The Strategy Unit.

Risk and Reward Sharing for NHS Integrated Care Systems

June 2018

Prepared by:
Steven Wyatt
Head of Strategic Analytics
T: 0121 612 3872
M: 07702 444029
swyatt@nhs.net

NHS
Midlands and Lancashire
Commissioning Support Unit

Document control

Document Title	Risk and Reward Sharing for NHS Integrated Care Systems
Job No	422
Prepared by	Steven Wyatt
Checked by	Peter Spilsbury, Fraser Battye, Sharon Townsend, Paul Seamer
Date	June 2018

Acknowledgements

This report was commissioned and funded by NHS Midlands and Lancashire Commissioning Support Unit.

Thanks to Paul Taylor, Provex Solutions Ltd, Jessica Boothroyd, NHS Transformation Unit, Peter Spilsbury and Fraser Battye, Strategy Unit, for their insightful comments on the draft report.

Executive Summary

Risk and reward sharing is a key feature of the policy agenda for Accountable Care Organisations in the US and Integrated Care Systems in England. It is a simple and attractive concept, offering a commissioner the opportunity to co-opt and incentivise a provider to moderate growth in healthcare demand by sharing in the savings or cost over-runs. Risk and reward sharing can be seen as an end in itself, or as step in the journey towards capitated budgets and the full transfer of financial risk to providers.

The rationale that underpins the development of Accountable Care Organisations in the US and Integrated Care Systems in England is similar: to moderate healthcare costs through service coordination and integration. However, US ACOs and English ICSs are vastly different in scale (on average, US ACOs provide services to c. 19,000 enrolled patients) and operate in radically different political, financial and cultural contexts.

The Centers for Medicare and Medicaid Services (CMS), a US government agency, has established a comprehensive approach to risk and reward sharing for US Accountable Care Organisations: the Shared Savings Program. This paper draws out the central themes from the Shared Savings Program and translates these into an NHS context.

In the US, ACOs are required to sign up to one of three risk-sharing 'tracks'. Track 1 is a one-sided risk-sharing model where providers have the potential to share in savings if priced activity falls below expected levels, but are not required to pay a share of any cost over-runs. Tracks 2 and 3 are two-sided models, exposing providers to an increasing proportion of upside and downside risks. Six years since the first ACOs were established, more than 90% of ACOs remain on track 1. This suggests that to date, providers have a limited appetite for risk. It also offers some insight into the level of confidence that US ACOs have in their ability to moderate demand growth.

Robust risk-reward sharing arrangement are complex and are likely to increase transaction costs above those associated with fee-for-service arrangements. The complexity arises as the commissioner or system designer attempts to ensure that the incentives accurately reflect the policy intention, and do not instead reward cost shunting, quality reductions or chance variations in costs.

It is possible to extend risk-reward sharing to multiple partners within an Integrated Care System and to organisations outside the scope of an ICS. But these extensions add further complexity.

The notion of a 'counterfactual' is central to risk-reward sharing. In this context a counterfactual is the price of healthcare activity that might be expected under normal circumstances. It is the benchmark against which priced activity levels are assessed at year end. If priced activity falls below this level, then the provider may be entitled to a reward payment. If it exceeds this level, then a penalty may be applied. There are many approaches to calculating and agreeing

counterfactuals, but none are simple. These calculations determine the allocation of significant sums of money.

If the NHS is to make best use of risk and reward sharing, then it must be aware of the complexities and hazards inherent in these arrangements as well as the potential benefits.

Contents

1 Introduction

Deciphering the jargon and decoding the rhetoric

This report outlines approaches to risk and reward sharing as they might apply to Integrated Care Systems in the NHS. It is designed for those involved in establishing NHS Integrated Care Systems and we anticipate it will be particularly relevant to NHS Directors of Finance, Directors of Strategy and their teams.

NHS England's ambitions to establish Integrated Care Systems in England are clear. In June 2017 the first wave of Integrated Care Systems were announced and by May 2018 coverage had been extended to fourteen areas serving 12 million people. But details about the operating model and financial arrangements for these systems, including those relating to risk and reward sharing, remain sketchy.

In contrast, the Centers for Medicare and Medicaid Services, a US government agency, has established and documented a comprehensive approach to risk and reward sharing for US Accountable Care Organisations. Although rich in detail, these documents are difficult to navigate without knowledge of US health insurance systems and legislation. We have reviewed these documents and many others to identify the central themes and critical issues relating to risk and reward sharing and placed these firmly in the NHS context.

Risk and reward sharing is a simple and attractive concept: offering a commissioner the opportunity to co-opt and incentivise a provider to moderate growth in healthcare demand by sharing in the savings or cost over-runs. But in practice, effective risk-reward sharing arrangements are complex and are likely to increase transaction costs. If the NHS is to make best use of risk and reward sharing, then it must be aware of the complexities and hazards inherent in these arrangements as well as the potential benefits.

Outcomes for a defined population within a fixed budget

In the US, where the term originated, Accountable Care Organisations (ACO) take a particular form and function defined in legislation, and in contracts. Elsewhere, including in the UK, the term takes on a looser meaning. In its recent review the King's Fund define the three key features of accountable care organisations and systems.[1]

[1] The King's Fund, Accountable Care Organisations (ACOs) explained, 14 June 2017, https://www.kingsfund.org.uk/publications/accountable-care-organisations-explained?gclid=EAIaIQobChMI5YalicCH2AIVbLXtCh0JqgrTEAAYASAAEgJOKvD_BwE .

First, they involve a provider or, more usually, an alliance of providers that collaborate to meet the needs of a defined population. Second, these providers take responsibility for a budget allocated by a commissioner or alliance of commissioners to deliver a range of services to that population. And third, ACOs work under a contract that specifies the outcomes and other objectives they are required to achieve within the given budget, often extending over a number of years.

The term Integrated Care System (ICS)[2] is commonly used to describe a collection of organisations which are similar to ACOs in terms of principles, but without (or prior to) formally merging or establishing a contractual joint enterprise.

When describing how ACOs or ICSs might achieve their objective, the following themes emerge:

- shifting care from acute hospital settings to community settings
- proactive management of patients with long term conditions, with a strong emphasis on self-management and shared decision-making
- predicting and pre-empting situations that might lead to a patient being admitted to hospital in an emergency
- sharing data to link-up care delivered by different professional groups
- accepting or sharing the risks associated with increasing demand for healthcare and receiving or sharing rewards should these demands be successfully offset

This report is concerned with the last of these themes.

Sharing risks and rewards to control cost growth

In the US, ACOs were designed to address the problem of rising healthcare costs. They are defined in contrast to, and are seen as a reaction against, the prevailing method of commissioning healthcare where providers are paid for each unit of service delivered and therefore have incentives to increase supply.

Commissioners might ultimately prefer to pass all the financial risk associated with cost growth to the provider by fixing the budget that providers receive. However, most providers would be

[2] Whilst NHS England initially used the term 'Accountable Care System', the preferred term is now 'Integrated Care System'

reluctant to accept a contract with a fixed budget, at least not one that commissioners could afford. Schemes in which savings and losses are shared between the commissioner and provider can therefore offer a compromise between two more extreme options where the risk is borne in full by either the commissioner or provider. Under these arrangements, providers continue to be paid on a fee-for-service basis, but savings or cost over-runs against some agreed target are shared between the parties. To some, risk and reward sharing is seen as a transition to capitated contracts, allowing providers time to integrate and develop innovative ways of controlling cost growth.[3] To others, risk and reward sharing schemes have virtues in their own right, aligning commissioner and provider incentives and encouraging a dialogue about the best balance between service costs and the scope, accessibility and quality of the service.

Risk and reward sharing is underpinned by a theory of change that expects a provider to adjust its behaviour in response to financial incentives. This is a well-trodden path, but the evidence base for incentives in healthcare contracts is somewhat mixed and the efficacy of financial incentives is highly dependent on context (see chapter 3).

Risk-reward sharing approaches can be described in relation to the degree of balance and symmetry. Balanced risk sharing implies that the risk and rewards are shared equally between the parties. Risk-reward sharing is symmetric if the same rules apply to sharing risk and to rewards. In the US, the term one-sided risk sharing is used to describe asymmetric arrangements where an ACO is able to access a share of the savings, but is not required to contribute towards any cost over-runs. In contrast, for two-sided risk sharing, ACOs are exposed to both upside and downside risks.

Figure 1.1: Who holds the risk (or reward)?

	Commissioner	Provider (ACO/ICS)
Fee-for-Service (e.g. national Tariff)	●	
Risk Sharing (One-sided)	●	● (reward only)
Risk Sharing (Two-sided)	●	●
Capitated Budget		●

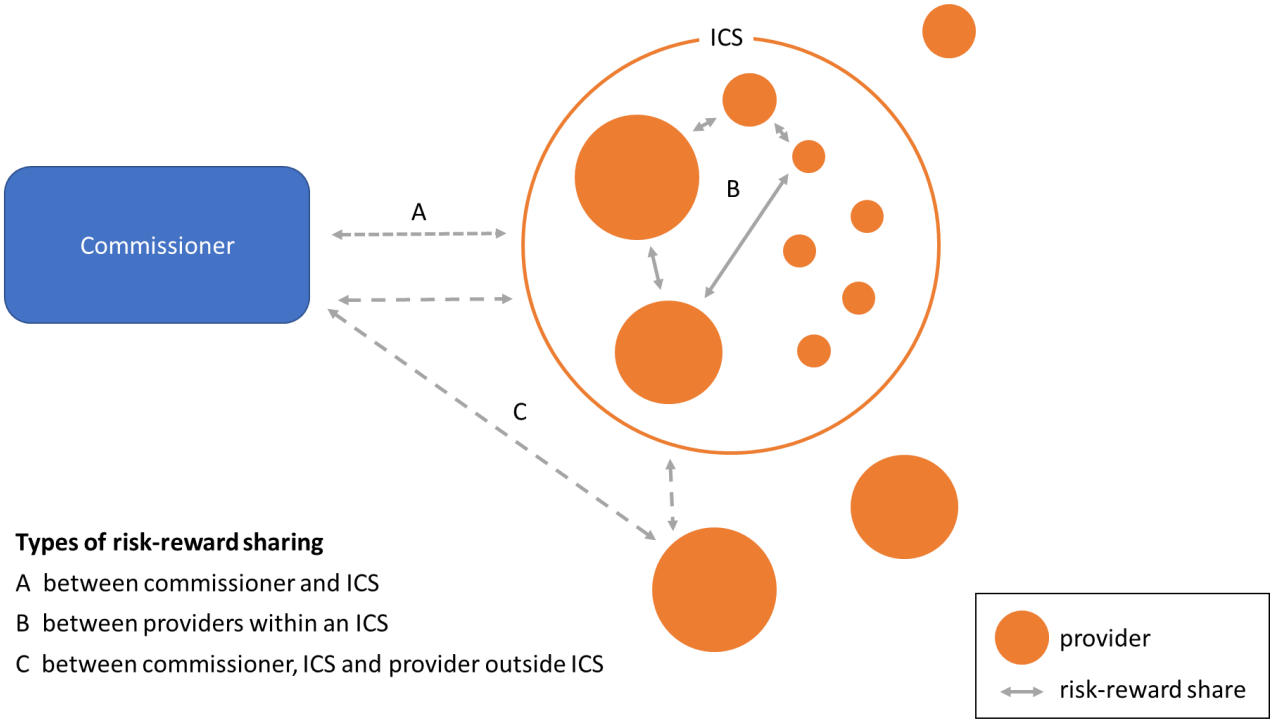
[3] We use the term capitated contract to describe a payment regime where a provider receives an annual payment for each patient. Whilst the contract value may vary as the number and characteristics of patients changes, it is fixed in each year and does not vary according to the quantity of healthcare delivered.

Risk-reward sharing can be seen as building up from a fee-for-service contract or down from a capitated contract. When the base contract is a fee-for-service arrangement then the commissioner would make an additional payment to the provider when savings are delivered, and the provider would repay some of the contract income if costs over-run. When a capitated contract is in place then the commissioner would make an additional payment to the provider if costs over-run and the provider would repay some of the contract income if savings are made.

Who gets a share?

Whilst most risk and reward sharing schemes focus on the transactions between commissioner and provider (ACO/ICS), similar models can be used to share risks and rewards within a supply-chain. We identify three forms of risk-reward sharing which differ in terms of the parties involved.

Figure 1.2: Parties in possible risk-reward sharing schemes



Risk-reward sharing in US ACOs

The first ACOs were established in the US in 2012. The Centers for Medicare and Medicaid Services (CMS), review applications for organisations to become ACOs on an annual basis. By January 2018, 561 organisations were accredited as ACOs, serving 10.5 million patients (referred to as beneficiaries).[4] On average, ACOs serve 19 thousand beneficiaries making US ACOs substantially smaller than the ICSs envisaged by NHS England.[5] When applying to become an ACO or when reapplying at the end of a contract period, ACOs are required to select from a range contractual models. Risk-reward sharing arrangements, known as shared savings and losses schemes, are a key feature of all models. For each ACO, CMS calculate an expected annual expenditure based on the characteristics of their beneficiaries, the expenditure of comparable ACOs and expenditure trends. Potential savings or losses are calculated by comparing actual and expected levels of priced activity. Three risk sharing tracks are available to ACOs.[6]

Track 1 is a one-sided risk sharing scheme: it does not expose the ACO to any risk. If its priced activity exceeds expected levels then the commissioner pays as before. But if savings are made, then ACOs may receive up to 50%. Track 1 is time limited (ACOs can apply to be on Track 1 for a maximum of two, three-year contract terms) after which ACOs must move onto Track 2 or 3. This track is seen as a safe entry point, allowing ACOs the necessary time to establish themselves and test their ability to control cost growth.

Track 2 is two-sided risk sharing scheme. It allows ACOs to receive up to 60% of any savings, while requiring the ACO to pay between 40% and 60% of any losses associated with cost over-runs.

Track 3 is also a two-sided scheme, but allows ACOs to receive up to 75% of any savings, but also requires the ACO to pay between 40% and 75% of any losses.

The actual saving that an ACO is entitled to or the loss an ACO is required to pay, depends on a number of factors including the ACO's performance against a set of quality measures, the number of ACO beneficiaries and the ACO's actual and expected levels of priced activity. ACOs can access the maximum level of savings or minimum level of losses only if they perform well against the quality standards. Larger savings or losses are required before smaller ACOs (with fewer

[4] Centers for Medicare & Medicaid Services , Medicare Shared Savings Program – Fast Facts, https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/SSP-2018-Fast-Facts.pdf .

[5] In 2016 the average number of beneficiaries per ACO was 18,000, range 2,000-140,000, interquartile range 7,000-21,000

[6] Centers for Medicare & Medicaid Services, Shared Savings and Losses and Assignment Methodology – Specifications Version #5, April 2017, https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/Shared-Savings-Losses-Assignment-Spec-V5.pdf .

beneficiaries) become eligible to receive and make payment, because smaller ACOs are subject to higher levels of relative cost volatility. For track 1 ACOs, the savings are capped at 10% of total expenditure, whereas for tracks 2 and 3, savings and losses are capped at 15% and 20% respectively.

Tracks 1, 2 and 3 give ACOs access to an increasing share of any savings. For track 1, all of the financial risk sits with the commissioner, the CMS, but an increasing level of risk is passed to ACOs in tracks 2 and 3.

In 2016, 432 organisations had signed up to the ACO Shared Savings Program. Almost all (95%) of these had elected for track 1 and many (31%) were operating under a second contract term. Under the rules of the Shared Savings Program, these second-term, track 1 ACOs would be obliged to move to tracks 2 or 3 in the third contract term and accept a share of the risk. At this point, CMS introduced **Track 1+** which offered ACOs a lower risk two-sided sharing model.

By January 2018, 87 (20%) of ACOs operating in 2016 had left the program and a further 216 organisations had joined. At this stage 460 (82%) of the 561 ACOs were operating on track 1, 55 (9%) on track 1+, 8 (1%) on track 2 and 38 (7%) on track 3. This implies a limited appetite for risk amongst US ACOs at this stage. It may also provide some insight into the level of confidence that ACOs have in their ability to moderate cost growth.

Early adopters of the ACO model, known as Pioneers, were permitted to move to a full capitated budget approach in the latter part of their contract. This state represents the full transfer of risks from the commissioner to the ACO. Whilst initially several of the pioneers indicated a preference for this endpoint, when the opportunity arose most elected to continue with a risk-reward sharing arrangement.[7]

Risk-reward sharing in English ICSs

In June 2017, NHS England announced that 10 areas serving approximately 9 million patients would be the first wave of Integrated Care Systems.[8] In April 2018, NHS England set out its intention to formally establish these 10 areas as ICSs. One month later four new ICSs were announced extending the population coverage to 12 million people. NHS England publications

[7] L&M Policy Research, Evaluation of the CMMI Accountable Care Organization Initiatives, December 2016, https://innovation.cms.gov/Files/reports/pioneeraco-finalevalrpt.pdf .

[8] NHS England, NHS moves to end "fractured" care system, 15 June 2017, https://www.england.nhs.uk/2017/06/nhs-moves-to-end-fractured-care-system/ .

indicate some flexibility around the services that could be delivered by an ICS. However, the following services appear to fall within the potential scope: primary care, prescribing, community services, mental health services, acute hospital care, specialised services/tertiary care, and adult social care.

Whilst healthcare organisations in England and the US operate in radically different political, financial and cultural contexts, the core rationale that underpins ACO and ICS policies is similar: to moderate healthcare costs through service coordination and integration.

Some indications of the financial arrangements for ICSs were published in August 2017.[9] It appears that all three forms of risk-reward sharing set out in figure 1.2, are being considered as part of 10-year ICS contracts. The term 'gain/loss sharing' is used to describe these arrangements.

NHS England's description of an ICS allows for a range of models. In particular, acute hospital services may fall inside or outside the scope of an ICS. (This is at odds with the US system where ACOs must include costs of hospital services). This is of particular relevance to the issue of risk and reward sharing since hospital services represent the largest source of financial risk to commissioners in England. This risk flows from the fee-for-service payment arrangements for hospital services, known as the National Tariff (previously Payment by Results), that operate in England.

Much like the US model, it is clear that the ultimate intention of NHS England is to pass all financial risk to ICSs by holding them to a budget which is adjusted annually but fixed in-year. But where an ICS incorporates acute hospital services, NHS England acknowledges that it may be necessary to pay for these services under National Tariff arrangements until the ICS matures, and use gain/loss sharing arrangements to facilitate the transition.

Where an ICS does not provide acute hospital services, NHS England envisages a series of bilateral agreements or a single multilateral agreement, which allow savings and cost over-runs to be shared between the commissioner, the ICS and the acute hospital provider.[10] In addition to incentivising a reduction in acute hospital cost, these arrangements would also seek to minimise the risk of an ICS shunting costs from its fixed budget to an acute hospitals fee-for-service contract.

[9] NHS England, Finance and Payment Approach for ACOs, 7 August 2017, https://www.england.nhs.uk/wp-content/uploads/2016/12/1693_DraftMCP-7_A.pdf .

[10] Integrated care policies often set out a preference to eliminate the commissioner-provider split. We take this to reflect a preference to transfer some of the planning and system design functions to an integrated provider. We anticipate some continued if diminished (or more strategic) commissioner function and use the term commissioner in this paper to describe the agency with whom an ICS holds a contract.

Finally, NHS England suggests that an ICS may wish to distribute any savings it accrues amongst its member organisations and in particular to GP practices.

At this stage, little information is available about the mechanisms, rules and calculations that might underpin these three forms of risk-reward sharing.

2 Basic Illustrated Examples

This chapter will illustrate how fee-for-service, block or capitated contracts and risk-reward sharing arrangements impact on commissioner and provider costs and income.

Let us imagine that a highly simplified scenario in which a healthcare commissioner and provider are considering how to contract for activity in the year ahead. Healthcare contracts commonly include several hundred service lines, currencies and tariffs, but for this illustration we will imagine that the commissioner and provider are negotiating a contract for a single type of activity. We also assume that the commissioner and provider are willing to share information about drivers of activity and costs.[11]

In the current year, the provider is expected to deliver 10,000 units of this activity. The commissioner and provider have agreed that increases in population size, age profile and health status and changes in certain clinical guidelines will serve to increase activity of this type by 2% next year, to 10,200 units. At a tariff of £600 per unit this activity would cost £6.12m. The provider estimates that this income would cover the costs of delivering the services and it would therefore break even. If activity is higher than expected, then it would cost the provider £180 per additional unit, and if activity is lower than expected then its costs will reduce by £180 per unit. This implies a marginal rate of 30% (180/600).

Whilst the commissioner and provider agree that demographic and non-demographic pressures will increase demand by 2%, they cannot be certain of this forecast. Furthermore, commissioner and provider have concluded that there are opportunities to reduce activity levels if patient care is managed more proactively.

The commissioner and provider are considering 3 alternative methods of compensating the provider for activity it carries out next year.

1 **a fee-for-service contract** where the provider is paid £600 per unit of activity

2 **a capitated** (fee-per-patient) **contract** for £6,120,000

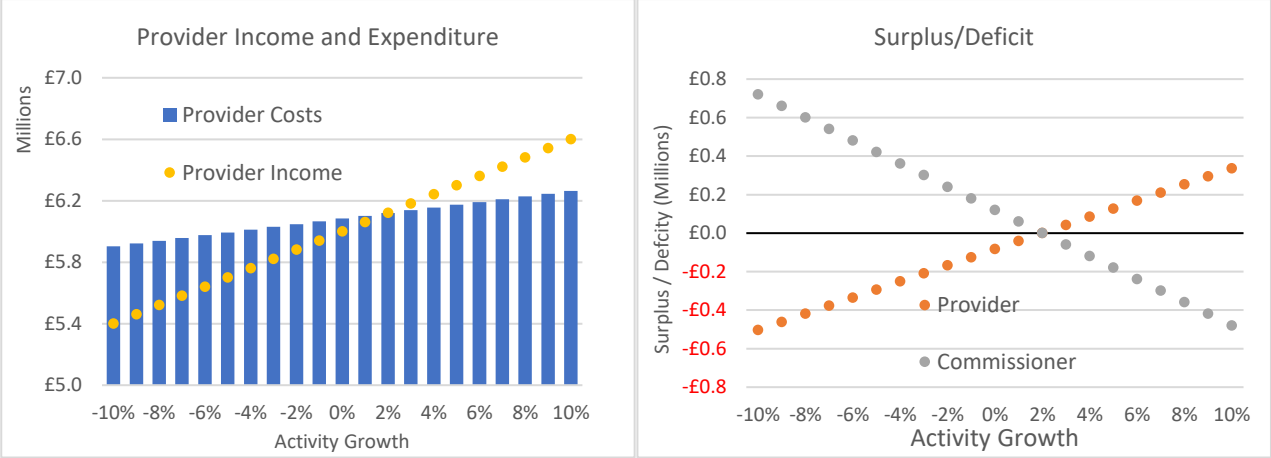
3 **a risk-reward sharing contract** where the provider is paid £600 per unit of activity, but in addition the commissioner will pay the provider £300 per unit of activity below the expected level (10,200 units) and the provider will pay the commissioner £300 per unit of activity above the expected level.

[11] For the purposes of this example we also assume that a provider marginal rate is fixed (i.e. that the cost of providing an additional unit of activity is fixed as the number of additional units increases) and symmetric (i.e. that the reduction in costs when activity is reduced by 1 unit is equal to the increase in costs associated with delivering 1 additional unit).

Fee-for-service models

The charts below show how commissioner costs and provider income under a fee-for-service model will vary under various activity growth assumptions. The chart on the left shows the provider costs and income vary with activity growth. The chart on the right shows the net position for provider (income minus costs) and commissioner (planned expenditure minus actual expenditure).

Figure 2.1: Fee-for-Service Contracts - Costs, Income, and Net Costs



Under a fee-for-service model, the commissioner covers the cost of additional activity at the average (unit) cost. The provider gains income at the same rate but must also cover the cost of delivering this additional activity at its marginal rate. Given that we have assumed that the provider's marginal rate is 30%, the provider's net income (income minus costs) increases by 70% of the average cost for each additional unit of activity delivered. If activity is lower than expected, the commissioner's costs fall by the unit cost and the provider's net income falls by 70% of the unit cost for each, for each unit of activity below the expected level.

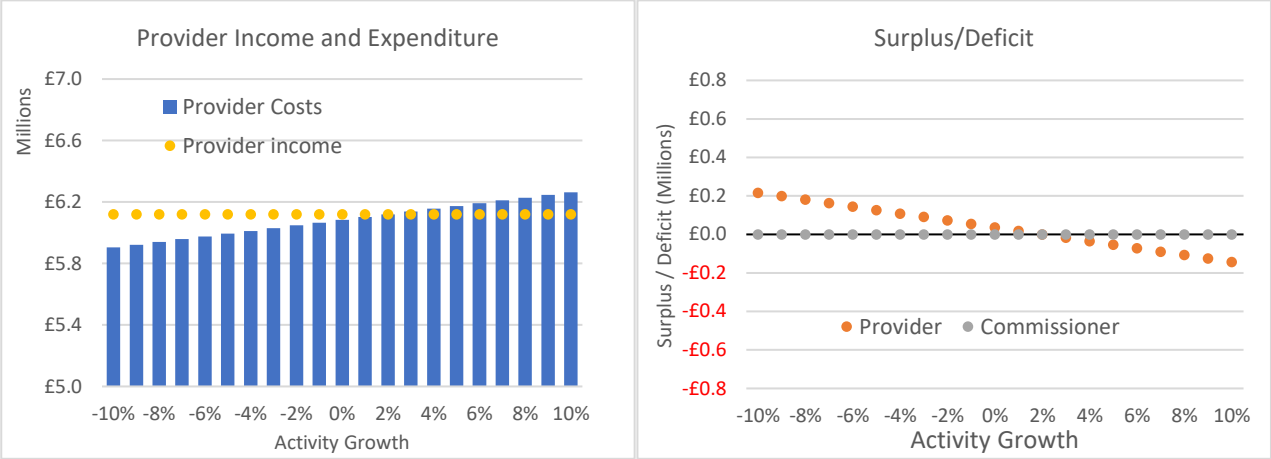
In this model, the provider has a financial incentive to increase activity whereas the commissioner has an incentive to reduce activity, below the expected level.

Capitated contracts

With a capitated contract, the provider's income and commissioner's costs are not dependent on the level of activity. Whilst the commissioner's costs are fixed, the provider must cover the cost of increased activity at its marginal rate. But the provider can reduce its costs at the same rate if activity falls below the expected level.

In this model, the provider has an incentive to reduce activity.

Figure 2.2: Capitated Contract - Costs, Income and Net Costs

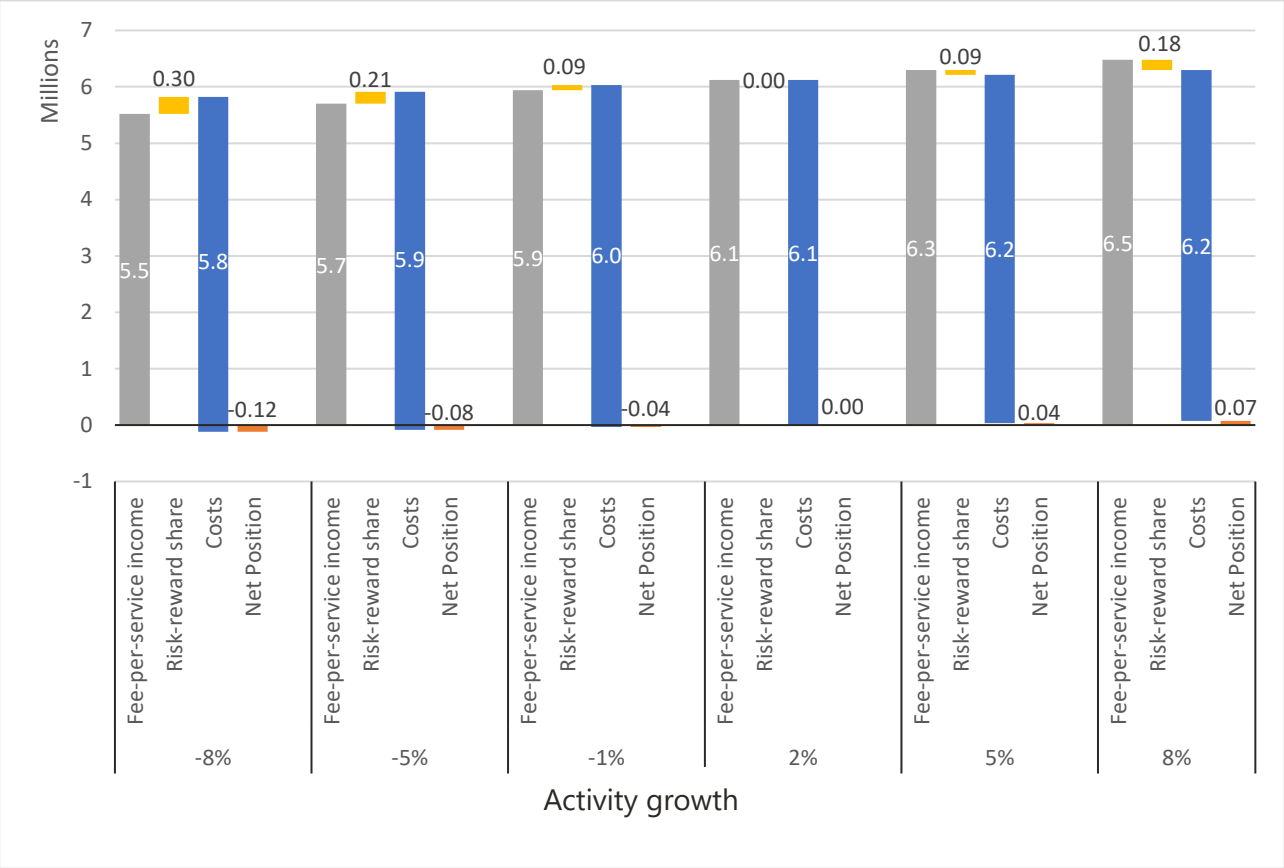


Risk-reward sharing

In a risk-reward sharing model the provider is initially paid on a standard fee-for-service basis, but a year-end adjustment is applied to reward the provider if activity falls below the expected level (growth of 2% over previous year) or penalise the provider if activity exceeds expected levels. In a balanced risk-reward model, the provider's reward (penalty) is set at 50% of the unit price.

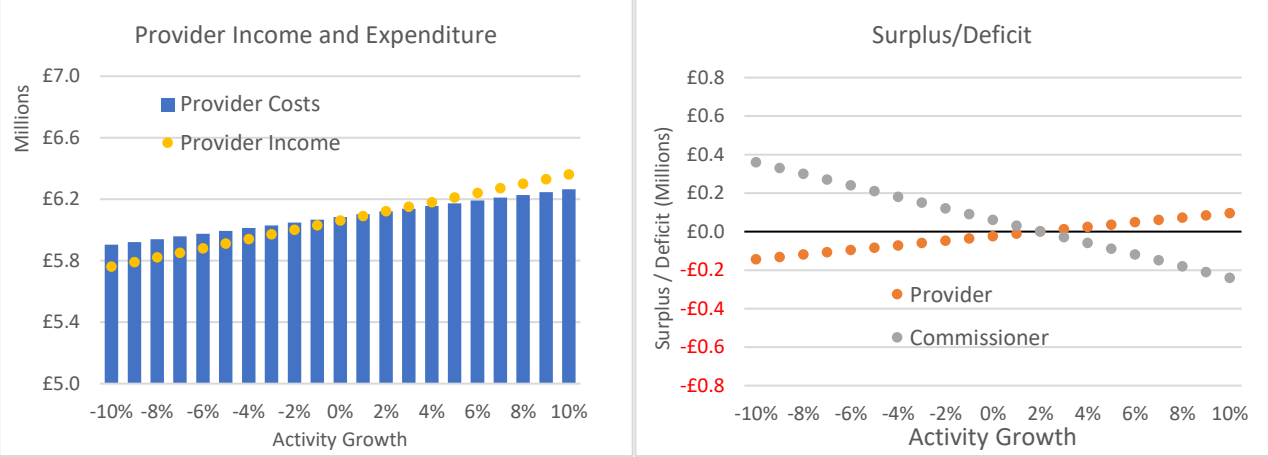
Figure 2.3 shows how these rewards or penalties are applied under different activity growth scenarios and the impact of these rewards on the providers net position.

Figure 2.3: Balanced Risk-Reward Sharing - Income and Costs



When the commissioner and provider agree to share the risks and rewards of activity changes relative to the expected level on a 50/50 basis, then the provider's net income increases by 20% (100% minus the sharing rate, 50%, minus the marginal rate, 30%) of the unit cost for each additional unit of activity and falls at the same rate for activity reductions. The commissioner's costs increase by 50% of the unit cost (the sharing rate) for each additional unit of activity.

Figure 2.4: 50% Risk-Reward Sharing - Costs, Income and Net Costs



In this model, the provider has a much smaller incentive to increase activity, whilst the commissioner has an incentive to reduce activity below the expected level.

Sharing rates and marginal rates

Whether the provider is incentivised to increase or decrease activity and the scale of these incentives is dictated by two factors, the sharing rate (the proportion of the risk-reward that the provider is required to cover) and the provider's marginal rate (the cost of delivering one additional unit of activity as a proportion of the average cost). The commissioner's incentives are dictated only by the sharing rate. The table below shows how incentives for commissioners and providers change as a function of the sharing rate and the marginal rate. Note the provider has no incentive to increase or decrease activity when the marginal rate plus the sharing rate equal 100%.

Commissioners wishing to remove provider incentives to increase supply, must set the sharing rate plus marginal rate above 100%.

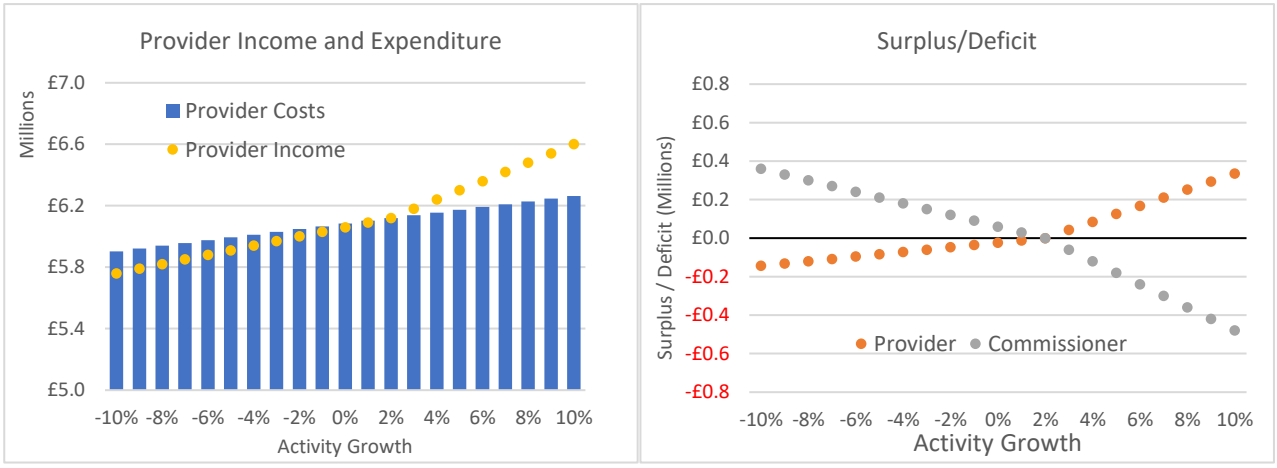
Figure 2.5: Commissioner and Provider Net Costs Under Alternative Sharing Rates and Marginal Rates

	Sharing rate		Net cost of additional activity per unit of activity as a proportion of the unit cost							
			Commissioner	Provider						
				Marginal rate						
	Provider	Commissioner		20%	25%	30%	35%	40%	45%	50%
Fee-for-service	0%	100%	100%	-80%	-75%	-70%	-65%	-60%	-55%	-50%
Risk-reward	10%	90%	90%	-70%	-65%	-60%	-55%	-50%	-45%	-40%
	20%	80%	80%	-60%	-55%	-50%	-45%	-40%	-35%	-30%
	30%	70%	70%	-50%	-45%	-40%	-35%	-30%	-25%	-20%
	40%	60%	60%	-40%	-35%	-30%	-25%	-20%	-15%	-10%
	50%	50%	50%	-30%	-25%	-20%	-15%	-10%	-5%	0%
	60%	40%	40%	-20%	-15%	-10%	-5%	0%	5%	10%
	70%	30%	30%	-10%	-5%	0%	5%	10%	15%	20%
	80%	20%	20%	0%	5%	10%	15%	20%	25%	30%
	90%	10%	10%	10%	15%	20%	25%	30%	35%	40%
Capitated contract	100%	0%	0%	20%	25%	30%	35%	40%	45%	50%

One-sided risk-reward sharing

Finally, the charts below show how asymmetric risk-reward sharing models impact on costs and income. In this example, the provider receives half of any savings associated with reduced activity but is not required to contribute should activity increase beyond the expected level.

Figure 2.6: 1-sided Balanced Risk-Reward Sharing - Costs, Income and Net Costs



In this model, the incentives to increase supply remain, but the disincentives to reduce activity have been moderated.

In these simple examples, an ACO's income is expressed as a formula (see appendix B1). Extensions to these simple examples and associated formulae are set out in chapter 5 and appendix B2.

3 Making Risk-Reward Sharing Work

In this chapter we describe some the characteristics of risk-reward sharing schemes that are likely to make them successful and some of the practical limitations of these schemes.

What should risk-reward sharing incentivise and discourage?

If risk-reward sharing schemes are a means of aligning incentives, what behaviours and activities might a good risk and reward sharing scheme want to encourage and discourage?

A good risk and reward scheme should encourage ICSs to **reduce healthcare costs the right way**. A commissioner would not want an ICS to reduce costs by rationing or restricting access to cost effective interventions (i.e. by creating or growing waiting lists or times), by lowering quality standards, or by deselecting patients whose needs are great or whose patterns of care are not considered amenable to care management strategies. The underpinning ethos of accountable care is that providers reduce costs by collaborating and improving allocative efficiency, by investing in prevention and by pre-empting escalations in patients' needs. So, a good risk and reward scheme needs to encourage some approaches to cost reduction and discourage others. It should operate over a sufficient duration to allow providers to reap the rewards associated with investment in prevention.

Second, a risk and reward sharing scheme should encourage an ICS to **plan and act ambitiously, but not recklessly**. Small incentives may not be sufficient to motivate an ICS to innovate and seek improvements. If the incentives are too large, then providers might act irresponsibly, tempted by the substantial potential rewards. This argument also applies to one-sided risk sharing. An ICS in a one-sided risk-sharing scheme stands to gain from any savings but is not required to pay if any losses are made. In these circumstances what is to stop an ICS gambling with the commissioner's money on high-risk projects? If the gamble works then the ICS wins a share, and if it fails, then the commissioner picks up the cost (see box 3.1 below).

Third, a good risk reward sharing scheme should encourage an ICS to **plan thoroughly whilst acknowledging uncertainties**. If incentivised appropriately, an ICS should recognise the benefits of accurately planning and forecasting, acknowledging that it pays the consequence if its plans are poor. The consequences might take the form of financial losses or missed opportunities to secure shared savings. Some uncertainties, however, cannot be resolved by analysis and the commissioner would want an ICS to be cognisant of how these uncertainties might affect their service. The impact of Brexit on UK healthcare provision is a good example of such a critical uncertainty. Exactly how Brexit might affect the UKs GDP, the share of tax-revenue that might be allocated to the NHS, the NHS's access to overseas workers, the cost of imported healthcare goods, etc. cannot be known with any certainty at this stage. An ICS needs to plan to accommodate a range of scenarios, and we should expect an ICS to price this uncertainty in its contracts.

Box 3.1: Do Payment Incentives Influence Service Utilisation Levels?

In 2007, the Health Foundation published a review of the evidence on the impact of payment incentives on healthcare services.[12] Whilst predominantly focused on the impact of incentives on service quality, the paper also explored the impact of incentives on service utilisation. The authors concluded that:

> *"… provider behaviour in the patient–provider decision-making process can be influenced by financial incentives."*
>
> *"… very few studies [describe] how incentives in payment arrangements between payers and provider organisations [are] actually transmitted to practitioners."*
>
> *"Beyond supporting the idea that incentives can be used to influence provider behaviour, the literature provides little useful direction for decision-makers regarding the design of incentive systems that would reward quality of care."*

For incentives to work, professionals taking decisions about the allocation of healthcare resources (most commonly clinicians) must not only be aware of the incentive arrangements, they must also have a clear understanding of how their decisions can influence the outcome at which the incentive is targeted. In particular, clinicians must have sound understanding of the short and long-term cost consequences of alternative treatment options.

Fourth, an ICS should **not benefit from cost shunting**. Cost-shunting refers to a range of strategies that organisations employ to minimise costs at the expense of another organisation. This may involve reclassifying activity so that it falls outside the scope of a contract or failing to act or invest, thereby forcing another organisation to do so. Whilst this represents a good financial outcome for the first organisation, it represents poor value for money for the system as a whole. Cost shunting could occur between organisations within an ICS or between providers inside and

[12] The Health Foundation, Financial incentives, healthcare providers and quality improvements, A Review of the evidence, 2007, www.health.org.uk/sites/health/files/FinancialIncentivesHealthcarePRovidersAndQualityImprovements.pdf .

outside an ICS. This issue is particularly relevant when some or all acute hospital services fall outside the scope of the ICS.

Fifth, an ICS should always benefit from **accurately recording its activity**. Resources are required to record activity accurately. If appropriate incentives do not exist, then providers may record activity poorly, leaving the commissioner with limited data and intelligence to assess performance and develop future plans. Conversely, incentives should not be so great as to encourage provider gaming, distorting recording and reporting processes such that data cannot be trusted.

Finally, a good risk-reward sharing system should encourage providers to **collaborate and share information** appropriately. It should never be in an ICS provider's interest to withhold information from its partners if this information could be used to deliver better outcomes for the population it serves or lower costs for the system as a whole.

Practical considerations

A review of the technical documentation for the CMS ACO Sharing Savings Programme, quickly demonstrates how **complex** these schemes can become. An evaluation of the US ACO Pioneers programme, concluded that many of these highly experienced healthcare providers did not foresee the implications of critical aspects of shared-savings programme and that this had a substantial impact on their financial sustainability.[13] It is unlikely that this complexity had been designed into the scheme unnecessarily, and so each layer of detail and each step in the process must have been developed to make the scheme work or to close a perceived loophole. Many ACOs might wish the scheme were simpler, but simple schemes are unlikely to exhibit all of the characteristics described previously. Given the complexity of healthcare service and associated legislation, a highly simplified scheme is likely to be ineffective, imprudent and potentially counter-productive in any health system including the NHS.

A successful risk-reward sharing scheme therefore must be **defined in sufficient detail** to handle the complexity inherent in the English healthcare system. Most notably, the scheme must reflect the scope of the ICS (patients and service inclusions and exclusions); it must recognise that the eligible population is not fixed (deaths, births, migration) and that medical technology, clinical standards and patient expectations are constantly evolving. Furthermore, given Government

[13] L&M Policy Research, Evaluation of CMMI Accountable Care Organization Initiatives, 2 December 2016, https://innovation.cms.gov/Files/reports/pioneeraco-finalevalrpt.pdf .

commitments to patient choice, it must also confirm how activity delivered by private or out-of-area providers should be treated in any financial reconciliation process.

Whilst a commissioner may not be able create a highly simple and elegant scheme, it can make the scheme details open to all and the processes **transparent**. The carries the dual benefits of allowing providers and independent organisations the opportunity to review and comment on the schemes methods, but it also allows potential ICS's to weigh up their options and plan appropriately. Transparency can also promote trust between the commissioner and ICS, limiting the opportunities for commissioners and ICSs to take divergent views on the rules that govern the scheme when information about the rules that operate are inadequate or ambiguous.

The timely provision of **management information** can help ICSs understand their exposure to risk in-year, affording them time to adjust their plans and activities. Management information may even moderate a provider's inclination to price uncertainty. US ACOs receive a suite of management information from their commissioner which includes detail on their assigned beneficiaries, their expected expenditure (against which shared savings and losses will be calculated) and their expenditure to date.

Special considerations for the NHS

Two-sided risk sharing and capitated budget schemes which push some or all of the financial risk to an ICS, carry the inherent hazard that the provider will be unable to cover its costs. Whilst the failure of a private sector healthcare provider might be inconvenient for a commissioner, it may be **politically untenable for an NHS provider to suffer substantial losses**, calling into question its financial viability, particularly if the provider delivers high quality services to its patients. Indeed the current NHS financial regime deems in-year provider deficits as unacceptable. Insulating NHS ICSs from financial risks, fundamentally blunts risk-reward sharing as a tool to manage cost growth, limiting the benefits that any such schemes might produce.

The US ACO Shared-Savings programme, attempts to avoid the potential for provider bankruptcy by requiring the ACOs demonstrate that they have sufficient means to pay for any shared losses that might be incurred. In an NHS setting, the **means to repay losses** must presumably be made up of reserves that the provider has set aside. These public funds would not be available to deliver health services to the public, but instead to provide the necessary assurance that losses could be covered. Whilst it is not uncommon for public sector bodies to hold reserves of public funds, the

scale of the funds required for risk and reward sharing may be politically unacceptable. Moreover the notion of holding reserves to offset deficits is somewhat at odds with those of control totals.[14]

In the US, ACOs receiving shared savings may decide to reinvest these savings or it may choose to assign savings to profits. For NHS ICSs, the options are more straightforward, but public sector finance rules constrain **the opportunity to use savings** from one year to offset operating costs in the next. If NHS ICSs are unable to carry forward savings with freedom, then the intended incentives will not function.

Where an ICS's services are delivered in full or in part by private sector organisations, we must acknowledge that **large financial rewards may flow out of the NHS**. It is not clear whether this would be seen as acceptable to the public, politicians or policy makers.

If activity levels deviate substantially from commissioner's planning assumptions, the financial consequences set out in long-term (10 year) ICS contracts may **limit the funding available for services that fall outside of the scope of an ICS.**

The limits and hazards of risk-reward sharing schemes

Whilst risk-reward sharing schemes have many potential benefits, there are limits to what the schemes can achieve, and to date, limited evidence of benefits.

If the core objective of risk-reward sharing schemes is to provide a safe intermediary stage on the journey to full risk transfer to providers, then the US schemes provide very **limited evidence of success** so far. Many providers have signed up to the one-sided ACO risk sharing model, which carries only upside risks, but very few have indicated their willingness to expose themselves to downside risks.

Whilst capitated budgets hold the promise of reducing transaction costs, we should expect the intermediary stage, risk-reward sharing, to produce an **increase in transaction costs** as commissioners and providers maintain the pre-existing fee-for-service arrangements in addition to administering new processes associated with risk-reward sharing schemes. Moreover, commissioners should expect providers to price the uncertainty arising from the novelty of the contractual form. At some level, these cost increases must be offset against the potential cost benefits of risk-reward sharing.

[14] The term 'control total' is used by NHS England to denote the spending limit that is negotiated and agreed between NHS England, NHS Improvement and individual NHS provider trusts and CCGs.

The US ACO Shared-Savings Programme provides assurance that ACOs are not making savings by cutting services and reducing quality by measuring a set of process, outcome and experience indicators. These approaches can provide **only limited assurance that providers will not game** the system and that quality will not suffer since only a subset of all possible outcomes, processes and experiences can be measured. Providers determined to maximise profits at any cost would seek to maintain and improve performance against the contract performance measures, whilst eroding service levels and quality in other areas.

Even the best designed risk-reward sharing schemes are **unlikely to anticipate all eventualities**. Would a commissioner stand-firm and impose shared losses on providers whose costs over-ran in response to an unexpected epidemic or an unforeseen political priority? Any post-hoc adjustment to the system rules, will invite further challenges from providers seeking to minimise losses.

4 Establishing a Counterfactual

What is a counterfactual?

In order to determine how effective an ICS has been in controlling cost growth, we must have some estimate of the level of priced activity that is expected under normal circumstances. This imagined outcome, describing what might otherwise have occurred, is often referred to as a counterfactual. If for example, the agreed counterfactual priced activity for an ICS was £52m and the year-end position showed priced activity of £50m then we might conclude that the ICS saved £2m. Whereas if the counterfactual was set at £47m then we would conclude that the ICS had caused an additional £3m of costs. In risk-reward sharing models, the £2m savings or £3m cost over-run would be shared between the ICS and commissioners in some agreed fashion. The counterfactual and the methods used to set it are crucial to the effective operation of risk-reward sharing schemes.

Whilst NHS commissioners and providers often attempt to set a contract value for a year ahead, these values rarely have any formal status. If risk-reward arrangements are introduced, then the methods used to set counterfactual priced activity levels have the potential to determine the allocation of many hundreds of millions of pounds each year.

The counterfactual should represent an unbiased estimate of future priced activity. It should not be deliberately challenging, lenient or be constrained by budgets. It must be fair, authoritative, credible and transparently calculated.

There are many ways in which this counterfactual outcome can be defined, and many methods of calculating counterfactuals. Whilst no single definition or method can claim to be definitive, some approaches appear to have more merit than others given the risk-reward sharing objectives. This chapter will explore who should set counterfactual expenditure levels, when, and how these should be set.

Who should set the counterfactual?

A counterfactual sets the threshold at which an ICS is deemed to have generated savings rather than cost over-runs. An ICS might therefore prefer counterfactual priced activity levels to be set high, to minimise the risk of shared loses whilst maximising the potential and level of shared savings. Commissioners on the other hand, might prefer counterfactual priced activity levels to be set at a low level. We describe three approaches to setting counterfactuals which aim to manage this tension.

The first option is for counterfactuals be set by a **third party** who is seen to be independent of the commissioner and ICS. Indeed, it may be sufficient for the third party to set out the methods and

data sources that should be used to unambiguously determine the counterfactual position for any given ICS. A well-regarded academic institution, health think-tank or a suitably qualified, specialist NHS team would be well placed to perform this function. Whichever option is selected, we should acknowledge that this exercise will require investment.

A second option is to set counterfactual priced activity levels **by negotiation between the commissioner and ICS.** If this approach is adopted then there may be value in establishing a framework that both parties use to set out their understanding of baseline priced activity, future (counterfactual) priced activity and the various assumptions that are made to transit from the baseline to the counterfactual. This would provide some structure for the negotiation, facilitate greater understanding of each other's position and minimise the risk of misunderstandings. There is a risk that this approach will drive up costs, with commissioner and provider investing greater sums to secure the best negotiators and analysts. A hybrid option might see the commissioner and ICS provider jointly appoint an independent facilitator to agree a method and to conduct the technical calculations on behalf of both parties.

Negotiations do not always lead to agreement. We should expect that some commissioners and ICS providers may be unable to agree on a counterfactual priced activity level. In these cases, a third party could be engaged to arbitrate and set the level. NHS England may wish to consider nominating an agency to fulfil this role for all commissioner-ICS negotiations.

Where the number of potential ICSs is large, then a third option is to set the counterfactuals through some **procurement process**. The commissioner would set the counterfactual or stipulate the method that should be used to calculate the counterfactual, and suppliers would engage in the procurement process only if they considered the level to be reasonable. The commissioner would be incentivised to ensure that it set accurately so as not to discourage potential suppliers from responding without exposing themselves to any unnecessary financial risk.

When should counterfactuals be set?

Counterfactuals are commonly used when researchers attempt to determine the impact of an intervention on specific outcomes. In this context, it is best practice for researchers to set out the method for calculating counterfactual outcomes in advance even though the counterfactual can usually only be calculated after the intervention has been delivered.

In the context of risk-reward sharing for accountable care systems, counterfactual priced activity levels calculated after year-end are unlikely to promote thorough and ambitious planning or promote trust between the commissioners and ICS. Instead, counterfactual priced activity levels

should be agreed before or as early as possible within a financial year, to provide the ICS with a clear financial target.

Where longer term contracts are put in place, commissioners are unlikely to want to agree specific counterfactual priced activity levels for all years of the contract, because greater uncertainties exist for long term forecasts. Where this is the case, commissioners should provide as much detail as possible on the methods that will be used to calculate counterfactual priced activity levels and must allow ICS's to withdraw from the ICS contract without penalty beyond the period where counterfactuals are agreed.

How should counterfactuals be set?

It may be tempting for commissioners to set counterfactual ICS priced activity levels with respect to their budgets, but there are many reasons to avoid this approach. A counterfactual should indicate the level of priced activity that might be expected under normal circumstances, rather than the level of priced activity that can be afforded. Commissioner allocations change as a function of the overall NHS settlement, which is set as much by the political context as by health need. A number of commissioners operate under structural deficits and commissioner allocations explicitly acknowledge the gap between actual and fair shares allocations. For all these reasons, commissioner budgets do not represent a fair counterfactual. Furthermore, budgets commonly incorporate ambitions for activity or cost control. These are the areas that the ICS will likely target to reduce their costs below the counterfactual. If these opportunities are included in the counterfactual, then this sets the bar unrealistically and unreasonably high for an ICS.

Counterfactuals should instead demonstrate how current levels of priced activity will change over the coming year(s) without ICS intervention, showing the assumptions and calculations that have been applied. We describe four approaches that can underpin counterfactual calculations; time series forecasts, cross-sectional analysis, actuarial analysis and judgements driven by expert opinion. These approaches are not entirely distinct and in practice they may be deployed in combination.

Time series methods use trends in historical priced activity for the population served by the ICS to forecast future expenditure. Increasing the number of historical data points allows more sophisticated forecasting methods to be used and is likely to increase the forecast accuracy. For planning purposes, there may be value in disaggregating historical data into a set of meaningful service lines and forecasting these separately. Similarly, there might be value in separating activity trends from unit cost trends. Data on factors which are thought to have influenced historical trends (e.g. population size, average earnings) can be incorporated into forecasts, where future values of these predictor variables can be sourced or estimated. Most forecasting techniques will

provide estimates of future priced activity (or activity or unit price) as well as some indication of the probability that future values might deviate from these estimates by 5%, 10% or 25%. Whilst the central forecast should be used to set a counterfactual, these confidence intervals may help in setting thresholds where shared savings or risks are triggered.

Cross-sectional approaches calculate the average priced activity level per head of population for areas similar to the ICS.[15] These average priced activity levels per head are applied to the ICS population to produce a counterfactual priced activity. As with time series models, cross-sectional approaches can be applied to priced activity as a whole, or by service line or separately to activity and unit cost. Cross-sectional approaches can also incorporate factors which are thought to explain differences in priced activity between areas (e.g. population age structure, levels of need). This process, known as casemix adjustment, helps a commissioner to understand what proportion of the variance from the unadjusted benchmark priced activity levels can be explained by differences in need. Cross-sectional approaches would not intrinsically recognise that some drivers of priced activity are structural and cannot be changed immediately. These factors would need to be adjusted for informally.

The third approach uses **actuarial analysis**. Whilst actuarial science has its origins in the health insurance industry, the underlying methods have been used extensively by the NHS for many years. This approach uses mathematical and statistical methods to predict the level of healthcare that an individual might require in the future based on their characteristics. The characteristics commonly include a patient's age, gender, diagnosed conditions and prior healthcare utilisation, but a wide range of other factors can also be used. These prediction models use historical data to assess the influence of these factors over future healthcare use. These predictions are rarely sufficiently accurate at an individual level to usefully guide decisions about a patient's care. But when the predictions are aggregated across large populations, some of the inaccuracies cancel out, leaving more accurate population-level predictions. These summed predictions across an ICS population can serve as a counterfactual.

The final approach relies on judgements and **expert opinion**. In this context, experts might include local commissioners, provider managers and clinicians, national experts/academics in healthcare commissioning trends, demographics, healthcare technologies or public attitudes to health and healthcare. Nominated experts may find it difficult to set counterfactual priced activity levels in total and breaking the question down into a number of smaller problems may make the exercise more tractable. We suggest the problem is broken down as follows:

[15] The benchmark might equally be some other point of the distribution (e.g. lower quartile).

- demographic effects on activity – changes in population size, age profile, and age-specific morbidity levels/health status
- non-demographic effects on activity – changes in clinical guidelines and service standards; the introduction of new technologies and the declining use of out-moded interventions; changes in public expectations
- price/cost effects – changes in labour or estate costs; improvements in service efficiency; changes in drug/device costs

Whilst these four distinct methods clearly differ, and will produce different results, no single approach is unequivocally superior.

A more detailed, technical report on this topic will be published by the Strategy Unit later this year.

Counterfactuals in the US ACO Shared Savings Program?

The US ACO shared saving programmes uses a combination of actuarial analysis and time series methods to estimate counterfactual priced activity levels. The process by which counterfactuals are set is highly detailed, although a substantial part of this process is required to handle the many complex public health insurance plans that exist in the US and deal with the high levels of patient switching between plans and providers. Setting these issues aside, there are three stages to the process for a new ACO joining the shared savings programme

- the average priced activity per head is calculated for the population to be served by the ACO for each the last three years. Some assumptions are required to estimate savings in the most recent (incomplete) year. The impact of very high cost patients is moderated by truncating extreme values in the average priced activity calculations.
- these average per head priced activity levels are casemix-adjusted to take account of the changes in the age and health status of the population to be served by the ACO. This case-mix adjustment uses established actuarial models.[16]
- the casemix-adjusted, per head expenditure levels are combined across the three years to produce a weighted moving average with the greatest weight given to the most recent year.

[16] The models used in the US ACO programme are not dissimilar to the risk stratification methods widely used in the UK.

It is this casemix-adjusted per head priced activity level that acts as the counterfactual and against which future casemix-adjusted priced activity levels are compared. An ACO must reduce expenditure below this level in order to access shared savings.

Whilst comprehensive and detailed, it does not necessarily follow that this approach represents a 'gold standard' against which alternative methods can be assessed.[17]

[17] Indeed, the question of how to assess the effectiveness or appropriateness of alternative methods of setting counterfactuals is by no means trivial.

5 Calculating Rewards and Losses

The difference between actual levels of priced activity[18] delivered by an ICS and the counterfactual level represents the level of saving or cost over-run that must be shared between the ICS and the commissioner. At the simplest level these savings or costs can be split equally between the ICS and the commissioner. However, commissioners may wish to make adjustments to this calculation to moderate some of the risks associated with the programme. We describe several adjustments, many of which feature in the US Shared Savings Program.

Unbalanced and asymmetric sharing

If we consider risk-reward sharing to be an intermediate step between fee-for-service and capitated budget contracts, then a risk-reward scheme that shares savings or losses 50-50 can be thought of as the mid-point between the two models. A commissioner and ICS that wishes to manage the journey from fee-for-service and capitated budget more slowly might opt for a scheme that assigns only 25% of the savings or losses to the ICS in years 1-3, 50% in years 4-6 and 75% in year 7-10. A similar transition is seen in the US where a greater proportion of the savings are available to ACOs as they move from track 1, to 2, and 3.

Asymmetric sharing models assign different proportions of the savings and losses to the ICS. For example an ICS might be eligible to receive 50% of the savings but is only required to pay 25% of any losses. Track 1 of the ACO shared savings programme represents an extreme case of asymmetric sharing, where the ACO can access shared savings but is not required to contribute towards any shared losses. Asymmetric models of this type offer a low risk opportunity for ICSs to enter the market. If counterfactuals are unbiased, (i.e. equally likely to understate and overstate future priced activity) then the commissioner should expect savings and cost over-runs to occur with the same frequency unless the incentives succeed in influencing provider behaviour. Because savings are shared with the provider and cost over-runs are not, savings on average will not be sufficient to offset cost over-runs. A commissioner that operates a number of ICS contracts might therefore expect one-sided risk-reward sharing to increase its costs compared to a symmetric risk-reward approach (or indeed to a fee-for-service model or a fully capitated model).

These approaches may be useful when the state is prepared to accept some cost increases in the short term to encourage provides to migrate onto integrated service contracts in the expectation that the investment will produce returns in the medium or long-term.

[18] Calculated using national tariffs and equivalent systems for non-acute care.

Foreseen in-year effects

Whilst demographic changes (births, deaths and migration) are to some extent predictable, some variation between population projections and actual future population levels and structures will inevitably arise. A risk-reward sharing policy should set out in detail how these differences should be accounted for. In the US shared savings program, counterfactual priced activity levels are adjusted after year end to reflect changes in the numbers and characteristics of beneficiaries.

Similarly, if changes are made to the underlying payment mechanism (e.g. changes to national currencies or tariff), then these changes will need to be accounted for in any final reconciliation; either by adjusting the counterfactual or stripping out the changes from actual priced activity calculations.

Unforeseen in-year effects

The purpose of risk-reward sharing is to co-opt a healthcare provider into the process of controlling cost growth. Commissioners may argue that in accepting a contract with risk-reward sharing an ICS is in effect agreeing to accommodate a proportion of the inherent uncertainties that influence healthcare expenditure across populations. However, certain rare and unpredictable events such as natural disasters and disease outbreaks have the potential to substantially increase healthcare expenditure in-year and an ICS may reasonably argue that counterfactual calculations should be adjusted post-hoc. The commissioner's approach to resolving these issues should be explicit in any contract with an ICS.

Similarly, whilst the commissioner can set out its requirements of an ICS in a contract, it may need to alter these requirements in-year in response to, for example, changes in government policy or the approval of a new high-cost or high-volume drug or medical device. Changes to an ICS's requirements should be reflected in the counterfactual.

The options for in-year adjustments to counterfactuals mirror those for setting counterfactuals (see previous chapter).

Taking account of random cost variation and volatility

Levels of healthcare expenditure vary significantly between patients and can change substantially for a patient from one year to the next. This variation and volatility poses a challenge to organisations that must deliver healthcare to a population within some limited budget. Some of

this variation can be explained by casemix or predicted using actuarial models, and so can be accounted for when setting counterfactual priced activity levels. But some unexplained and unpredictable variation will always remain. When risk-reward sharing arrangements are in place, the financial risk associated with this variation is shared between the commissioner and the ICS. One means of managing this risk is to increase scale, delivering care to a larger number of patients. This variation when summed across groups of patients will tend to be small relative to total priced activity. Indeed, the larger the group, the smaller the residual variation as a proportion of total priced activity. However, these effects only apply when the variation is independent across patients, such that one patient's unpredictable cost is not correlated with another's. An epidemic, may cause large and unpredicted costs increases for many patients. Because this effect is not independent across patients, the variation does not 'cancel out'.[19]

When assessing an ICS's performance at year end, there are therefore three potential explanations for the difference between actual priced activity and the agreed counterfactual. If the counterfactual has been set appropriately and adjusted for in-year effects, then the difference can be explained either in terms of an ICS's success/failure to control activity or in terms of chance variation in patient costs. Risk-reward sharing arrangements are designed to reward (or penalise) ICSs for successfully (unsuccessfully) controlling activity, but a commissioner may not want to reward (or penalise) an ICS when its costs are low (or high) in a given year simply by chance.

The probability that chance will cause a certain level of cost over-run for a given population can be quantified using statistics. The larger the population, the smaller the probability of a large relative cost over-run (or saving) occurring by chance. Commissioners can use this knowledge to minimise the potential of rewarding or penalising an ICS for savings or cost over-runs which occur simply by chance, by setting a buffer either side of the counterfactual such that savings or loses are shared with the ICS only when expenditure falls outside of this buffer. The size of the buffer would be determined by the number of patients served by the ICS.

If an ACO's priced activity falls just below the counterfactual, then CMS contends that this may have arisen through chance alone and as such the ACO is not entitled to a share of the savings. The minimum savings rate (MSR) sets the levels of savings that a provider must achieve in order to access shared savings. The MSR is expressed as a proportion of the counterfactual and is set with reference to the number of ACO beneficiaries. The smaller the number of beneficiates the larger the MSR.

[19] These patient dependent but time independent effects can be managed by increasing the time interval under consideration.

For ACOs on tracks 2 and 3 a minimum losses rate (MLR) is also calculated and applied. These ACOs have some freedoms to determine how minimum savings rates and minimum losses rates are calculated. The selected approach is locked in at the outset of the contract and must apply symmetrically to savings and losses.

It is worth noting that most of the US ACOs are small relative to the size of ICSs that are likely to emerge in England. Commissioners in England may determine that chance variation in costs across the very large populations served by ICSs is likely to be small and does not warrant the introduction of additional programme complexity.

Capping risks and rewards

Whilst we might expect an ICS's priced activity to fall close to the counterfactual, an ICS's outturn can in theory take any positive financial value. Very large deviations from the counterfactual should raise concerns. From a commissioner's perspective, large differences might indicate some problems with the risk-reward sharing process. A significant variation from a counterfactual might signal some deficiency in the process used to set the counterfactual. Alternatively, when an ICS's expenditure is considerably lower than its counterfactual, this might indicate that the ICS has found a way of reducing priced activity without contravening its contractual obligations, which is technically legitimate, but not in keeping with the spirit of the contract. And when priced activity is substantially higher than the counterfactual, this represents a large (if shared) financial problem and might indicate imprudence or incompetence on behalf of the ICS.

Commissioners may elect to cap shared savings or losses, either in absolute terms or as a proportion of the counterfactual expenditure. Capping shared losses also limits an ICSs financial exposure and may make market entry more attractive.

In the US ACO shared savings programme, shared savings are capped at 10% of the counterfactual for ACOs in track 1, at 15% for ACOs on track 2 and at 20% for ACOs on track 3. Shared losses are capped at 10% of the counterfactual for ACOs on track 2 and at 15% on track 3.

Quality pre-requisites and quality adjustments

Commissioners may wish to make access to, or the level of, shared savings and losses dependent on the ICS's performance against a set of quality or outcome measures. This would provide an incentive for ICSs to improve or maintain quality and assure the commissioner that cost savings

have not been achieved at the expense of an erosion of service standards (e.g. by creating or growing waiting lists or times).

The most straightforward way to achieve this is to set out the minimum service standards that must be delivered by the ICS in order to access shared savings if its costs fall below the counterfactual. Alternatively, the level of access to shared savings might be increased on some sliding scale depending on the level of performance against the standards. A similar approach could be used to moderate the level of shared losses that an ACO is required to pay when its expenditure exceeds the counterfactual, when it has achieved high service standards.

In the US ACO shared savings programme, ACOs are monitored against 29 measures: a mix of service standards, patient satisfaction measures and outcome measures.[20] In year 1, ACOs are simply required to accurately report their performance against these standards. This is sufficient to gain access to shared savings and moderate shared losses at the maximum rate. In future years, ACOs are awarded a score which reflects both their level of performance compared to set of national benchmarks and the rate of quality improvement. If an ACO's priced activity falls below the counterfactual, then the quality score determines the proportion of shared savings that it is entitled to: from none where quality is low to the maximum shared savings rate allowed in each track when quality is high. If an ACO's priced activity is higher than the counterfactual, but its quality is also high, then the quality score determines the proportion of shared losses that can be avoided.

Adjusting for prior performance

Historical data is used to establish an ACO's counterfactual priced activity levels. As such, a high performing ACO that has successfully controlled priced activity in one contracting period may argue that its counterfactual the next contracting period is unrealistically low; that it is in effect penalised in the second round because of its performance in the first contracting round. To ensure that high-quality, successful ACOs are always encouraged to continue delivering services, a commissioner may elect to set counterfactuals in future rounds, using historic counterfactual priced activity levels rather than actual historic priced activity levels. Adjustments of this type are incorporated into the US ACO shared sharing programme.

[20] Centers for Medicare & Medicaid Services, Quality Measure Benchmarks For The 2018 And 2019 Reporting Years, Guidance Document, December 2017, https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/sharedsavingsprogram/Downloads/2018-and-2019-quality-benchmarks-guidance.pdf .

The extensions and refinements discussed in this chapter can be incorporated into formulae for an ACO's income. These are set out in appendix B2.

6 Multilateral Risk-Reward Sharing

Distributing risks amongst ICS partners

Where an ICS is made up of a number of partner organisations, working together to deliver care to a defined population, mechanisms will be required to determine how any shared savings or losses will be distributed amongst the partners. These 'within ICS' risk-reward sharing arrangements can be seen as a means of operationalising risk-reward sharing arrangements between the commissioner and ICS. Even, however, when a system control total is in place, an ICS may wish to operate a 'within ICS' risk-reward sharing arrangement.

We describe three approaches to distributing risks and rewards amongst ICS partners.

One solution involves **identifying one of the partner organisations that is nominated to carry the financial risk** should priced activity over-run and retain any savings should these be delivered. This simple solution has the benefit of minimising transaction costs and negotiations between partners. However, the lead partner may be concerned that its partners are not sufficiently incentivised to minimise activity and may seek additional leverage over its partners to manage its financial risk. The power imbalance which this arrangement introduces may diminish the sense of partnership which many would consider fundamental to the ICS model.

An alternative solution is to **allocate savings or losses between the partners according to the partners income** from the ICS contract. A partner receiving 60% of the income from the ICS contract would be entitled to receive 60% of any shared savings and would be required to pay 60% of any shared losses. Again, this solution requires minimal negotiations and transaction costs. However, the size of an organisation does not necessarily correlate with the contribution it makes to securing savings or cost over-runs. Organisations may be concerned that their partners will seek to gain from their efforts to improve efficiency without making similar efforts.

A third approach entails **detailed agreements between partners organisations** setting out the steps that each will take to secure efficiency and quality improvements. A partner's access to shared savings would be dependent on completing these actions. These arrangements could mirror the wider agreement between the commissioner and ICS or could take a simpler form to minimise transaction costs.

However risk and rewards are distributed between ICS partners, it would be prudent at the outset to describe the mechanisms and organisations that will be used to arbitrate on disputes between partners. Decisions about the construction of within-ICS risk reward sharing arrangement have significant implications for the composition of the ICS supply-chain. Issues of scheme complexity and cost volatility become more relevant as ICS's attempt to incentivise small-scale supply partners such as GP practices and local voluntary sector organisations.

In its recent paper on the financial arrangements for ICSs, NHS England mooted the possibility that risk and reward sharing arrangements may be established between a commissioner, an ICS and an acute provider that is not party to the ICS. In these circumstances, the commissioner would want to incentivise the ICS and acute provider to work together to reduce acute sector activity. It is certainly possible that the ICS and acute provider will hold differing views about the extent to which acute care activity can be reduced. Without hindsight, these differing views may be equally valid and arise from the organisation's unique commercial perspectives on the opportunities that exist. In particular, we might expect that the ICS will hold more optimistic views about the extent to which acute care activity can be avoided, with the acute provider adopting a cautious position more consistent with the history of steady demand growth. Designing risk-reward arrangements in these circumstances is not a trivial undertaking. We present in Appendix A our early thoughts on a novel solution to this problem.

In Conclusion

Risk and reward sharing is a simple and attractive concept, offering a commissioner the opportunity to co-opt and incentivise a provider to moderate growth in healthcare demand. However, robust schemes are likely to be complex and require careful construction.

If the NHS is to implement risk and reward sharing schemes then it must invest in developing, testing and documenting the underpinning methods and processes.

Given the paucity of evidence on the effectiveness of risk and reward sharing schemes, the NHS should commit to evaluating schemes as they are implemented.

Appendix A – Taking Account of Differences in Savings Expectations

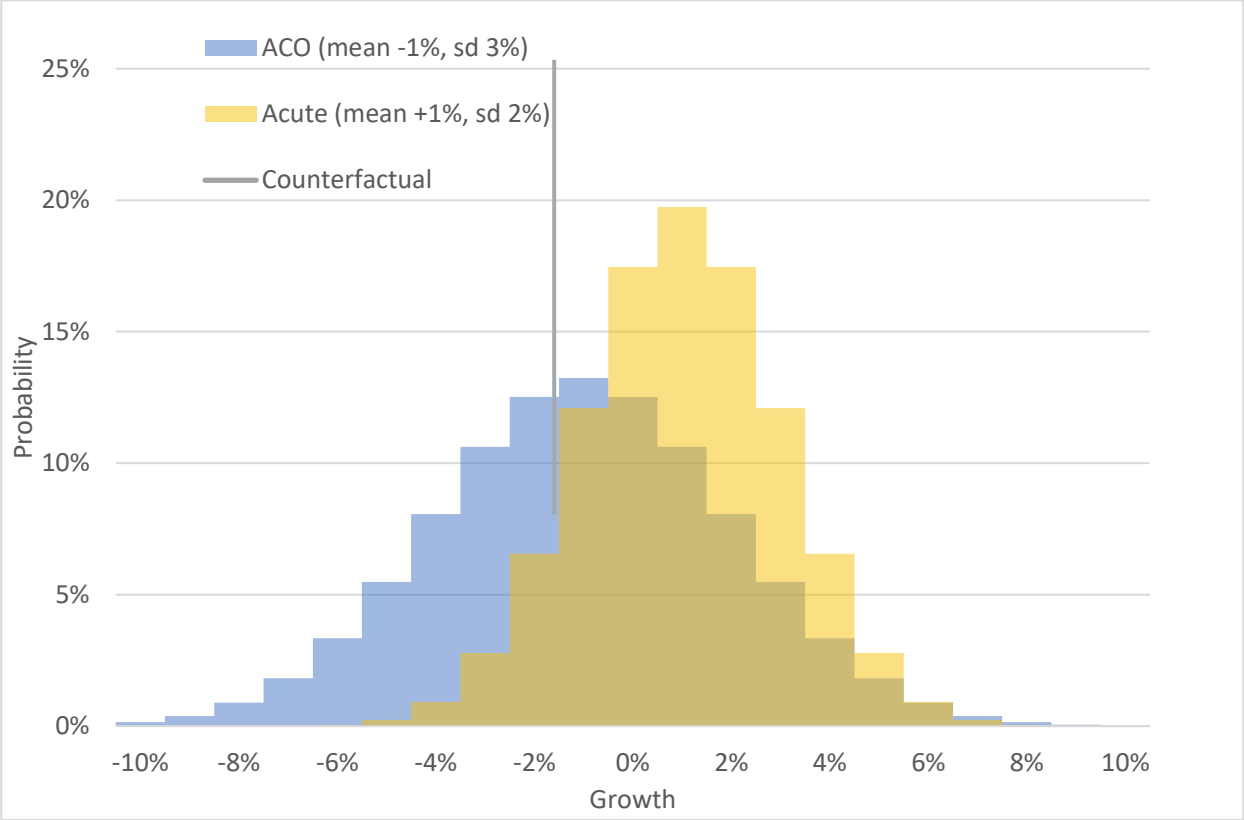
Although not strictly a prerequisite, the approaches to multi-lateral risk-reward sharing described in chapter 6 assume that ICS partners hold similar views about the scale of the opportunity to reduce costs, the extent to which this can be realised and the best ways to achieve it. In its recent paper on the financial arrangements for ICSs, NHS England mooted the possibility that risk and reward sharing arrangements may be established between a commissioner, an ICS and an acute provider that is not party to the ICS. In these circumstances, the commissioner would want to incentivise the ICS and acute provider to work together to reduce acute sectors activity. It is certainly possible that the ICS and acute provider will hold differing views about the extent to which acute care activity can be reduced. Without hindsight, these differing views may be equally valid and arise from the organisation's unique commercial perspectives on the opportunities that exist. In particular we might expect that the ICS will hold more optimistic views about the extent to which acute care activity can be avoided, with the acute provider adopting a cautious and position more consistent with the history of steady demand growth.

In these circumstances, how might a commissioner set up a risk reward sharing arrangement between the ICS and acute trust to incentivise the behaviours set out in chapter 3? Namely to reduce healthcare costs (the right way), to plan ambitiously and thoroughly and with due consideration for uncertainties. In particular, how might we distribute risks and rewards between the commissioner, the ICS and acute trust, when the ICS and acute trust do not share a view about the reduction of activity growth that could be delivered.

We present here our early thoughts on a novel solution to this quandary.

In order to take account of the ICS's and acute trust's differing views on the opportunity to deliver reductions in acute activity it will be necessary to quantify the ICS's and acute sector beliefs about future activity growth. In figure 6.1 below, the two curves encode these beliefs as two Normal distributions. Against a counterfactual growth of 2%, the peaks of the two curves indicate that the ICS believes that a 1% reduction in activity could be achieved (i.e. 3 percentage points lower than the counterfactual), whereas the acute trust believes a 1% growth is most likely (i.e. 1 percentage points lower than the counterfactual). The spread of the curves suggests that the acute trust is more confident with its prediction than the ICS. These curves are each defined by two parameters; the mean which indicates the most likely outcome and the standard deviation which indicates the degree of certainty of the prediction.

Figure 6.1: Spend Growth Expectations

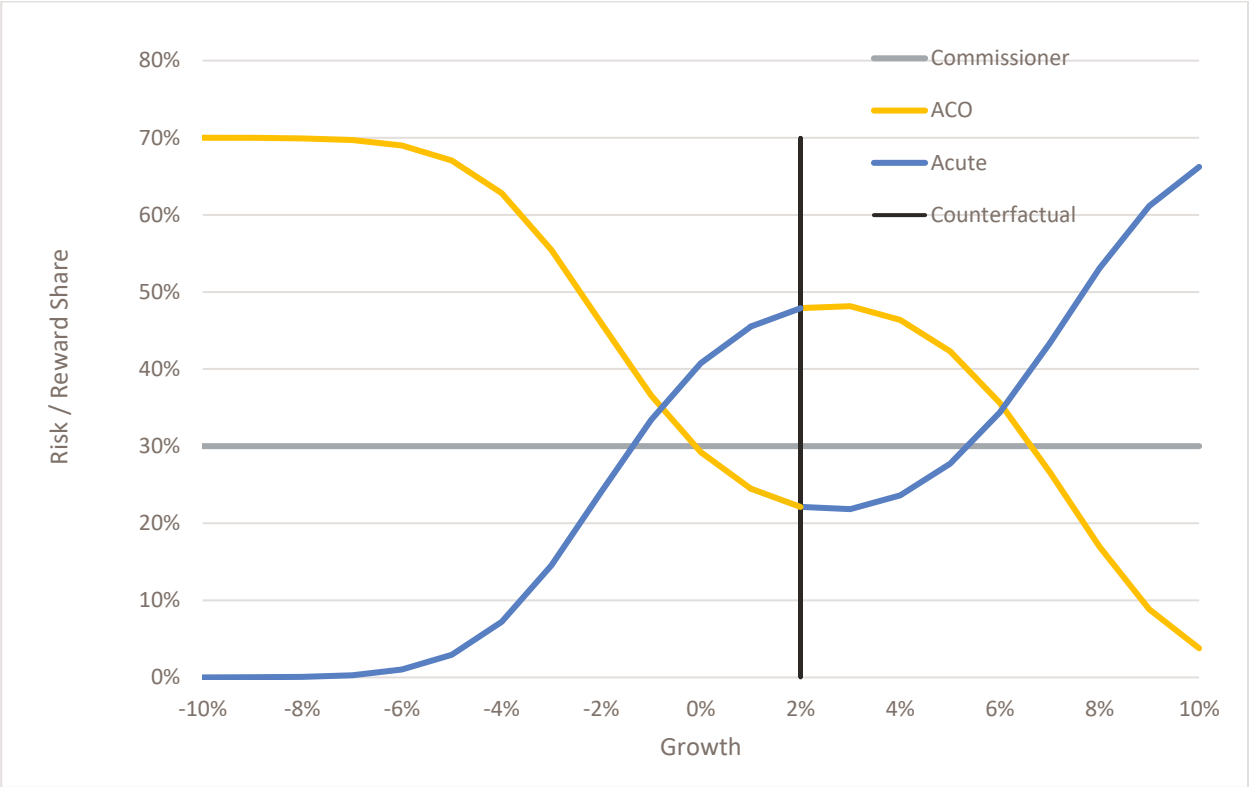


By comparing the height of the two distributions we can see that the acute trust is more confident than the ACO that spend growth will fall between -1% and +6%, whereas the ACO is more confident than the acute trust that spend growth will be lower than -1% and greater than +6%. Moreover, for any particular level of spend growth we can compare the ACO and acute sector expectation that growth of that level can be achieved (e.g. the ACO is nine times more confident that a spend growth of -4% will be achieved).

Once the year is passed, we will be able to determine whether the spend growth that was delivered, was more in keeping with the ICS's or the acute trusts prior belief. If the actual growth is less than 2%, then the commissioner will reward the ICS and acute trust with its share of the savings delivered. If growth falls between 2% and -1% then we argue that the acute trust should receive a larger share of these savings since the acute trust predicted this outcome with greater certainty than the ICS. If growth is less than -1% then the ICS should take the greater share of any savings. If growth is above 2% then the ICS and acute trust will be required to compensate the commissioner. If growth falls between 2% and 6% then the acute trust should pay the smaller share of these costs since it was more confident of this outcome than the ICS, whereas if growth is greater than 6% then the ICS should pay the smaller share.

If we assume an ICS-acute trust sharing rate of 70% (i.e. that the commissioner retains 30% of savings and pays 30% of cost over-runs), then the chart below demonstrates how savings or losses might be distributed based on the of the acute trust's and ICS's prior views about spend growth.

Figure 6.2: Share of savings & cost over-runs



This novel approach permits an ICS and an acute provider (outside the ICS) to hold differing views about the potential to constrain growth without forcing some unwelcome compromise. It requires both parties to be explicit about their assessments and to live by their judgements whilst rewarding efforts to curtail growth in the interest of the wider healthcare system.

The formulae to encode these risk-reward sharing arrangements is set out in appendix B3.

Appendix B1 – Basic Risk-Reward Sharing Formulae

$$ACO\ income = \begin{cases} \sum_{i=0}^{n} x_i\ t_i + s_u \left[c - \sum_{i=0}^{n} x_i\ t_i \right] & when \quad \sum_{i=0}^{n} x_i\ t_i < c \\ \sum_{i=0}^{n} x_i\ t_i + s_d \left[c - \sum_{i=0}^{n} x_i\ t_i \right] & when \quad \sum_{i=0}^{n} x_i\ t_i \geq\ c \end{cases}$$

Where:

x_i is the level of ACO activity for currency i
t_i is the tariff paid for currency i
c is the counterfactual level of spend
s_u is the proportion of savings allocated to the ACO when spend is less than the counterfactual
s_d is the proportion of losses allocated to the ACO when spend is greater than the counterfactual
$0 \leq s_u, s_d \leq 1$

Appendix B2 – Comprehensive Risk-Reward Sharing Formulae

$$ACO\ income = \begin{cases} \sum_{i=0}^{n} x_i\ t_i + min\left(q.s_u\left[c - \sum_{i=0}^{n} x_i\ t_i\right], c\ .m_s\right) & when \quad \sum_{i=0}^{n} x_i\ t_i < c(1-v) \\ \sum_{i=0}^{n} x_i\ t_i & when \quad c(1-v) \leq \sum_{i=0}^{n} x_i\ t_i < c(1+v) \\ \sum_{i=0}^{n} x_i\ t_i + max\left((1-q).s_d\left[c - \sum_{i=0}^{n} x_i\ t_i\right], -c\ .m_l\right) & when \quad \sum_{i=0}^{n} x_i\ t_i \geq\ c(1+v) \end{cases}$$

Where:

x_i is the level of ACO activity for currency i
t_i is the tariff paid for currency i
c_{adj} is the counterfactual level of spend having applied in-year adjustments
s_u is the proportion of savings allocated to the ACO when spend is less than the counterfactual
s_d is the proportion of losses allocated to the ACO when spend is greater than the counterfactual
m_s is the maximum savings that an ACO can receive as a proportion of the counterfactual
m_l is the maximum losses that an ACO can pay as a proportion of the counterfactual
q is the ACO quality score
v is the minimum absolute difference in spend from the counterfactual before an ACO can receive savings or pay losses as a proportion of the counterfactual
$0 \leq s_u, s_d, m_s, m_l, q, v \leq 1$

Appendix B3 – Risk-Reward Sharing Formulae – Differences in Savings Expectations

ACS, acute trust and commissioner share of savings or cost over-run.

$$ACS\ share = \quad s.\left[\frac{\Phi\left(\frac{g_r + 0.005 - x_{acs}}{d_{acs}}\right) - \Phi\left(\frac{g_r - 0.005 - x_{acs}}{d_{acs}}\right)}{\Phi\left(\frac{g_r + 0.005 - x_{acs}}{d_{acs}}\right) - \Phi\left(\frac{g_r - 0.01 - x_{acs}}{d_{acs}}\right) + \Phi\left(\frac{g_r + 0.005 - x_{at}}{d_{at}}\right) - \Phi\left(\frac{g_r - 0.01 - x_{at}}{d_{at}}\right)}\right]$$

$$Acute\ trust\ share = \ s.\left[\frac{\Phi\left(\frac{g_r + 0.005 - x_{at}}{d_{at}}\right) - \Phi\left(\frac{g_r - 0.005 - x_{at}}{d_{at}}\right)}{\Phi\left(\frac{g_r + 0.005 - x_{acs}}{d_{acs}}\right) - \Phi\left(\frac{g_r - 0.005 - x_{acs}}{d_{acs}}\right) + \Phi\left(\frac{g_r + 0.005 - x_{at}}{d_{at}}\right) - \Phi\left(\frac{g_r - 0.005 - x_{at}}{d_{at}}\right)}\right]$$

$$Commissioner\ share = \ 1 - s$$

Where:

s is the proportion of savings allocated to the ICS and acute trust
x_{acs} is the mean spend growth expected by the ICS
x_{at} is the mean spend growth expected by the acute trust
d_{acs} is the standard deviation of spend growth expected by the ICS
d_{at} is the standard deviation of spend growth expected by the acute trust
g is the observed spend growth
$g_r = \lfloor 100g + 0.5 \rfloor / 100$ is the observed growth rounded to the nearest percentage point
$\Phi(x)$ is the cumulative distribution function of the standard normal distribution
$0 \leq s \leq 1$

These arrangements may be generalised as in chapter 5 and appendix B2 to incorporate unbalanced and asymmetric sharing, quality adjustments, risk-reward capping and to take account of variation and volatility.

Strategy Unit
Tel: 0121 612 1538
Email: strategy.unit@nhs.net
Twitter: @strategy_unit

The Strategy Unit
Tel: +44(0)121 612 1538
Email: strategy.unit@nhs.net
Twitter: @strategy_unit