficiency factor, some initial testing around the presence of different supply and efficiency conditions should be possible. However, any evidence is likely to require further refinement before being sufficiently accurate to be used for price setting.

9.1.2 Summary recommendation

These limitations suggest that in the short term, a sector wide efficiency approach is likely to be more appropriate. However, this approach will need to carefully consider how to account for those providers who are furthest away from the efficient frontier. Moreover, an efficiency target which seeks to establish pricing at the frontier could risk the sustainability of some providers. To mitigate this, Monitor will need to exercise a degree of discretion along with an appropriate impact assessment.

9.2 Estimation method

9.2.1 Limitations relevant to 2015/16

In the long term, it is recommended that top-down econometric approaches and bottom-up modelling be deployed as core methodologies to estimate the efficiency factor. However, there may be a number of challenges with regard to the availability of robust underlying information which may limit the scope of their application for 2015/16. These are set out in the remainder of this section.

Econometric methods

The ability to undertake such analysis is contingent on the availability of significant data. In particular, information is required on:

- Cost incurred to deliver services;
- Volume and mix of activity provided;
- Uncontrollable factors including, local health needs and supply side conditions; and
- Quality of service delivered.

Whilst there is sufficient data covering these areas for acute secondary services, there are significant gaps for other services. For example, mental health reference costs are currently captured around care clusters.[72] However, due to the underlying variation in coding across providers, cluster based reference costs may not be a reliable basis for efficiency estimation.[73] Given these gaps, it is recommended that the econometric analysis focuses on providers of acute services for 2015/16.

The potential limitations of this focus are set out at the end of this section.

Bottom-up modelling

As discussed in Section 8, bottom-up models should be developed for different service groups to gain an understanding of the different sources of efficiency. It is recommended that a single secondary acute provider model is developed for 2015/16. Focussing on an acute provider for this year is likely to be preferred given:

[72] Clusters are the currency for adult and elderly mental health services, and were introduced in the 2011/12 reference cost collection.

[73] Appendix E summarises the available sources of cost and activity information across different service groupings.

1. The efficiency factor applies directly to nationally determined prices, which mainly comprise acute services; [74] and

2. There is a more consistent service definition when compared to other settings of care, such as community services.

In subsequent years, Monitor could look to develop a more comprehensive suite of models covering mental health, community and ambulance providers.

9.2.2 Summary recommendation

It is recommended that econometric analysis and bottom-up modelling are conducted to support the estimation of the 2015/16 efficiency factor. Econometric analysis could provide both a measure of the frontier shift and relative efficiency. The bottom-up model could more explicitly identify the efficiency opportunities to improve relative efficiency. The scope of the investigation, however, will be limited for 2015/16 and can focus only on acute providers.

Econometric methods

Econometric techniques could be used to estimate both the frontier shift and catch-up component. These approaches allow for a range of differences across providers to be controlled for when estimating efficiency. Such factors could include scale, case-mix, quality, local health needs and uncontrollable cost conditions.

The econometric analysis would focus on measuring efficiency based around two sources of cost and activity data; reference cost and PLICS. Reference cost data covers around 250 NHS providers with the HRG4 currency from 2009 onwards. Despite some concerns regarding the quality of this information at a more granular level, aggregate costs and activity could form a more reliable basis for estimation. PLICS information has been collected by Monitor on a voluntary basis for 2012/13. This database provides a more granular breakdown of cost pools and activity, given these fields are recorded at a patient level. However, this has only been completed by 66 providers and the lack of a time dimension to the data limits its use for the estimation of potential frontier shift.

A number of econometric estimation methods are likely to be appropriate given the datasets available. These could include COLS, SFA and other panel based approaches.

Bottom-up modelling

A bottom-up estimation of efficient costs could be undertaken to understand the potential opportunities for efficiency improvement for a provider. This would help to define achievable efficiencies and support Monitor's decision around the efficiency requirement and glide path.

[74] Whilst not mandatory, the efficiency factor is often used as an anchor point for local price setting.

It is recommended that an actual provider in the sector is used as a basis for developing the model. This case study approach would support a more realistic estimation of the costs, and potentially efficiencies, within the model. Whilst a 'typical' provider is difficult to define, consideration of a range of provider characteristics including financial, quality and operational metrics would support an informed choice for the case study.

A baseline model of service provision would firstly be constructed around average efficiency assumptions. A number of levers would then be applied to estimate achievable efficiency improvements. For a secondary acute provider these levers are likely to include:

- Improving the utilisation rate of theatres and wards;
- Reducing length of stay, whilst maintaining or improving patient quality and readmission rates;
- Optimising staffing ratios; and
- More effective procurement of drugs and consumables.

A suite of efficiency opportunities would be developed through a review of current CIPs and relevant literature. The choice of efficiency levers for inclusion within the model would need to reflect the one year tariff lifecycle. For example, whilst better utilisation of estates could lead to efficiency savings, these may be realised only over a number of years. As such, the model would make a distinction between controllable and non-controllable costs relevant to the one year time-frame.

Figure 8: Overarching architecture

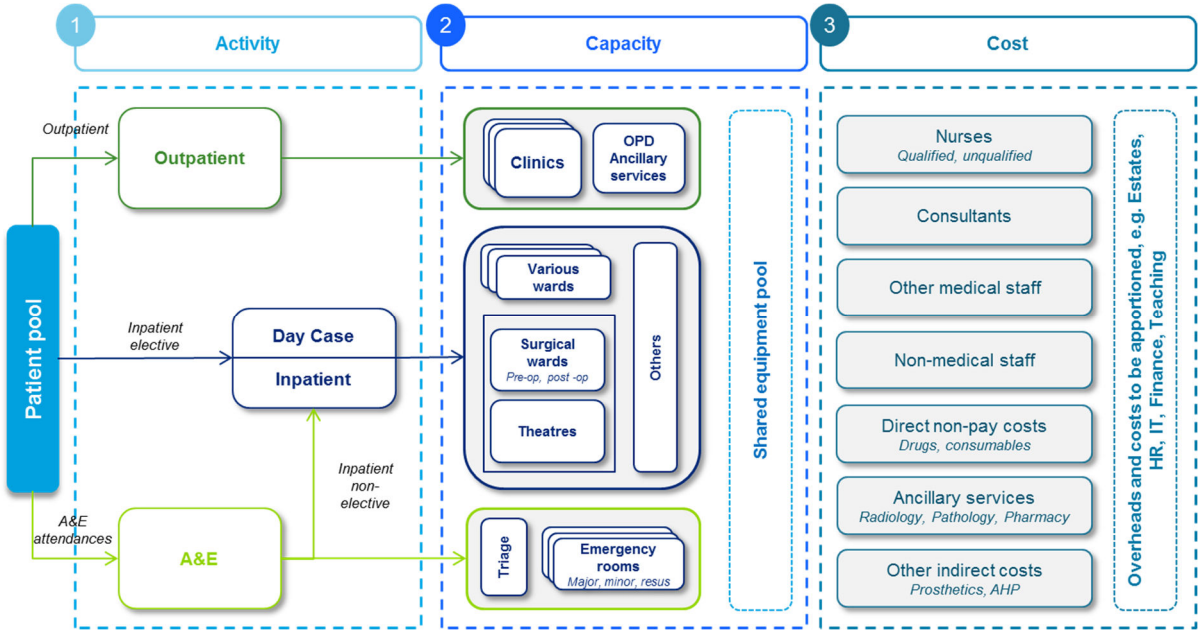


Figure 8 illustrates the overarching model architecture. The baseline model will comprise three stages:

1. **Definition of activity.** Activity will be defined across the different points of delivery. The model will be flexible to changes in patient flow that may lead to efficiency improvements, for example, increasing the proportion of day case admissions.
2. **Modelling of capacity.** The model will then estimate the capacity required to meet the activity levels defined. This could include physical capacity, such as the number of beds, shared equipment and medical or non-medical staff.
3. **Estimation of costs.** Costs will be estimated on a bottom-up basis across different points of delivery. For example, when constructing a bottom-up model of inpatient ward costs, ward establishment ratios for medical and non-medical staff and other non-pay costs such as drugs and consumables will be considered. Ancillary services and shared equipment costs could be apportioned on a weighted basis, reflecting activity.

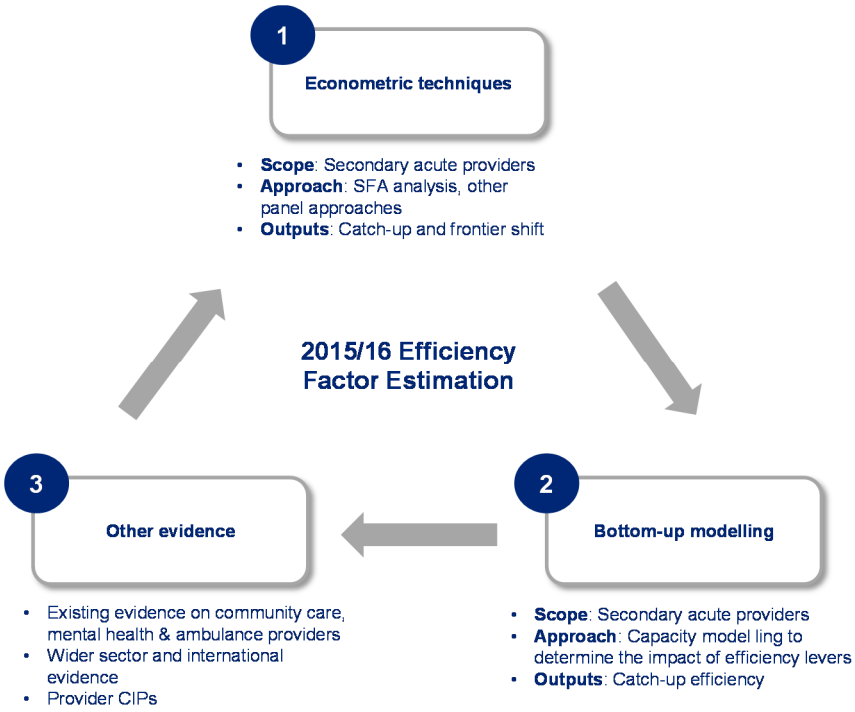
9.2.3 Limitations

The two estimation procedures suggested are focussed around measuring efficiency in secondary acute provision. This is driven by greater data availability, more established currency and service definition and the greater direct application of the efficiency factor in current national pricing. For these reasons, it is likely that estimating efficiency using acute provision will be more accurate. However, it should be acknowledged that the services and models of care can be quite different across other service groupings, potentially causing some challenge regarding its wider application. It is therefore recommended that Monitor engage with stakeholders to identify the suitability of the estimate derived from the acute sector for community, mental health and ambulance providers.[75] Further, engagement with these providers will also support a better understanding of any limitations and mitigations.

Given these limitations it is recommended that the proposed approach should triangulate various evidence for the efficiency factor, as summarised in Figure 9.

[75] It is important that these additional measures are, however, weighted appropriately in determining the overall efficiency factor, given they may be less robust.

Figure 9: Triangulation of estimates



Triangulation is likely to produce a range of estimates. Monitor will therefore need to exercise a degree of discretion to select the appropriate factor. It is recommended that discretion is undertaken in a transparent manner with reference to specific criteria. Particular considerations could include the impact on the sector and Monitor's and NHS England's joint overall pricing strategy, as set out in Section 7. For example, in order to establish the sector impact, Monitor could measure changes in provider revenue and net surplus across the range of efficiency estimates.[76]

9.3 Estimation challenges

Estimation of efficiency is challenging and Monitor needs to be clear about the limitations associated with any estimates made. This section considers the limitations and suggested mitigations, associated with the recommended approach for 2015/16.

9.3.1 Econometric techniques

The econometric methodology should consider a number of measurement and estimation challenges.

- **Measuring the quality of service.** Quality is challenging to measure in health care given that it is largely unobserved. There are a number of indicators that can be used to proxy for quality, but these cannot always be uniformly applied. For instance, high standardised

[76] This analysis would need to be based around activity assumptions and scenarios regarding the level of efficiency achieved.

hospital mortality rates may be indicative of quality, or due to the local needs of health economy.

Mitigations

- ✓ Multiple proxies. The estimation process should consider multiple quality measures covering different dimensions including patient safety, experience and outcomes.
- ✓ Population needs. Estimation should control, where possible, for underlying population needs by including relevant demographic indicators including indices of multiple deprivation, the percentage of people over the age of sixty-five and ethnic composition.

- **Controlling for case-mix.** Complexity of case-mix can impact the costs of service provision. Providers with a more complex case-mix may incur higher treatment costs, which may not be due to inefficiencies in service provision. Case-mix is challenging to control for and providers with a more complex case-mix often have higher costs for other reasons, such as additional teaching activities.

  Mitigations

  Case-mix can be captured using alternative proxies.

  - ✓ Demographic indicators. Demographic indicators partially capture the complexity of case-mix. For example, there is higher diabetes prevalence among the South Asian population which could contribute to increased treatment costs.[77]
  - ✓ Stratified sample. Providers could be stratified to account for differences in case-mix. For example, specialist teaching hospitals may have a more complex case-mix than district general hospitals.
  - ✓ Complexity index. Indices could be constructed to proxy for complexity. This could be captured by using HRG specific costs. A weighted inpatient activity index, where the weights reflect HRG national average costs, has previously been used in the literature.[78]

- **Dealing with causality issues**. The causality of the relationship between costs and its drivers is sometimes ambiguous. For example, the complexity of case-mix is a driver of higher costs, but providers with high costs or inefficiencies may choose to try and reduce case complexity by treating lower complexity patients. Similarly, achieving a higher quality

[77] British Medical Journal (2000), Association of glycaemia with macro-vascular and micro-vascular complications of Type 2 diabetes: prospective observational study.

[78] As in Jacobs, R., Smith, P., and Street, A. (2006), "Measuring Efficiency in Health Care: Analytic Techniques and Health Policy".

of service typically requires greater expenditure; but inefficient providers may choose to provide services at a lower quality in an attempt to make cost savings.

Mitigations

- ✓ Advanced econometrics. A number of econometric techniques are available to deal with such estimation issues, for example, instrumental variable analysis and panel fixed effects. However, it should be noted that these techniques are often challenging to implement.

9.3.2 Bottom-up modelling

A number of estimation challenges may arise when building a bottom-up model.

- **Complexity of patients.** Complexity of patients will impact on ward and theatre costs. For example, some wards may require higher intensity of staffing due to the higher acuity of patients.

  Mitigation

  - ✓ Modelling multiple ward or theatre types. This could be mitigated against by modelling multiple ward or theatre types, grouped according to complexity. For example, theatres could be modelled separately for emergency or trauma and elective or scheduled procedures. This classification could be validated by clinicians.

- **Ward staffing.** Recommended staffing levels and grade-mix is very specific to organisations. Whilst there are Royal College of Nursing guidelines for different ward types, they are not mandated and are often significantly different from what is observed amongst providers. This variation makes it difficult to define efficient and clinically sustainable staffing levels.

  Mitigation

  - ✓ Validated establishment ratios. Ward establishment ratios for the provider could be validated by clinicians or sector experts.

- **Indirect and overhead costs.** There may be scope for efficiency savings within indirect and overhead costs. However, these are difficult to incorporate in a bottom-up model.

  Mitigation

  - ✓ Exclude or apportion. These costs are unlikely to be a major source of efficiency savings in the short-run and hence could be excluded from the model; particularly in the case of overheads.

Appendices

Appendix A Review of precedent

As discussed in Section 2, a range of precedent across a number of sectors has been considered to inform the framework for the estimation of the efficiency factor. A list of the main references, split into each regulated sector considered, is provided below.

Electricity and gas

1. Cambridge Economic Policy Associates (2009), Ofgem, The use of RPI-X by other network industry regulators;
2. Ofgem (2010), RIIO, a new way to regulate energy networks, Final decision;
3. Ofgem (2010), Handbook for implementing the RIIO model;
4. Ofgem (2012), RIIO-GD1: Final Proposals – Overview;
5. Ofgem (2012), RIIO-GD1: Final Proposals - Supporting document - Cost efficiency
6. Ofgem (2012), RIIO -T1/GD1: Real price effects and ongoing efficiency appendix
7. Ofgem (2012), Electricity Distribution Annual Report for 2010-11;
8. Ofgem (2012), RIIO-GD1: Initial Proposals – Step-by-step guide for the cost efficiency assessment methodology;
9. Ofgem (2013), Price controls explained;
10. Oxera (2013), Recommendations on cost assessment approaches for RIIO-ED1;
11. London Economics (2011), Estimating the Value of Lost Load, a report for Ofgem.

Water and sewerage

12. Ofwat (2009), Relative efficiency assessment 2008-09 – supporting information;
13. Ofwat (2009), PR09 final determinations key headlines;
14. Ofwat (2012), The form of the price control for monopoly water and sewerage services in England and Wales – a discussion paper;

Communications

15. Ofcom (2012), Mobile call termination, Adoption of revisions to SMP Conditions in accordance with the directions of the Competition Appeal Tribunal of 8 May 2012;
16. NERA (2008), The Comparative Efficiency of BT Openreach, a Report for Ofcom;

Post and mail

17. Ofcom (2005), 2006 Royal Mail Price and Service Quality Review.

Railways

18. Cambridge Economic Policy Associates (2013), Office of Rail Regulation, Update report on the scope for improvement in the efficiency of Network Rail's expenditure over CP5;
19. ORR (2013), PR13 Efficiency Benchmarking of Network Rail using LICB;

Other precedents

20. Oxera (2011), When Less is More: reducing regulatory judgement.
21. BIS (2011), Principles of Economic Regulation;
22. HM Government (2013), Streamlining Regulatory and Competition Appeals;

Appendix B Level of efficiency

Section 4 discussed the various levels of efficiency that could be applied when estimating the efficiency factor. This appendix sets out some of the potential groupings for provider and service groups.

B.1 Service groupings

A number of potential groupings at a service level are provided below in Table 3.

Table 3: Example service groupings for efficiency setting

Service grouping	Description
Point of delivery	Services could be grouped by point of delivery, reflecting difference efficiency opportunities. For example, reducing the number of did-not-attends is likely to be a key source of efficiency savings in outpatient departments. There are five points of delivery considered in the national tariff; A&E, non-elective, day case, elective and outpatient.
Clinical division/Specialty	Groupings could be based around clinical divisions or specialty. Clinical services may be grouped into divisions for internal management purposes and may vary by hospital. For an acute provider these typically include: • Women and Children • Acute Medicine • Surgery • Critical Care These divisions are likely to have shared indirect costs.
Currency/payment systems	The efficiency factor could also be disaggregated at the individual currency level, for example, for acute services this could be by individual Health Resource Group (“HRG”) chapter or subchapter, or for mental health at the cluster level.
Service baskets	Efficiency targets could be set for particular service baskets, such as for pathways. Pathway based efficiency targets could lead to significant cost and efficiency savings whilst incentivising integrated care.
Other	Other potential groupings could be developed in alignment with Monitor’s and NHS England’s joint long term pricing and payment strategy.

B.2 Provider type groupings

A level of possible disaggregation for the efficiency factor is at a provider type level. Some of the potential provider type groupings are listed in Table 4 below.

Table 4: Grouping by provider type

Factor	Description	Potential grouping/variable
Incentive to invest	There may be certain distortions in the market which limit providers' ability to invest, for example the ability to retain and re-invest surpluses, as foundation trusts are currently able to do.	• Foundation trust status and non-foundation trust
Differential cost base	Providers may also have different cost bases, which may impact the ability to make efficiency savings. For example, NHS providers can claim back value added tax ("VAT") on certain contracted out services (catering, child-care, laundry, purchasing and procurement services); however the independent sector providers cannot reclaim this cost.[79]	• Independent and third sector providers
Differential cost of financing[80]	Providers may have differential costs of financing. Cost of financing private finance initiative ("PFI") debt is significantly higher than other sources. PFI contracts are usually inflexible, involving long term leases and non-negotiable payments. Further, PFI charges on debt interest are linked to the RPI measures of inflation which tend to be higher than other measures such as the GDP deflator, which is the economy wide measure used for public services.	• Trusts with or without PFI funding
Organisation type	Costs of service provision may be higher in teaching hospitals due to the additional resources devoted to education and training.	• Teaching hospital • District General Hospital • Specialist hospitals
Banding	Providers could be banded based on their relative inefficiency, as measured by their distance from the efficient frontier. In order to promote provider sustainability, alternative glide paths could be considered by each band.	• Various bands

[79] Street *et al* (2008). "Establishing a Fair Playing Field for Payment by Results"

[80] Monitor (2013), A fair playing field for the benefit of NHS patients, Monitor's independent review for the Secretary of State for Health

Appendix C Stakeholder Engagement

Incorporating the views of providers, commissioners and technical experts from both the health care and other regulated sectors has played a key role in determining the overall framework. The key findings from this process have been summarised in Table 5 below.

Table 5: Stakeholder engagement feedback

Main themes	Key points
Current state	Transparency. Acknowledging progress made, there were some suggestions that the determination of the efficiency factor could be more transparent. For instance, some providers indicated that they believed the efficiency factor in the past reflected several savings which are more under the control of commissioners. Multi-year tariff. Setting a multi-year tariff rather than single year could give providers more flexibility to plan effectively and potentially achieve greater efficiency savings by introducing initiatives that require more fundamental changes Simplicity. It is generally regarded that a key advantage of the current efficiency factor framework is its simplicity and ease of implementation. Commissioners and providers use the efficiency factor more widely in pricing off national tariff.
Disaggregation of efficiency factor	Service-specific. Determining the efficiency factor by service group could be appropriate where demand or supply conditions are expected to be sufficiently different in the future or where there are large differences in current efficiency. There were concerns, particularly in the short-run, that any disaggregation may be inaccurate due to data quality issues at a service-specific level. Provider-specific. There was an acknowledgement that the range of efficiency is likely to be significant. Given this range, it could be argued that different catch-up requirements could support providers in more sustainably adjusting to the frontier. However, concerns over data quality to allow for provider specific factors were raised.
Measurement challenges	Case-mix. It was generally agreed that accounting for case-mix is the greatest challenge in comparative benchmarking analysis in the health care sector. Triangulation. It was generally agreed that combining information from alternative sources and using different methodologies to estimate the efficiency factor generally reduces estimation uncertainty. Glide path. The length of the glide path to the efficient frontier should be determined by a range of factors, including critically the time to implement and receive benefits from efficiency initiatives. The glide path could be within or beyond the period of the tariff. Estimation sample. The estimation of the efficiency factor could be carried out using data from a sub-set of providers. This could support the use of some PLICS analysis from the current trial of PLICS data collection.