UNIVERSITY of York

RESEARCH

CHE

Centre For Health Economics

NIHR | Policy Research Unit in Economics of Health Systems and Interface with Social Care

Payment Reform, Purchaser and Provider Decisions and the Performance of Emergency Healthcare Systems: The Case of Blended Payment in the English NHS

Martin Chalkley, Hugh Gravelle,
Nikita Jacob, Rita Santos, Luigi Siciliani

CHE Research Paper 187

Payment reform, purchaser and provider decisions and the performance of emergency healthcare systems: The case of blended payment in the English NHS

[a]Martin Chalkley
[a]Hugh Gravelle
[a]Nikita Jacob
[a]Rita Santos
[b]Luigi Siciliani

[a]Centre for Health Economics, University of York, UK
[b]Department of Economics and Related Studies, University of York, UK

August 2022

Background to series

CHE Discussion Papers (DPs) began publication in 1983 as a means of making current research material more widely available to health economists and other potential users. So as to speed up the dissemination process, papers were originally published by CHE and distributed by post to a worldwide readership.

The CHE Research Paper (RP) series takes over that function and provides access to current research output via web-based publication, although hard copy will continue to be available (but subject to charge). Results and ideas reported in RPs do not necessarily represent the final position. Work reported in some RPs should be seen as work in progress and may not have been subject to peer review at the time of publication.

Acknowledgements

This project is funded by the National Institute for Health Research (NIHR) Policy Research Programme (reference PR-PRU-1217-20301). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

No ethical approval was needed.

Further copies

Only the latest electronic copy of our reports should be cited. Copies of this paper are freely available to download from the CHE website www.york.ac.uk/che/publications/. Access to downloaded material is provided on the understanding that it is intended for personal use. Copies of downloaded papers may be distributed to third parties subject to the proviso that the CHE publication source is properly acknowledged and that such distribution is not subject to any payment.

Printed copies are available on request at a charge of £5.00 per copy. Please contact the CHE Publications Office, email che-pub@york.ac.uk for further details.

Centre for Health Economics
Alcuin College
University of York
York,
YO10 5DD, UK
www.york.ac.uk/che

Abstract

This paper constitutes the first and foundational output of the ESHCRU2 project 3 - Analysis of purchaser-provider contracts: modelling risk sharing and incentive implications. In this project, we have focused on the implications of payment reform of what is called blended payment for emergency care. This paper sets out the theoretical model developed to understand how behavioural choices could be influenced by this payment reform. We construct a framework in which two organisations - a hospital and a purchaser - influence respectively admissions from, and attendance at, emergency departments. These decisions are each influenced by the payment system and interact to determine an equilibrium. We show how the equilibrium is affected by the characteristics of the hospital and the purchaser and how it will be changed by shifting towards a greater reliance on fixed payment. We further establish what outcomes (in terms of admissions and attendances) can be sustained as an equilibrium under different parameters of the payment system.

1. Introduction

In fixed price hospital payment systems, often termed Diagnosis Related Group (DRG) systems, providers of hospital services are paid a specific amount for each patient they treat with a particular medical condition (diagnosis). In respect of acute hospital services, for example, this approach has been adopted across many jurisdictions, starting with the prospective payment system for Medicare in the US (Guterman and Dobson, 1986), having been adopted widely in Europe (Reinhard et al., 2011) and are now being adopted in low- and middle-income settings (Mathauer et al., 2012). Prices are usually set to reflect the average cost of the treatment of a condition or diagnosis. In the English NHS the payment mechanism for hospital services is set out in the National Tariff Payment System (NTPS)1, and for elective care this mandates prices (termed tariffs) that are calibrated to average costs. However, emergency hospital care funding was reformed starting in 2019 and is now funded through a mixture of a national tariff, with adjustments to the price for treatments above an indicative volume, and some element of a fixed budget agreed between commissioners and providers of care2 who also have discretion in adjusting the national tariff to local circumstances. This approach is referred to as blended payment. The essence of blended payment, combined with local discretion in terms of setting a price for each unit of activity, is to establish a two-part tariff in which the prices of activity are reduced below the previously mandated national price, and the provider is compensated for that by a fixed budget. This policy change has brought into focus the potential incentive and risk-sharing properties of different payment mechanisms for the emergency care system.

There is a long tradition in health economics of examining the incentive properties of payment mechanisms following the analysis of incentive contracts (Laffont and Martimort, 2001) and considering the particular concerns to ensure cost control and high quality care (Ma and Mak, 2019). The common approach is to specify alternative payment systems as contracts and to examine under what conditions these contracts can achieve what a *purchaser* of health services desires in regard to a number of possible dimensions of care (Chalkley and Malcomson, 1998a; Ma, 1998). One key lesson from this literature is that what constitutes an appropriate contract depends on the nature of the services being contracted for, the motivation of the providing institution, the goals and objectives of the purchaser of services, and the information that is available both in specifying the contract and that which can be relied upon for enforcing it.

There has hitherto been no specific application of this approach to consider the particular circumstances and imperatives of emergency health care delivery in a nationalised health system such as the English NHS or its adoption of a two-part tariff. The key elements of this context are: service providers are independently managed health care organisations and are predominantly an integral part of the public service but required to operate within their budgets; emergency health care in hospitals is delivered partly at the discretion of hospitals in response to patients delivered to them through a number of different pathways (walk in, sent by GP, delivered by ambulance); demand for services is perceived to sometimes exceed the capacity of service provision; emergency care is part of an integrated system in which patients have a number of potential avenues to having their urgent needs met; the purchasers of emergency hospital services also have responsibility for ensuring care in other settings. These last two features are particularly pertinent since they indicate that a whole system approach is required, and that constitutes an important focus for our investigation.

1 NTPS is summarised here https://www.england.nhs.uk/pay-syst/national-tariff/#:~:text=The%20national%20tariff%20is%20a,cost%20effective%20care%20to%20patients.
2 See https://www.england.nhs.uk/wp-content/uploads/2021/02/20-21_National-Tariff-Payment-System.pdf

In this paper we set out and analyse a simple theoretical model in which *both* purchasers and providers of emergency care make important decisions that will impact on how the emergency care system performs, both in terms of attendances at emergency departments and subsequent admissions to hospital. Those decisions are influenced by the nature of the payment that the purchaser makes to the provider of hospital services but also crucially are interdependent, since each organisation will be impacted by the decision made by the other. The emergency care outcome is viewed as an equilibrium between these decisions, conditional on the form of payment agreed or mandated between the two organisations. We are particularly interested in how changing the prices paid for hospital treatment will impact on emergency care in equilibrium because this lies at the heart of the adoption of blended payment - a reduced reliance on prices and an increased reliance on fixed transfers.

This model serves a number of purposes. First, it establishes the potential avenues through which blended payment can be expected to influence the performance of emergency care systems in the English NHS and highlights that there may be important trade-offs between reducing hospital admissions and increasing hospital attendances. Second, it establishes a framework for understanding and analysing the broader determinants of admissions and attendances - a lens through which to view these data. Third, it provides a means for evaluating the potential of blended payment to achieve improved performance in emergency care systems.

Our model can be summarised in general terms, as follows. There is a given population of individuals for whom emergency health care is being organised. For simplicity, we consider only a single illness that individuals may incur, and we incorporate uncertainty via probability distribution over the number of people who will become ill. There are two organisations involved in the delivery of care, which we term the purchaser and the provider. In the context of the NHS in England these correspond to a Clinical Commissioning Group (CCG) and an NHS hospital trust respectively. The purchaser puts in efforts to arrange some treatments for ill patients outside of the hospital whilst, of the remainder who go to hospital, the provider determines how many are treated intensively (admitted). The purchaser pays for hospital treatments according to a price per attendance and admission, and also makes a fixed transfer to cover costs. The model proceeds by examining how the prices (for both attendance and admission) and fixed payment elements affect both decisions separately and utilises an optimising framework in which each organisation chooses their effort to maximise an objective function. Those decisions are conditioned on both the payment and the choice made by the other agency. This latter element suggests that equilibrium should be characterised by the coincidence of each organisation's choice in respect to the other. We examine the characteristics of this equilibrium and its determinants. We can thereby consider questions regarding outcomes can be sustained by different payment mechanisms. In this setting, the overall performance of the system is characterised by the vector of choices made by the agencies and the key questions addressed concern how performance is influenced by payment. To examine that in a way relevant to the movement towards blended payment, we focus on how changing prices influences outcome. We show that moving towards blended payment reduces the propensity of a hospital to admit patients but increases expected number of attendances at hospital. The model also naturally gives rise to a specification of the other factors determining equilibrium, and this set of factors guides our subsequent empirical investigations into the likely range of influence that payment systems will have in practise.

The main literature that we build on is the economics of contracts for healthcare and incentives, which is summarised by Ma and Mak (2019**).** An alternative Operations Research/Management Science perspective on incentives and contracting is in Fainman and Kucukyazici, (2020). In the economics literature the role of fixed prices contracts has been extensively studied under the assumption that a contract is specified by a welfare motivated purchaser. Following Chalkley and

Malcomson (1998a, 1998b), Ma (1998, 1994) a substantial focus has been placed on whether fixed price contracts can deliver incentives to simultaneously promote quality and low-cost service provision under a great variety of assumptions and settings, such as when there are many dimensions to choice, or providers have unknown characteristics. Eggleston (2009, p. 2020), Jack (2005), Kaarboe and Siciliani (2011) serve as examples of this approach. A lot can hinge on the objectives of the provider and we follow an approach which assumes that providers have a degree of altruism (Jack, 2005; Ma, 1997). In accordance with this general approach, our model emphasises the role of prices in affecting behaviour but we depart from the standard model in a number of ways. First, we assume that the purchaser is an agency with its own objective (which need not conform to social welfare) and that it too makes decisions impacting on the delivery of care. Second, we explicitly account for a specific policy shift from purely fixed prices towards a two-part tariff. Within the theoretical literature, it is often assumed that a two-part tariff is feasible and there is no presumption that prices will be set to exactly cover costs. This is in contrast to the practical implementation of fixed price systems. Our approach to this issue is therefore to consider the impact in a practical system of unconstraining prices such that they do not cover costs and compensating the hospital via a fixed budget. We do, however, make allowance for constraints such as non-negativity of prices and transfers. A very useful summary of the institutional context for our model and its analysis is in Pauline and Kath (2020).

In the following section we set out our model by detailing the decision problems of a provider and a purchaser, and then considering the determination of equilibrium in an emergency care system. The subsequent section provides the main analysis of how blended payment might affect outcomes and the relationship between those outcomes and system performance. That is followed by a discussion of some possible extensions of our model and the likely robustness of our findings to generalisations. The final section provides a summary of the policy implications of the analysis.

2. The Model

Our focus is on an emergency health system in which we assume there is fixed population of n individuals. A proportion ρ of that population will become ill and require emergency care, where ρ is a continuous random variable with support $\{0,1\}$ and density $g(\rho)$. This proportion can be expected to vary according to the characteristics of the population with older, more socially deprived populations or those an otherwise higher incidence of disease characterised by a higher expected value of ρ. We examine below how the number of ill patients that arrive at a hospital is determined but for now simply denote that number by $N < n$, which is assumed to be a random variable with density $f(N)$, mean μ and variance σ^2.

The hospital's decision

Prior to patients arriving we assume that the hospital sets a policy that determines what proportion α of these patients will be admitted for inpatient care - α is the hospital's admission rate. This policy could be operationalised by setting a threshold of severity of illness above which the patient will be admitted.

There are costs and benefits to the hospital that vary with the value chosen for α. These arise both from the processes and actions that the hospital has to take for either admitted or non-admitted patients and also come from the contract that it has with the purchaser, which will determine how much it will be paid for the treatments it carries out - varying according to whether a patient is admitted or not. We assume that all of these influences can be captured in a function of N and α which is denoted $v(\alpha, N)$ which reflects the hospital's concern for patients and its financial surplus or deficit. The hospital is assumed to choose α to maximise the expectation of this function, so that its optimal choice is satisfied.

$$\max_{\alpha} V(\alpha) = \int_0^n v(\alpha, N) f(N)\, dN. \qquad (1)$$

Provided that $V(\alpha)$ is a concave function the optimal value of α which is denoted α^* can be found from the solution to

$$V_\alpha(\alpha^*) \equiv \int_0^n v_\alpha(\alpha^*, N)\, f(N)\, dN = 0 \qquad (2)$$

The questions of interest are how the circumstances of the hospital (such as the costs of its treatments and how it values those), and the form of the payment it receives, affect its choice α^*. The answers depend on precisely what assumptions are made regarding the functions $v(\alpha, N)$ and $f(N)$. In order to proceed in as transparent a way as possible we, therefore, make simplifying assumptions regarding these functions. The assumptions are not intended to be general, but rather capture the most essential aspects of the hospital's decision problem and ensure there is a readily interpretable solution for α^*.

Turning first on the objective of the hospital, one key element of the English NHS which is the focus of our study is that hospitals are regarded as a part of a national system and are motivated in general by the goals of that system to provide necessary health care. Hence, we assume that the hospital inherently values the treatments it provides. Specifically, we assume that the provider has attached a value b_1 for each patient treated in the emergency department but not admitted, and a value b_2 to each patient treated through admission. Given the hospitals policy there will be $(1 - \alpha)N$ patients for whom the benefit perceived by the hospital is b_1 and αN that the hospital values at b_2. Hence the total benefit of the treatments delivered, as perceived by the hospital provider, is

$$B(\alpha, N) = b_1(1-\alpha)N + b_2\alpha N, \qquad (3)$$

where $b_2 > b_1 > 0$ are constants.

Care is needed in considering the concept of costs since not all costs are observed. Whilst it is commonplace to assume that there is a constant cost of treating each additional patient, this does not capture the idea that at higher levels of activity treatment becomes more difficult. This may be reflected in the stress of staff and the deprecation of facilities, more than in reported financial figures, and is a reasonable factor to include. To do that we assume that the cost of treating patients is increasing in the squared value of the number of patients treated (a convex cost function) and by analogy with the formulation for benefits write

$$C(\alpha, N) = F + (N - \alpha N)^2 c_1 + (\alpha N)^2 c_2, \quad (4)$$

where F, c_1 and c_2 are positive constants and, reflecting the fact that inpatient care is more intensive, $c_2 > c_1$.

We assume that the payment the hospital receives[3] is comprised of a fixed sum per patient treated in the emergency department of p_1 a price per patient treated as an admission of p_2 and a fixed financial transfer of T. So, the total revenue of the hospital is given by

$$R(\alpha, N) = T + p_1(1-\alpha)N + p_2\alpha N \qquad (5)$$

The notion of activity-based contract is captured by the potential for p_2 to vary with the volume of treatments. In practice it is expected that this relationship will be piecewise linear and specify differing prices applying to different volumes of admitted patient activity. In such cases the function p_2 will be defined by a number of constant prices and the thresholds at which they apply. To capture, in as simple as way as possible, the essence of such an arrangement we consider the difference between a contract in which p_2 is constant and T is zero (a pure activity-based approach) and a contract where p_2 is zero and T is positive (a simple blended contract in which only attendances are subject to activity based payment), and conceptualise the contract choice decision as being a choice of p_2 within this range. Although not specifically an element within the context of blended payment arrangement we also allow for the possibility that p_1 could be varied.

In the case of all payments, it is anticipated that the hospital will receive enough to cover the anticipated (expected) costs of its activity. Hence there is an overall budget requirement, given the hospital's choice of α^* of,

$$\int_0^n [R(\alpha^*, N) - C(\alpha^*, N) f(N)]\, dN = 0 \qquad (6)$$

which will determine restrictions on the values of T, p_1 and p_2.

For a given specification of the hospital's revenue function it is now possible, given the remaining assumptions, to solve the hospital's optimal choice α^*. For simple linear specifications of revenue, a closed form solution is possible and is a useful benchmark for analysis. In this section, we assume that p_1, p_2 and T are fixed constants. Given the assumptions made, the function $v(\alpha, N)$ is quadratic in both of its arguments. Hence, when taking expectations over N, only linear and squared terms

[3] The current payment system specifies a price for an attendance, which we have denoted p_1 and an additional payment if the patient is admitted. So in practise the payment system specifies p_1 and a differential payment. For the model we label the sum of p_1 and that differential by p_2.

appear. Using the fact that $E[N^2] = \mu^2 + \sigma^2$ for any density function $f(N)$, the hospital's objective can be written as

$$\int_0^n v(\alpha, N)\, f(N)\, dn = b_1(1-\alpha)\mu + b_2\alpha\mu - [F + (1 + \alpha^2(\mu^2+\sigma^2) - 2\alpha\mu)c_1 + \alpha^2(\mu^2 + \sigma^2)c_2] + T + p_1(1-\alpha)\mu + p_2\alpha\mu \quad (7)$$

Differentiating (7) with respect to α and equating to zero gives the condition satisfied by α^*as,

$$-b_1\mu + b_2\mu - 2c_2\,\alpha^*(\mu^2+\sigma^2) + 2c_1(1-\alpha^*)(\mu^2+\sigma^2) + (p_2 - p_1)\mu = 0. \quad (8)$$

Equation (8) can be solved to yield,

$$\alpha^* = \frac{(p_2 - p_1 + b_2 - b_1)\mu + 2c_1(\mu^2+\sigma^2)}{2(c_1+c_2)(\mu^2+\sigma^2)} \quad (9)$$

Equation (9) indicates that the hospital's optimal choice of the proportion of patients to admit depends upon its valuation of inpatient and emergency treatments, the costs of these treatments, both the magnitude and variability of the demand for its services (the number of patients who present for treatment), and the contract that it has with the purchaser. In the particular instance analysed here, that contract is characterised by two prices, one for each of the emergency department and inpatient treatments it provides.

Some properties of the optimal choice as a function of these various parameters can be determined easily. For example, by inspection it can be seen that α^* is increasing in the difference between p_2 and p_1 increasing in the difference between b_2 and b_1 and decreasing σ^2. The α^* increases with p_2 but not in a linear form, so the increase of α^* due to an increase in p_2 is moderated by the provider costs and A&E demand. Other comparative statics results can be derived more formally by differentiating the expression on the right hand side of equation (9) with respect to a parameter of interest. It follows that α^* is decreasing in μ. The intuitive explanation of all of these results comes from considering the impact of a changing parameter on the marginal benefit and marginal cost of α. As α increases, the number of admitted patients increases with consequent costs captured by c_2 and benefits in the form of increased payment (through p_2) and perceived patient welfare (through b_2), with the magnitude of these marginal costs and benefits in turn influenced by the environment the hospital faces in terms of the magnitude and variability of demand. The more volatile is demand, the greater the variation in cost that the hospital faces and so, other things equal, it reduces admissions.

Whilst the solution of the model is most naturally presented in respect of the hospital's choice variable, there is value from a practical and policy perspective in examining the *consequences* of that choice. Most notably, as the driving focus of our analysis is on the performance of an emergency care system, it is relevant to consider implications in respect of the costs of treatment. Overall, the costs of treatment will have to be met out of public funds, with the parameters of whatever contractual system is in operation adjusted so as to compensate the hospital for these costs.

The optimised cost of treating N patients can be found by substituting the optimal value of α into the equation (4). This gives

$$C(\alpha^*, N) = F + \left(N - \frac{(p_2 - p_1)\mu + (b_2 - b_1)\mu + 2c_1(\mu^2+\sigma^2)}{2(c_1+c_2)(\mu^2+\sigma^2)}\, N\right)^2 c_1 + \left(\frac{(p_2 - p_1)\mu + (b_2 - b_1)\mu + 2c_1(\mu^2+\sigma^2)}{2(c_1+c_2)(\mu^2+\sigma^2)} N\right)^2 c_2 \quad (10)$$

which defines the cost of emergency hospital care purely in terms of parameters that can be viewed as exogenous to the hospital. Taking expectations of the expression in equation (10) with respect to N provides the expected value of system costs. It is, therefore, possible to infer how this cost varies with those parameters. For example, considering an increase in the price p_2 then since α is increasing in p_2 and given we would expect c_2 to be greater than c_1 then expected cost will increase with p_2. This finding is more important and subtle than at first sight. What it illustrates is that, even if the increased price is accompanied by a reduction in the fixed financial transfer T, the overall cost to the purchaser must increase. This is because the price causes a change in behaviour - a move towards greater inpatient treatment - that has real resource implications that will have to be met by the payment system.

This reasoning begins to reveal the value of a model as a means for understanding how the performance of a system may be influenced by the *form* of payment, not just the level of payment. In this instance, the model predicts that a smaller reliance on fixed prices may reduce system costs. That is not to say such a change is desirable. It will also result in some seriously ill patients not being admitted for the care they perhaps need, but, nevertheless, the basic insight is important.

This avenue of analysis can be pursued further. Suppose it was possible to influence a hospital's own valuation of treatments by reducing b_2. The model predicts that this change will reduce α and hence reduce the expected overall costs of emergency care. Following such a change, the purchaser would be able to reduce any fixed budget element without causing the hospital to incur an expected deficit.

These basic insights will be explored further below in the context of thinking about system performance as the interaction of hospital and purchaser decisions, and how different contract mechanisms can influence those.

As discussed above, the model, the solution of which is encapsulated in equation (9), provides both a guide to how to understand the admission rate of hospitals empirically and a lens for viewing the empirical results. In respect of guiding an empirical strategy the model suggests that having controlled for as many characteristics of patients as possible, any remaining variation in admission rates can be interpreted as indicating differences in the cost and valuation parameters of hospitals. It also suggests that demand parameters have a role to play and we pick up this strand of thinking when looking at purchaser's decisions and how those might in turn influence demand.

A crucial question from a policy perspective is how much variation there is and what effect it has, since that gives an indication of both how choice of contract might be directed (towards those hospitals that have the most unfavourable parameters) and what it can hope to achieve.

Purchaser choice of effort to reduce attendances

In the English NHS, the agency charged with overall responsibility for facilitating the supply of healthcare to a given population has generically been referred to as a *purchaser*, following the institution of the separation of purchasing and supplying functions in the 1990s. The specific name given to purchasers, and the extent of the populations they are charged with serving, has changed over time - and continues to undergo development. The data that we will subsequently consider relates to purchasers who are called Clinical Commissioning Groups (CCGs), but the framework we consider can be applied quite generally to any organisation unit that has the function of purchasing health care on behalf of its population.

Individuals requiring urgent medical attention can be treated in a number of different settings, but a crucial distinction is often made between those treated within a hospital and those treated in a

community or primary care context. This is an important distinction because hospital care is resource intensive and there is often perceived to be potential savings from ensuring that health care needs are met outside of hospitals, if that is both possible and beneficial to the individual concerned.

The previous section considered how treatment choices emerge once an individual arrives at a hospital, and draws a distinction between individuals who are admitted for further intensive treatment and those who are treated within the emergency department context and returned to other modes of care. This section focuses on a decision that a purchaser of hospital services can make in respect of how much to invest in out-of-hospital services, where these can serve as an alternative for an individual arriving at a hospital for treatment. Our focus is on understanding how different payment mechanisms impact on decision making that gives rise to a configuration of a local healthcare system, and especially on the role that so-called blended payment arrangements can make, if used to replace pure activity-based (National Tariff Payment System) arrangements. In any contractual relationship, a financial transfer is made between the purchaser and the provider of services, and usually attention centres on the latter of these. However, for the reasons set out above, in emergency healthcare decisions of both hospital providers and purchasers are important.

Models of provider behaviour are commonplace in economic approaches to healthcare delivery and hence we set out a specific model, without much recourse to fundamental considerations of how such organisations are modelled or the domain of their decisions. CCG purchasers are a more NHS-centric institution. A CCG is charged with ensuring the provision of a broad range of health care to its population. To achieve that it has a given budget. Population healthcare needs are many and diverse, and a key function of purchasers in the NHS is to determine priorities as to which needs can be met and how. Hence, budget allocation across competing demands for service provision is a key task.

Whilst meeting some healthcare needs is discretionary, a fundamental organising principle of the NHS is that individuals' urgent and pressing health concerns are addressed, and these concerns are exemplified in respect of emergency care. Hence, there is little discretion in respect of emergency care - the purchaser has to ensure that all individuals who present themselves for urgent care are treated. There is further little overt choice regarding the intensity of care that emergency cases should receive - it needs to be sufficient to restore them, wherever that is medically possible, to its previous state and, if necessary, access further discretionary healthcare in the future.

Therefore, we regard funding the provision of emergency care as being essential, and largely non-discretionary, from a purchaser perspective - this avoids the need to explicitly model the choice between emergency care and other service provision. The issues to be addressed are how and in what setting that emergency need is met, not whether it will be met or what it is designed to achieve. Given its overall budget, a purchaser therefore needs to allocate a provision for emergency care, and whatever remains is available to be used to determine the configuration of other healthcare services. This suggests that a key goal for the purchaser is to ensure the provision of emergency services at the lowest cost to *its own budget*. This is the approach that we adopt.

As previously stated the number of individuals in the purchaser's population is assumed fixed and equal to n and a proportion ρ of that population will require emergency care where ρ is a continuous random variable with support $\{0,1\}$and density $g(\rho)$. We assume that the emergency care can be provided either within or outside a hospital setting. The purchaser makes investments out of its fixed budget that will determine what proportion $(1-\beta)$ of individuals can be treated outside of hospitals, with the proportion β then seeking care through a hospital emergency department. β is the purchaser's emergency attendance rate. We denote the expenditure on these

investments as purchaser effort e. The more effort the purchaser exerts the higher the cost it incurs, but with a smaller proportion that will attend hospital.

The purchaser expends effort so as to minimise the expected cost of emergency care for its population. If the expected cost *to the purchaser* of $y$patients attending hospital is $C_h(y)$, then the cost minimisation problem can be written

$$\min_e [e \; + \; \int_0^1 [C_h(\beta(e)\rho n)] g(\rho)\, d\rho] \quad (11)$$

The function $\beta(e)$ is assumed to be continuous, invertible and decreasing with domain {0,∞} and range {0,1}.

The cost C_h depends on the contract that the purchaser has with the hospital that supplies emergency care. Consistent with the analysis in Section 2.1, we assume that this contract takes a simple linear form of blended payment so that

$$C_h(\beta(e)\rho n) \; = \; T \; + \; p_1(1-\alpha)\rho\, n \; + \; p_2 \alpha \rho n. \quad (12)$$

Substituting the expression on the right-hand side of (12) into (11) and differentiating, the condition characterising the purchaser's optimal (cost minimising) choice of effort, e^* is

$$1 + \; n\bar{\rho}\beta_e(e^*)[(p_1(1-\alpha) \; + \; p_2\alpha] = 0 \quad (13)$$

where $\bar{\rho}$ is the mean of the distribution $g(\rho)$. We assume that $\beta_{ee}$is positive for all e and hence condition (13) is sufficient to define cost minimising effort.

From (13) it follows that the purchaser's choice of effort depends on $n, \bar{\rho}, \alpha, p_1$ and p_2. Denoting the expression in (13) by $h(e^*, z)$ where z is the factor of interest, the direction of effect of z on e^* can be determined by equating the differential of h to zero and is given by

$$\frac{de^*}{dz} = sign[-\frac{h_z}{h_e}] \; = \; sign[-h_z] \;] \quad (14)$$

where the last equality holds on the assumption that the second order condition for minimisation of cost is satisfied.

Applying the evaluation implied in (14) indicates that the purchaser's optimal effort is increasing in all of $n, \bar{\rho}, \alpha, p_1$and p_2 and hence that β is decreasing in all of these. The intuition being that anything that increases the cost of a marginal patient treated in a hospital setting (that being the consequence of reducing effort) increases the return to the purchaser of ensuring care outside of the hospital setting. An increase in effort (e) implies a smaller β.

Hence, the model again indicates that the outcome of decision making in respect of emergency care depends on the *form* and precise details of the payment mechanism between the purchaser and the provider. Here, according to the model, a greater reliance on the fixed element of a 'blended' contract will reduce the effort that a purchaser exerts to reduce hospital attendance.

The fixed payment T has no impact on the purchaser's choice of effort, but nevertheless will need to be set to cover the costs that a hospital incurs in treating the individuals who attend it. By way of comment, a purely fixed payment system, in which prices are set to zero, will in this context minimise the purchaser's effort and hence result in the greatest volume of hospital attendances.

This is worth noting in that it illustrates that in this decentralised decision making framework, the purchaser's optimal decision conditional on zero prices being mandated would be highly likely to increase the costs of the emergency care system.

As discussed earlier, the solution encapsulated in equation (13) provides a framework for structuring an empirical investigation of variation across CCGs in respect of the differing hospital attendance rates of their populations. The model distinguishes between exogenous factors - how sick the population is, the contractual mechanism in place etc - and the purchaser's response to these, which is determined by its perception of the value of increasing out of hospital care. If we could be sure that we had captured all relevant exogenous factors, any remaining variation across purchasers could be attributed to differences in their respective functions $\beta(e)$, giving rise to differences in chosen effort.

Equilibrium of the emergency care system

The sections above have set out models of two decision making agencies in respect of emergency care provision. A provider has been modelled as making a decision between treating individuals as inpatients, or treating them in the emergency department. The purchaser has been modelled as making a decision that affects how many individuals seek emergency care from a hospital in contrast to utilising other, predominantly primary care, facilities. In each case, the context for the decision being made is influenced by the choice made by the other agency. For a hospital, how many patients it receives determines how it wishes to balance inpatient and other treatment and, for the purchaser, the cost it incurs of patients attending hospital depends on the proportion of patients that will be admitted. Our approach is to suppose that the emergency care system comprises one purchaser and one provider.

The overall outcome of emergency care, the proportion of individuals attending hospital, and what proportion of those are admitted, depends on the interaction of the separate decision making processes of a purchaser and a provider. As always, in such an interdependent setting, there are questions regarding how much of the interdependency decision makers take into account when formulating their choices. Formally, the models set out above have considered the purchaser and provider to have taken as given the choice made by the other. Hence, one approach that can be taken to establish an overall description of decisions is to look for coincidences of choices that do not cause either purchaser or provider to wish to revisit their decision. In the terminology of game theory this approach captures the notion of a Nash equilibrium. It is the approach that we follow here.

A Nash equilibrium constructed on this basis presumes that the underlying structure of the decision making processes is one of simultaneous choices, and that the decision makers act unilaterally (without consulting each other). There are alternatives, such as to consider one decision maker moving first and to commit to their decision, which then anticipates the response of the other decision maker. In such a sequential choice setting, the *leader* can form an assessment of the likely response of the *follower,* and factor that response into their decision. In our setting with, for example, the purchaser as leader, the purchaser might anticipate the choice of α being a function of their choice of effort, and includes that response of the provider into their assessment of costs and benefits. A further possibility is that, rather than acting unilaterally in making decisions in a so-called non-cooperative game, the purchaser and provider negotiate over some aspects of their decisions. Such an approach would invoke theories of bargaining to attempt to describe the likely outcome. We leave such alternative formulations to future research and consideration.

In considering the number of individuals that the provider receives, we can refer to the number of individuals that the purchaser will expect to send, conditional on how much effort it has expended

on other emergency care provision. Hence, expressions (6) - (9) that involve N can be rewritten in terms of $n\beta\rho$ and expectations over N now taken over ρ so that μ is replaced with $\beta(e^*)n\bar{\rho}$ and σ^2 replaced with $(\beta(e^*)n)^2{\sigma_\rho}^2$. Hence, both the mean and variance of the attendances the providers faces are dependent on the purchaser's effort as that influences β.

Making these substitutions into (9) and writing this as a condition for equilibrium (equating to zero) gives

$$\alpha^* - \frac{(p_2-p_1)\beta(e^*)n\bar{\rho}+ (b_2 - b_1)\beta(e^*)n\bar{\rho} + 2c_1(\beta(e^*)n\bar{\rho}^2+(\beta(e^*)n)^2{\sigma_\rho}^2)}{2(c_1+c_2)(\beta(e^*)n\bar{\rho}^2+(\beta(e^*)n)^2{\sigma_\rho}^2)} = 0 \quad (15)$$

The second condition for equilibrium can be written directly from (13) replacing α with $\alpha^*$to give

$$1 + n\bar{\rho}\beta_e(e^*)[(p_1(1-\alpha^*) + p_2\alpha^*] = 0. \qquad (16)$$

Equations (15) and (16) constitute a pair of simultaneous conditions in terms of two unknowns, e^* and α^*. It is convenient to re-express (16) in terms of the optimised value of β^*. Denoting the inverse of the function $\beta(e)$ by $e(\beta)$ equation (16) can be written as a function of β^*as

$$1 + n\bar{\rho}\frac{1}{e_\beta(\beta^*)}[(p_1(1-\alpha^*) + p_2\alpha^*] = 0, \qquad (16')$$

whilst $\beta^*(e)$ in (15) can be replaced by β^* so that equations (15) and (16') define two unknowns $\alpha^*$and β^*.

Standard methods of comparative statics utilising the differentials of the right hand side expressions in equations (15) and (16') can be used to determine how the emergency care system (the propensity to attend an emergency department and the propensity for patients to be admitted) will respond to changes in the parameters of that system, including the contract that exists between purchaser and provider. However, it is both intuitive and instructive to proceed using a geometric analysis of the system equilibrium. Figure 1 illustrates the Nash equilibrium of the emergency care system by graphing (15), which gives the providers best choice of α^* to any given choice of β^* by the purchaser. It is labelled the provider's *best response* (red line). By analogy (16') defines the purchasers best choice of β^* to any given α^* chosen by the provider. It is labelled the purchaser's *best response* (blue line). As depicted in α^*, β^* space the provider's best response is shown as the steeper of the two. This conforms with an equilibrium that is stable if the two decisions are made sequentially i.e. in 'response' to each other. For present purposes we assume that such stability is satisfied. The two choices are consistent with each other at the intersection of the two best response lines and this is labelled by $\alpha^{*E}, \beta^{*E}$to denote an equilibrium.

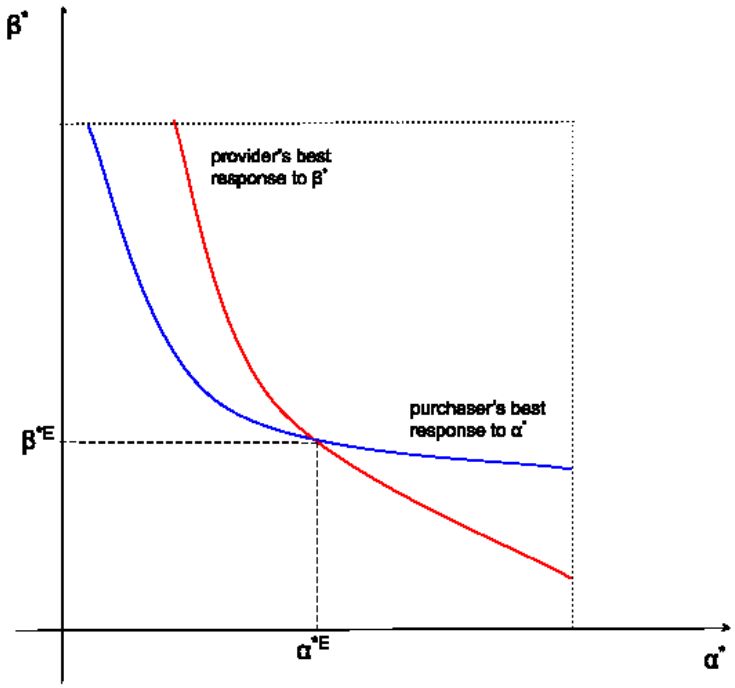


Figure 1: Equilibrium of the provider's and the purchaser's decisions

Any parameter that enters exclusively into the expression in (15) only affects the provider. Hence, a change in that parameter shifts only the provider's best response curve. Figure 2 illustrates the impact on the emergency care system of such changes in a provider's benefit or cost parameters.

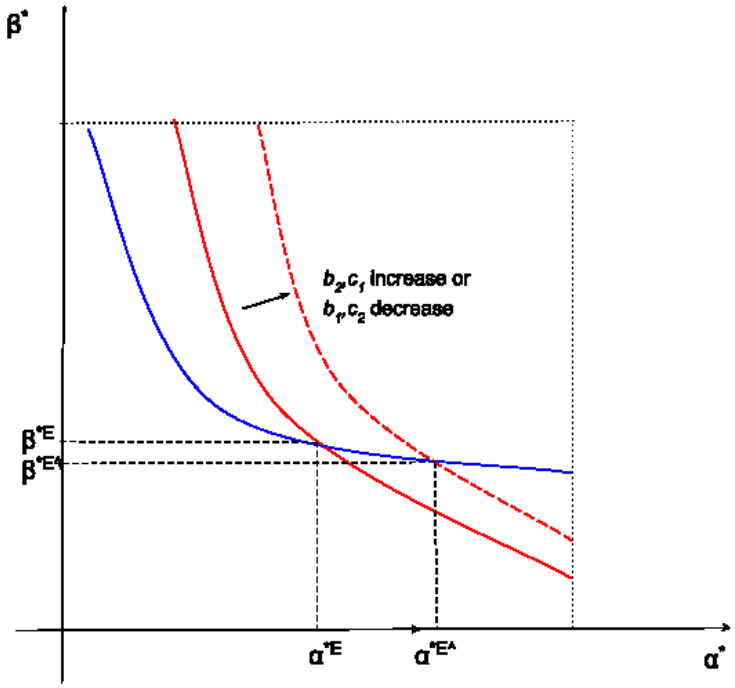


Figure 2: The impact of changes in the provider's circumstances on equilibrium

As illustrated in the figure above, increases in b_2, c_1 or decreases in b_1, c_2 will lead to an increase in α^* and a decrease in β^* in equilibrium.

Purchaser specific effects occur only through the function $\beta(e)$ and factor that causes effort to be more effective at reducing hospital attendances will increase effort e and shift the purchaser's best response curve down and to the left. In equilibrium this will result in lower hospital attendances (β^*) but a higher admission rate from those who attend (α^*).

3. Analysing the role of the payment contract

Achieving different system outcomes

The role of the payment mechanism between purchaser and provider can be understood in terms of comparative statics of equilibrium in respect of p_1, p_2 and T. As conditions (15) and (16') illustrate, the fixed transfer T has no effect. The reason is that it does not enter either the provider's or purchaser's marginal benefit or marginal cost of their respective efforts. As noted earlier T can be viewed as a balancing item by which the purchaser can ensure that the expected cost of delivering emergency care in the hospital is covered, and this will be especially important if less reliance is placed on prices.

Both prices enter into both the purchaser's and provider's first order conditions for optimal choice. Hence, they shift both the purchaser's and provider's best response curves. Since a key motivation for our investigation is a movement away from purely activity-based payment towards a greater component of fixed budgeting, we consider here the effect of decreasing one or both of p_1 and p_2. Both prices work in the same way as far as the purchaser is concerned. A reduction in either price reduces the cost it has to bear of marginal patients attending hospital and, therefore, either price reduction is an incentive to reduce effort e and hence increase β^*. Decreasing either or both prices therefore shifts the purchaser's best response curve up and to the right.

For the provider, as we have previously indicated it is the difference between p_1 and p_2 that is crucial in terms of influencing α^*. The reasoning here is that the provider's choice determines which of these prices is more likely to apply. Hence, if admissions become more valuable relative to treatment in the emergency room (p_2 increases relative to p_1), the provider will wish to increase their α. Hence, an increase in p_2 or a decrease in p_1 (or both together) will shift the provider's best response up and to the right.

If we take the example of a shift towards greater blended payment as being synonymous with a reduction in p_2 (with the associated requirement at Twill need to be increased so as to cover the provider's costs), Figure 3 illustrates the impact on equilibrium. A lower p_2 shifts the provider's best response down and left, and the purchaser's best response up and to the right. The equilibrium with more reliance on the fixed element T and a lower p_2 is therefore characterised by a lower $\alpha^*$and a higher β^*. The emergency care system will thus be more reliant on hospital provision, but that provision will be less focused on admissions.

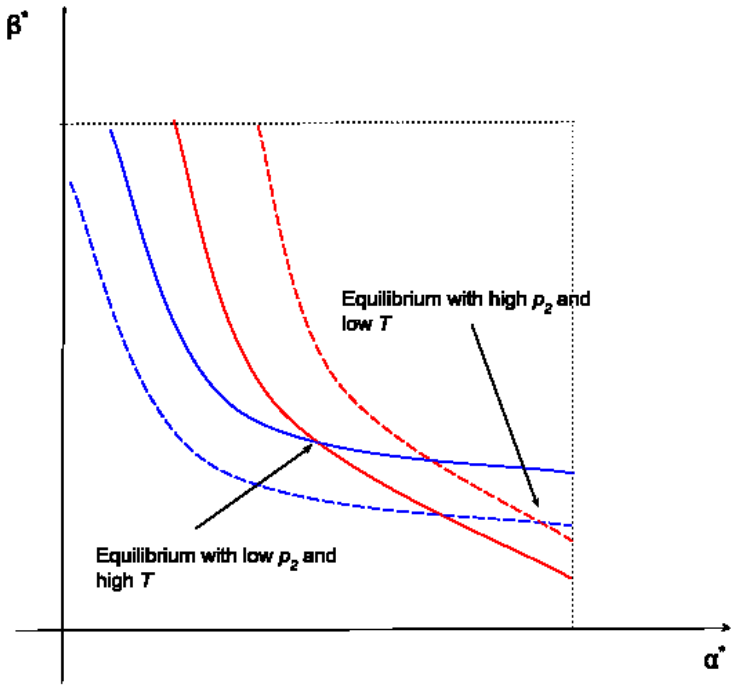


Figure 3: The impact of changing the price of admissions on the provider, the purchaser and equilibrium

Thinking of the configuration of the system as being the combination of $\alpha^*$and β^*, the model indicates that moving towards blended payment by reducing the price of admissions and correspondingly increasing the fixed payment a hospital receives will result in trade-off of lower admission but higher attendances at hospital. The reduced price for an admission under blended payment reduces the incentive for hospitals to admit patients so that there is a reduction in α, for a given β arising from the leftward and downward shift from the hashed red to the solid red best response curve. In isolation, that shift would result in an increase in attendances as the purchaser would be moved along their hashed blue best response. But the purchaser is also affected, since it now pays less for an emergency inpatient, and reduces their effort hence shifting to the solid blue best-response curve. This further reduces admissions but increases attendances. Given the response functions are downward sloping, α and β are substitutes, which exacerbates these effects so that emergency admissions and attendances end up even higher.

In practise, there are limitations on prices and the limits of contract choice might be characterised as a pure fixed price contract (with no fixed transfer) and a pure transfer (with a zero price per admission). In between these limits would define the feasible range of blended payment contracts which in term would define an achievable set of emergency care system configurations. This is illustrated in the green line in Figure 4 below.

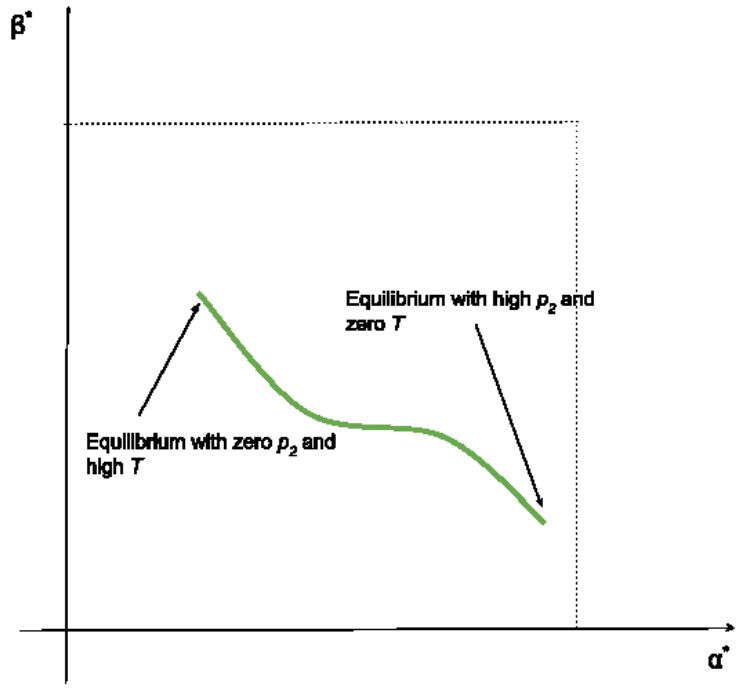


Figure 4: The set of possible equilibria as the price of admissions is varied

Each of the figures above is drawn conditional on the cost and valuation parameters of the purchaser and provider responsible for emergency care in that particular health system. It is useful to also consider the implications of provider and purchaser heterogeneity. By way of illustration, a system in which the provider places a higher intrinsic value on the treatments it gives to admitted patients, relative to the treatments given in the emergency department, will choose higher values of α and will therefore imply a different set of achievable system configurations. This is illustrated in Figure 5 below.

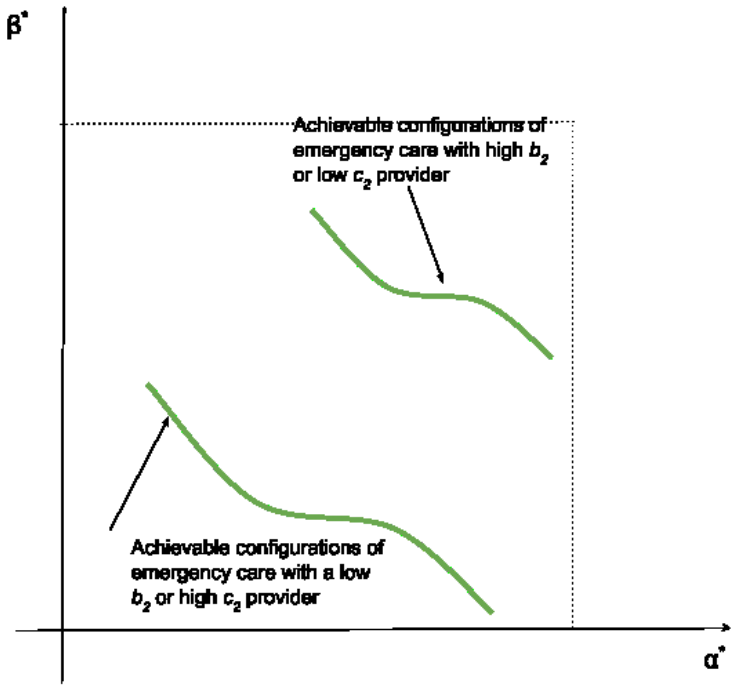


Figure 5: Achievable equilibria for two providers with different benefits or costs of treatment

The analysis above has considered payment reform only in respect of changing p_2. We can broaden the analysis to allow for variation in both p_1 and p_2. Here, the anticipated scenario is that blended payment is synonymous with reducing both prices while increasing the transfer T to ensure that a provider's costs are covered.

This policy implementation would open up many more possibilities in terms of configurations of p_1 and p_2 but there are some natural constraints. From the perspective of theory, negative prices, where the provider remunerates the purchaser for any increase in activity, are perfectly acceptable but are unlikely ever to be operationalised. Hence, we constrain price to be non-negative. A similar issue arises in respect of the transfer T which could conceivably be negative and constitute a payment from the provider to the purchaser. Again, such arrangements are, as far as we are aware, never implemented in practise and have not been suggested in the case of the NHS. A non-negativity constraint on T places an upper bound on p_1 and p_2 since, if these prices are increased without limit, the provider will more than cover costs, and the overall budget constraint would necessitate a negative value for T. Intuitively, the higher one price is set, the lower is the maximum value that can be chosen for the other. There is, therefore, a set of feasible prices which we can expect to have a boundary defined by a negatively sloped relationship between p_1 and p_2. This is illustrated in the shaded area of the figure below. If the practical implementation of blended payment involves setting a premium for an admission, over and above the payment to the hospital for an attendance, then the feasible prices are further restricted to lie above the 45^0 line in the figure below.

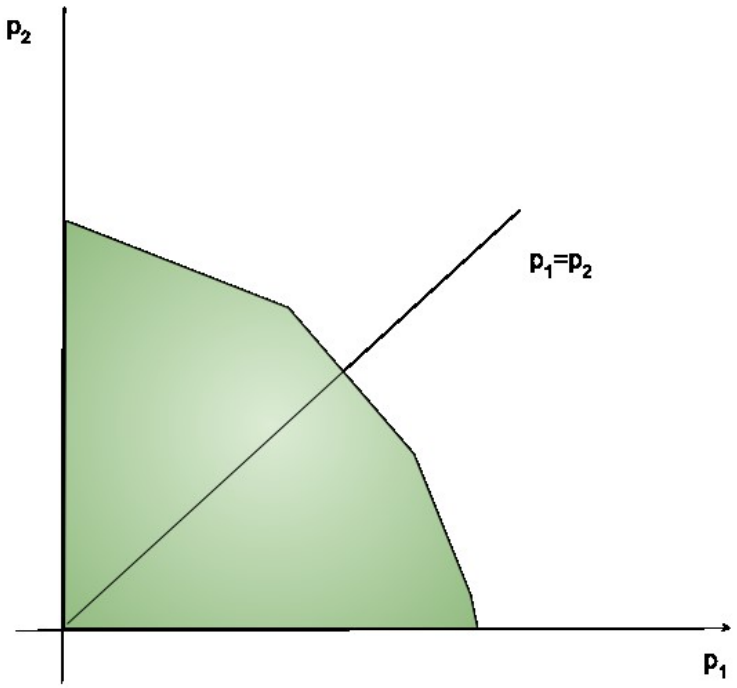
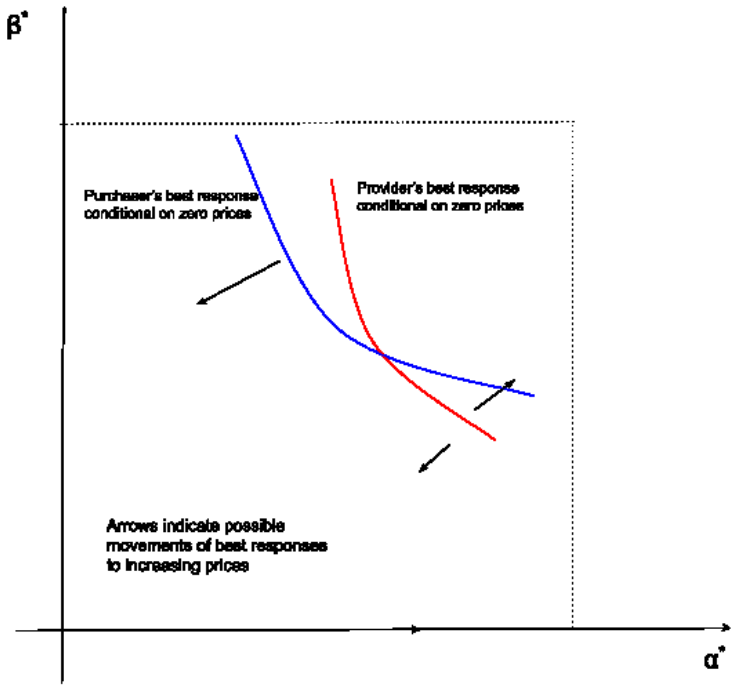


Figure 6: Possible combinations of prices (shaded area) and if the price for an admission is greater than the price of attendance (shaded area above 45^0 line)

The purchaser's and provider's best response curves can be viewed as being shifted according the choice of p_1 and p_2 and, as noted previously increasing prices will shift the purchaser's best response up and to the left, whereas the shift in the provider's best response will be determined according to whether the difference between p_1 and $p_2$increases or decreases. Nevertheless, the compactness of the set of feasible prices places limitations of the achievable configurations of emergency care.

The figure below indicates best responses for the purchaser and provider conditional on both p_1 and $p_2$set to zero whilst the arrows indicate the possible movements of these best responses conditional on increasing prices.



The implication of the limitations on prices and the potential translations of best responses that are then feasible, is that there is a compact subset of configurations of the emergency care system that can be supported by different payments systems (ranging from fully blended payment to simple fixed prices). This is illustrated in the shaded area in the figure below.

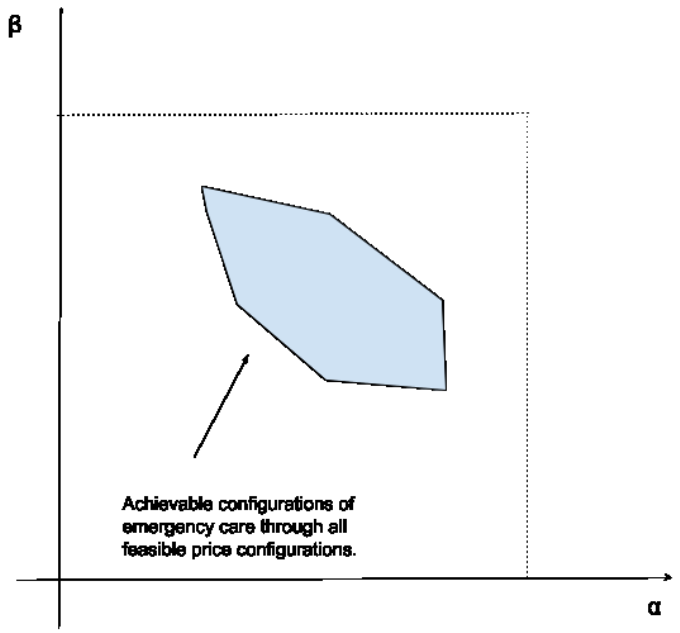


Figure 7: Possible configurations of admissions and attendances from all possible price configurations under blended payment.

The shaded area above shows how the model predicts that there is a set of outcomes that *could* be achieved through different implementations of blended payment systems. We next turn to the question of whether a desired configuration might lie within this set of possibilities.

Emergency care system performance - costs

The question as to what constitutes a well-performing emergency care system is a complex one. In practice it is to be expected that the benefit that patients derive from their treatment will depend on a large number of factors that are not being modelled here. For example, where treatment is delivered, what kind of inpatient and emergency services individuals receive and how timely the treatment is will all contribute to any full assessment of system performance. Our model does, however, provide insight into one aspect of performance - the overall resources used in delivering emergency services.

It is important to distinguish between societal costs and those costs that impact on decisions. According to how it is paid and specifically what price it receives from the purchaser, the provider faces a cost that is the actual marginal cost net of revenue of the treatment it delivers.

Expression (10) gives the expected cost of hospital treatment before any transfer from the purchaser. Using the substitution of μ replaced with $\beta^* n\bar{\rho}$ and σ^2 replaced with $(\beta^* n)^2 {\sigma_\rho}^2$ yields an expression for this expected cost in terms of $\alpha^*$and β^*. Using the inverse function $e(\beta^*)$, the overall cost of the emergency system can then be written as the sum of provider and purchaser resources as in terms of $\alpha$and $\beta$and is $e(\beta) + C(\alpha, \beta)$.

The conditions for $\hat{\alpha}$ and $\hat{\beta}$to minimise the combined cost of treating patients and avoiding attendance are

$$e_\beta(\hat{\beta}) + C_\beta(\hat{\alpha}, \hat{\beta}) = 0 \quad (17)$$

$$C_\alpha(\hat{\alpha}, \hat{\beta}) = 0 \quad (18)$$

Since these conditions do not correspond with the conditions for equilibrium, it follows that the equilibrium resulting from purchaser and provider choices conditioned on an arbitrary form of contract will not result in cost minimisation for the system.

The condition for the purchaser's choice β^*can be rewritten as

$$1 + e_\beta(\beta^*) \frac{1}{[(p_1(1-\alpha^*) + p_2\alpha^*]\, n\bar{\rho}} = 0, \quad (19)$$

whilst written in terms of the function $C(.)$the provider's first order condition for the choice of α^* is

$$C_\alpha(\alpha^*, \beta^*) - (p_1 - p_2 + b_1 - b_2)\beta^* n\bar{\rho} = 0. \quad (20)$$

There is then a question of whether appropriately chosen values of p_1, p_2 can ensure that the pair α^*, β^* satisfying (19) and (20) correspond to the pair $\hat{\alpha}, \hat{\beta}$ satisfying (17) and (18). Conditions for this to be true are set out in the following Proposition.

If there are prices p_1 and p_2 that simultaneously satisfy (i). $p_1 - p_2 - b_2 + b_1 = 0$ and (ii) $[p_1(1 - \hat{\alpha}) + p_2\hat{\alpha}]\, n\bar{\rho} = C_\beta(\hat{\alpha}, \hat{\beta})$ then a blended payment contract based on those prices will achieve an emergency care system that in equilibrium minimises the overall cost of emergency care provision. The demonstration of this comes from substituting prices satisfying (i) into (20) and replacing α^*, β^* with $\hat{\alpha}, \hat{\beta}$ yields (17). Substituting prices satisfying (ii) into (19), replacing α^*, β^* with $\hat{\alpha}, \hat{\beta}$and

multiplying by $C_\beta(\hat{\alpha}, \hat{\beta})$ then yields (17). Hence, with the given prices, the conditions characterising an equilibrium correspond to the conditions for minimising system cost.

However, the existence of such a set of prices is not assured but depends on the values of the various parameters describing the system; the cost minimisation outcome need not lie in the shaded area of Figure 8.

To see why this might be the case, note that condition (i) in the Proposition, together with the assumption that $b_2 > b_1$ implies that $p_1 > p_2$. Hence, a necessary condition for prices to ensure cost minimisation is that the price of an admission is lower than the price paid for treatment in the emergency department. This initially counter intuitive result is a consequence of altruism on the part of the provider. A provider that places more weight on the value of treating admitted patients is inclined to treat more patients in that way than would be consistent with minimising costs. To counteract that, prices must penalise admissions relative to treatments in the emergency department.

Emergency care system performance - social welfare

Overall system cost is only one metric of performance. A conventional conception of social welfare in the context of healthcare would define it as the sum of patient benefits, as perceived by the social welfare maximiser, net of system resources costs, with the latter possibly weighted to reflect the welfare implications of meeting those costs out of public funds (Jones, 2005; Slemrod and Yitzhaki, 1996). It is a matter of some debate as to whether the providers (or in our case both the providers and the purchasers) should be included separately as a reflection of their own utility from delivering care, or whether this constitutes double counting (Culyer, 1989). In the context of the general structure we have modelled however, these nuances are not critical. The key idea is that there exists some preference ordering over outcomes (accounting for both costs and benefits) which can be described by a strictly quasi concave function. Since in our model there are two dimensions to a healthcare configuration, such a function can be denoted $W(\alpha, \beta)$. The function $W(.)$ defines a set of social indifference curves and a potentially interior optimum in respect of α and β, as illustrated in the figure below, where the social welfare optimum is indicated by α^{*W}, β^{*W}.

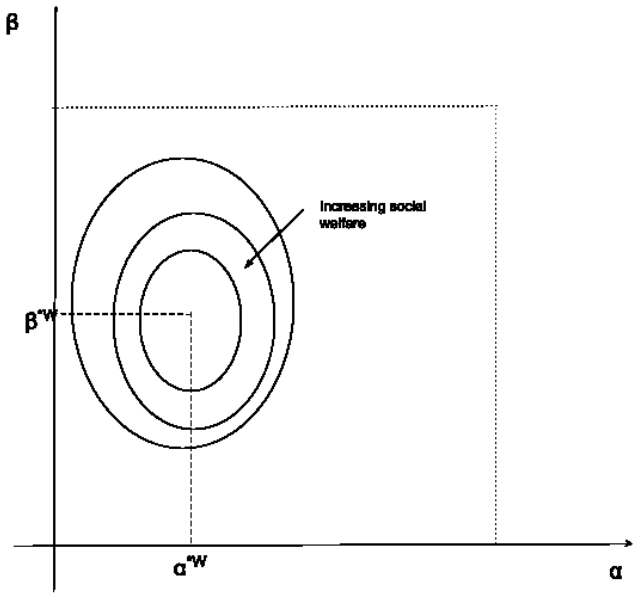


Figure 8: Social indifference curves in admissions and attendances.

The purpose of the section above was in part to show that, even when a desired outcome is anchored in some natural formula arising from the costs of delivering care, there is no guarantee that the outcome can be achieved through an appropriate configuration of blended payment - the necessary prices may simply not be feasible.

It is therefore clear that any more general desired outcome - potentially lying anywhere in the space defined by the unit square of $[\alpha, \beta]$- need not be achievable. Nevertheless, unless the current configuration, with a feasible set of prices, achieves a welfare optimum it will, in most cases, be possible to make a welfare improvement by adjusting prices. The figure below shows this to be the case for all points interior to the feasible configuration set (the shaded area) and gives an example of when it will not be possible - point A where the social welfare indifference curve and boundary of the achievable configuration set share a common slope.

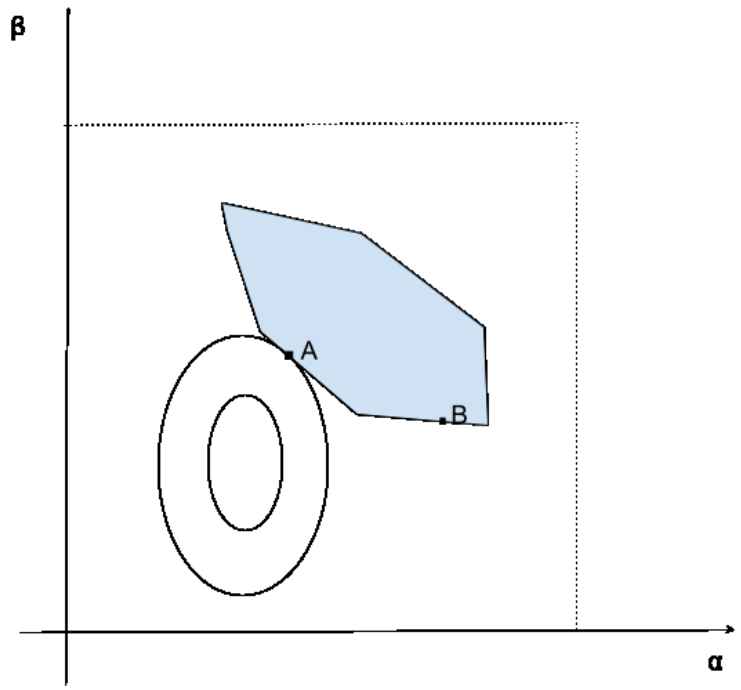


Figure 9: The possibilities for improving social welfare under blended payment

4. Extensions and generalisations

Different payment contracts

The formulation used in the model assumes that the provider is paid for each patient treated in the emergency room, or for each patient admitted. In other words, it is assumed for the purposes of the model that the provider receives one payment per patient but that the price differs according to the patient's treatment pathway. As noted earlier, practical implementation of payment may be made by specifying a price for attendance of p_1 and an additional payment Δ_p for an admission. As the model makes clear it is the magnitude of the difference between p_2 and p_1 - i.e. Δ_p- that is crucial from the perspective of incentives, so our analysis is unaffected by exactly how these prices are described. However, as the model also makes clear there is a benefit to considering the possibility that Δ_p is negative, and the practical implementation may preclude that. In this case, the model usefully draws attention to a constraint that may arise out of a convention (to pay a premium for an admission) that may make influence over the emergency healthcare system more difficult.

The model also considers only linear pricing rules, where each additional attendance or subsequent admission is paid a fixed price. A variation on this that has been discussed is a piecewise linear system in which, above or below certain thresholds, different prices apply. Accounting for these systems is simply a matter of replacing the expression $T + p_1 N + p_2.\alpha N$ with a formula that takes account of different prices applying at different thresholds. For example, if the price paid for an admission change from p_2^+ to p_2^- at a threshold of $\bar{N}$ then the provider's expected revenue is given by $T + p_1\,\mu(1-\alpha) + p_2^+.\alpha \int_0^{\bar{N}} N\,dF(N) + p_2^-.\alpha \int_{\bar{N}}^{\infty} N\,dF(N)$. Since this expression remains linear in α the analysis is unaffected except through the change in the expressions involving prices in the first order condition and subsequent solution for α.

Extended provider choices

The focus for our analysis is the proportion of patients that a provider chooses to admit for further treatment. In many analyses of fixed price payment contracts of the type we have considered, the concern is with other elements of provider choice impacting on cost-reducing effort and quality of care.

In respect of the former, the intuition from those other models carries over to the present setting. A fixed-price system makes the provider a residual claimant in respect of any cost savings it makes (Ma, 1994). Hence, it will choose those cost savings in a manner that will be conditionally efficient - that is for any given values of other choices cost-reducing effort will satisfy the requirements for minimising overall social costs.

In respect of quality, an issue that arises is whether a fixed price system induces skimping on quality of service since, relative to being reimbursed directly for the costs of the treatments it provides, a fixed price system does not compensate the provider for additional quality. Our model is based upon the assumption that a provider cares for its patient, i.e. it exhibits altruism so that it will retain an intrinsic motivation to provide quality of care, albeit that such a motivation may not align with a societal perspective.

Risk sharing

The primary purpose of our framework is to understand the trade-offs that emerge under different contracts between purchaser and provider. We have assumed that payments will be made so as to cover the mathematical expectation of the costs incurred by the provider in delivering treatments. Depending on the realisation of demand, the provider may incur a loss or surplus, whilst the

purchaser's expenditure will be variable. The extent of these respective variances also depends on the specification of the payment contract. This is a further consideration in the choice of contract. There are two reasons for separating risk sharing concerns and considering them separately.

First, there is no universally accepted approach to evaluating the impact of risk on decisions. The most utilised approach in economics is expected utility theory. That suggests that the purchaser's and provider's evaluation of risk is captured through them respectively maximising the expected utility of their objective functions. Such an approach raises a question as to what might be a reasonable specification of utility and why large organisations, such as hospitals and CCGs, are unable to accommodate risk or place a premium on its avoidance.

Perhaps more compelling, however, is the observation that the allocation of risk between purchaser and provider is secondary to the appropriate functioning of the healthcare system because risk allocation is a zero sum procedure - the risk avoided by the purchaser is a risk faced by the provider. Consider for example the goal of minimising societal costs that we have examined above, and suppose there exists a set of prices that achieve this goal. Those prices establish the extent of variation in the providers surplus or loss and the purchaser's expenditure in a straightforward manner. Since the source of risk in our model is variation in the number of patients requiring treatment, higher prices that link payment more to variation in activity imply less risk for the provider and more risk for the purchaser. If that distribution of risk is judged to be unacceptable, the model suggests both how a more acceptable distribution of risk can be achieved through lower prices and what the true cost of doing that is in terms of increased societal cost. This suggests that a formal treatment of risk allocation can be achieved by appropriately modifying the system goal to include a measure of the distribution of risk between the purchaser and provider.

Provider effort in reducing admissions

In the interests of parsimony, the model has not included explicit consideration of the measures that the provider might take to avoid admission, nor the costs of those measures. The conventional terminology to apply to these measures is *effort,* and the costs that the provider incurs would be analogous to the cost incurred by the purchaser in avoiding attendances. We could add a cost of effort to the expression for cost (4) which would take the form $E(\alpha)$, where $E(.)$ would be a decreasing function and capture the fact that avoiding more admissions (lower α)would imply a higher cost to the provider. Technically, such a function may be necessary to ensure the concavity of the provider's objective in respect of α but, in our chosen formulation with a convex treatment costs increasing in α, this is not required. Hence, adding this cost would serve only to complicate the algebra and preclude an explicit expression for α^* without adding any new element to our model.

5. Discussion and policy implications

The policy reform entailed in blended payment contracts for emergency care breaks the link between prices and costs. The lower price is compensated for with a fixed transfer, but the two elements - price and transfer - can be subject to discretion and negotiation between the purchaser and provider. This form of two-part tariff has long been considered a possibility in the theoretical economics literature, but its adoption in the English NHS for emergency services is both novel and challenging. The challenge arises from a need to understand how the newly empowered discretion can or should be used to influence the delivery of emergency care. With the new-found discretion over pricing comes a responsibility to ensure that it is adopted appropriately.

We have approached this topic by constructing a simple economic model that captures what seem to be important but hitherto neglected aspects of the real-world context of emergency care; in this setting *both* the purchaser *and* the provider make choices that impact on the performance of an emergency care system. Both of those decisions are likely to be affected by the precise payment terms chosen, and both also interact with each other -- what a purchaser finds it desirable to do in terms of arranging care outside of a hospital setting depends on how care is organised in that setting and vice versa.

Our first key insight from formally modelling this setting is that incentives can be a double-edged sword. One rationale for reducing the price of hospital admissions is to weaken incentives to admit patients to costly care. In our framework this also weakens the incentive to treat patients in the community and prevent them going to hospital. In the equilibrium of the model, there emerges a trade-off between payment that results in high admissions but low attendance, and payment that results in the opposite (low admissions but high attendance). As in so many settings, the model provides an articulation of the caution required in adopting reform in order to avoid unintended consequences.

The model goes much further in establishing more generally what the influences on the combination of equilibrium attendances and admissions are likely to be. Other than the already highlighted role of the payment and subsequent behavioural responses, these other influences are the nature of the conditions being contracted for and the values and costs that the agencies (both purchaser and provider) place upon different treatments. This provides a way of understanding the likely drivers of observed differences between systems, and could be extended to generate a rich set of testable hypotheses relating observed system differences in attendance and admission to potentially observable features of the patients and agencies in the system.

However, our primary focus has been to explore how the adoption of blended payment can potentially improve system performance. A general point that emerges is the recognition that the more instruments or levers of policy there are, the better the potential for improvement. This manifests in the setting of blended payment for emergency care in the recognition that allowing discretion over both the price paid for admitted patients, and the price paid for patients who simply attend the emergency department (but are not admitted), is better than focusing only on admissions. The less constrained are prices the better in this context, and the theory indicates there are circumstances in which the price for admissions could usefully be set lower than the price for attendance - even though costs of care lie in the opposite relativity. The intuition here is that a high attendance price provides a strong incentive for the purchaser to ensure out-of-hospital care, while a low admission price discourages the provider from admitting too many patients.

Quite generally it would seem that blended payment has the *potential* to improve system performance, and that is true almost regardless of what metric of performance is considered. But

here we encounter an issue since there is not an obvious consensus as to what performance metric ought to be pursued. Our analysis suggests that there are many possibilities - minimising the cost of emergency care or maximising one (of many possible) social welfare formulations. There is a further note of caution: whilst generally a move to blended payment may improve performance against these many goals, it is certainly not ensured that it will permit an 'ideal' configuration to emerge as an equilibrium. The reason is that, even with as many prices made discretionary as possible there are still some natural constraints that limit the outcomes that can be achieved - for example the requirements for prices and transfers to be non-negative.

Our model was motivated by the adoption of blended payment for emergency care in England. As with so many aspects of healthcare systems, policy has evolved in the light of the COVID-19 pandemic. Some of the institutional structures described in this paper are due to be abolished, notably CCGs. However, the fundamental reform that is embodied in blended payment -- less reliance on activity-based payments -- will continue to be a guiding principle of financial transfers within the reformed system.[4] Hence, the analysis that we have presented remains relevant.

Whilst the model we have constructed is simple, its most important predictions and insights appear to be quite robust to generalisation. However, it is a theoretical model and simply because a model predicts that behaviours change in response to changing prices does not guarantee that this will be observed in practise. There is a further possibility that changes do emerge but they are so slight as to be of no practical relevance. This provides the final value of our model. It serves as a framework for guiding empirical investigations of these important practical issues. Those empirical investigations are the focus of our subsequent papers under this project.

4 See https://www.lgcplus.com/services/health-and-care/nhs-england-recommends-law-to-abolish-ccgs-by-2022-27-11-2020/

References

Chalkley, M., Malcomson, J., 1998a. Contracting for Health Services with Unmonitored Quality. *Econ. J.* 108, 1093–1110.

Chalkley, M., Malcomson, J.M., 1998b. Contracting for health services when patient demand does not reflect quality. *J. Health Econ*. 17, 1–19. https://doi.org/10.1016/S0167-6296(97)00019-2.

Culyer, A.J., 1989. The normative economics of health care finance and provision. Oxf. Rev. Econ. Policy 5, 34–58.

Eggleston, K., 2009. Provider payment incentives: international comparisons. *Int. J. Health Care Finance Econ*. 9, 113. https://doi.org/10.1007/s10754-009-9065-3.

Fainman, E.Z., Kucukyazici, B., 2020. Design of financial incentives and payment schemes in healthcare systems: A review. *Socioecon. Plann*. Sci. 72, 100901. https://doi.org/10.1016/j.seps.2020.100901

Guterman, S., Dobson, A., 1986. Impact of the Medicare prospective payment system for hospitals. *Health Care Financ. Rev*. 7, 97–114.

Jack, W., 2005. Purchasing health care services from providers with unknown altruism. *J. Health Econ*. 24, 73–93.

Jones, C., 2005. Why the marginal social cost of funds is not the shadow value of government revenue.

Kaarboe, O., Siciliani, L., 2011. Multi-tasking, quality and pay for performance. *Health Econ*. 20, 225–238. https://doi.org/10.1002/hec.1582.

Laffont, J.-J., Martimort, D., 2001. The Theory of Incentives: The Principal-Agent Model. Princeton University Press.

Ma, C.A., 1998. Cost and Quality Incentives in Health Care: A Reply. *J. Econ. Manag. Strategy* 7, 139–142.

Ma, C.A., 1997. Cost and Quality Incentives in Health Care: Altruistic Providers.

Ma, C.A., 1994. Health Care Payment Systems: Cost and Quality Incentives. *J. Econ. Manag. Strategy* 3, 93–112.

Ma, C.A., Mak, H.Y., 2019. Incentives in Healthcare Payment Systems. *Oxf. Res. Encycl. Econ. Finance*. https://doi.org/10.1093/acrefore/9780190625979.013.61.

Mathauer, I., Wittenbecher, F., Organization, W.H., 2012. DRG-based payments systems in low-and middle-income countries: Implementation experiences and challenges. World Health Organization.

Pauline, A., Kath, C., 2020. Commissioning Healthcare in England: Evidence, Policy and Practice. Policy Press.

Reinhard, B., Alexander, G., Wilm, Q., 2011. Diagnosis-Related Groups In Europe: Moving Towards Transparency, Efficiency And Quality In Hospitals. McGraw-Hill Education (UK).

Slemrod, J., Yitzhaki, S., 1996. The costs of taxation and the marginal efficiency cost of funds. Staff Pap. 43, 172–198.