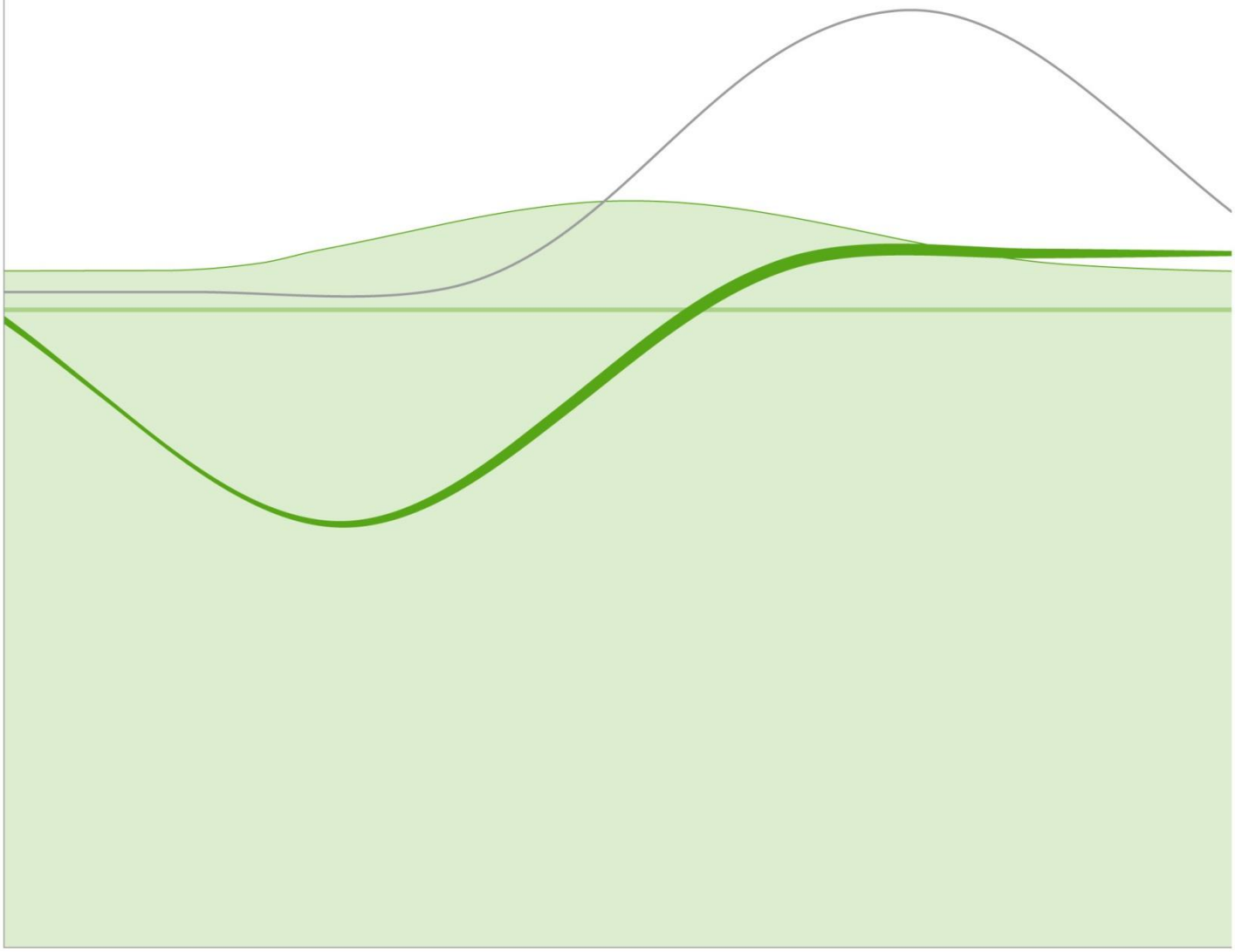
Office of Health Economics Research

Research Paper 15/03

Model of Behaviour within Fuzzy Budget Constraints

December 2015

Jon Sussex and Karla Hernandez-Villafuerte



Model of Behaviour within Fuzzy Budget Constraints

Jon Sussex[1] and Karla Hernandez-Villafuerte[2]

[1]RAND Europe [2]Office of Health Economics

December 2015

Research Paper 15/03

For further information please contact:

Jon Sussex

jsussex@rand.org

RAND Europe

Westbrook Centre, Milton Road

Cambridge CB4 1YG

United Kingdom

Tel: +44(0) 223 353 329

About OHE Research Papers

OHE Research Papers are intended to provide information on and encourage discussion about a topic in advance of formal publication.

Any views expressed are those of the authors and do not necessarily reflect the views or approval of OHE, its Editorial or Policy Board, or its sponsors.

Once a version of the Research Paper's content is published in a peer reviewed journal, that supersedes the Research Paper and readers are invited to cite the published version in preference to the original version.

Abstract

A major preoccupation in health policy is the displacement of existing health care services when new technologies are mandated. This opportunity cost is viewed as an inevitable consequence of the fixed budget constraints applying to the health service nationally and locally. However, local health care purchasers have some limited scope to overspend or underspend in relation to their annual budgets, and they may have some scope to increase efficiency. Our aim is to present a model that formally addresses all of the options that decision makers may have in addition to the displacement of existing health care services when faced with a mandate to pay for a new health technology. Based on stochastic frontier analysis (SFA), two concepts are included in the discussion: (1) minimum cost and efficiency; (2) the effect of the limited government tolerance of reduced underspends and higher overspends relative to budget. Based on the interaction of these concepts and the SFA, a new definition is developed: the margin of tolerance. This is an approximation to the willingness of local health care purchaser decision makers to overspend their budgets.

Keywords: new health technologies approval, budget constraint, opportunity cost, efficiency, overspend, underspend.

1. Introduction and policy context

In this paper we set out a model to capture the behaviour of the health sector decision makers (HSDMs) who must decide on allocation of resources and the level and mix of production while taking into consideration the maximum level of expenditure (budget) allowed by the government or the health system regulator. Development of the model is prompted by two perennial policy and practical discussions in the UK National Health Service (NHS) (and elsewhere): how to cope with financial austerity; and what are the opportunity costs of including new health technologies in the list of those the NHS is required to offer.

We observe that in practice HSDMs have three broad options when faced with demands to fund new health care technologies/services: (1) displace other services, (2) increase efficiency and (3) obtain increased funding. If local health care purchasers have fixed budgets and if they cannot squeeze increased efficiency out of health care providers, then the cost of providing a new medicine can only displace other health services they reimburse. But in practice average efficiency of providers is less than 100% (Bojke et al., 2013). Moreover, budgets are not entirely fixed: an underspender might choose to underspend less; an overspender might choose to overspend more.

Where the opportunity cost falls depends on the balance struck between the three options. With displacement of other services the opportunity cost is the forgone benefit of those services. Where efficiency can be increased, the opportunity cost is the loss of utility to staff being pressured into an increase in intensity of effort (Cyert and March, 1992). Where the HSDMs simply spend more, the opportunity cost falls outside the NHS. There is then a displacement of spending on some other part of the public sector and/or a marginal increase in future or current taxation to cover that spending.

Economists routinely accept the couching of economic problems in terms of making best use of a pre-determined endowment of resources: the fixed budget constraint. In the public sector, organisations given citizens' money to spend on providing public services are typically set fixed annual expenditure limits – fixed or 'hard' budgets – by central government. In high income countries health services are to a large extent financed centrally, from taxation or insurance, but the quantity and mix of services to provide is decided locally.

Recent empirical studies of the NHS in different parts of the UK have highlighted that even where budgets are presented by central government as 'hard', local commissioners of health care services may not treat them as such when faced with the problem of finding resources for a new health technology or service that has just been added to the basket of services that are required to be offered by the NHS (Karlsberg Schaffer et al., 2014; 2015).

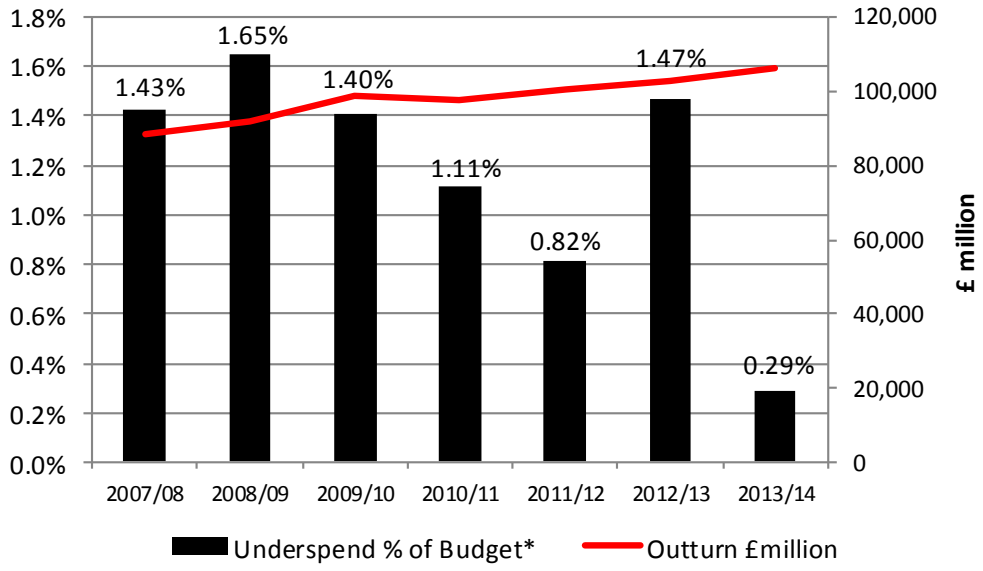
Since April 1999, 100% of NHS expenditure in England has been required by central government to stay within a pre-determined "cash limit" (Sussex, 1998). This requirement has been passed on to the sub-national organisations responsible for spending the total NHS budget. But do public sector organisations really behave as if their budgets are wholly fixed, or is there some, limited, flexibility? Are budget constraints fuzzy rather than being treated as fixed?

Newspaper headlines about public bodies overspending their budgets, and struggling to eliminate or reduce those overspends, are common. Contrastingly, in total the NHS in England has in most years underspent the funds awarded to it by Parliament. There are

both overspending NHS organisations and underspenders. In recent years there have been more of the latter. In future years with continued fiscal austerity there are likely to be increasing numbers of overspenders. Either way, failure to spend budgets precisely 100% and no more is to be expected, given the impossibility of spending a multi-million pound annual budget in full, but without overspending. There is inevitable uncertainty, not least in the amount of patient care that will be demanded, but also in the costs of delivering that care. Consequently at the start of each financial year, commissioners usually aim to underspend, so as to leave a contingency to pay for unexpected cost pressures that arise during the year.

It is perhaps surprising how close to 100% of the budget is spent every year by NHS organisations in England. The Audit Commission reported in 2012 that local NHS organisations in England accumulated a combined surplus of over £2 billion, equal to more than 2% of the total NHS budget for financial year 2011/12. Figure 1 shows how NHS spending in England has compared with the annual budget. Behind those average figures are much wider variations in percentage underspends, and overspends, against budget by individual local NHS organisations (Audit Commission, 2012).

Figure 1. Total NHS operating expenditure in England



* Underspend % is based on the difference between Revenue Department Expenditure Limit budget and spending outturn.
Sources: Department of Health, 2008, 2009, 2010 and 2014.

Over the course of a budget period an NHS organisation with a notionally fixed budget is likely to start the year aiming to undershoot that budget by a small amount. In other words it may, in effect, put aside a contingency fund. It will take a view on what demand it is likely to face over the coming year and what it will cost to meet that demand and hence what, if any, actions might be needed to improve efficiency or change the scope of services it offers or the range of patients it offers services to, in order to stay within that budget. Moreover, the amount of the contingency fund will depend on the perception of the HSDMs relating to possible negative consequences of going above the budget (i.e. possible sanctions by the government). As the year progresses, the organisation will monitor how far its expenditures are above or below its original expectations. Expenditure plans over the rest of the year can then be adjusted: increasing

discretionary expenditure if there have been underspends hitherto, or trying to cut back expenditure if it is running ahead of plan.

Our aim is to present a model that formally addresses the different possibilities that the HSDMs have in addition to the displacement of health care services in response to the mandating of a new health technology. The model will reflect the three factors that the HSDMs consider during the decision process: 1) budget, 2) efficiency and 3) health service demand. We will introduce a concept that partially explains the variation in HSDM responses: *the margin of tolerance*.

2. The basic model

For simplicity, we assume that the production of health care services for each of the local health commissioners, from here on referred to as "locations", can be classified into k different kinds of health services. Hence, q_{ik}^t corresponds to the health care output for health location i $(i \in 1,2,\dots,n)$ in year t in the production of the health service k $(k \in 1,2,\dots,K)$. Moreover, q_i^t represents the total output of health services $(\sum_{k=1}^{K} q_{ik}^t)$ produced by location i during year t.

The output level depends directly on the health care demand of location i; therefore, during the year the HSDM can have only an expectation of the values of q_i^t at the end of the year t. These expectations depend on the information available to the decision makers, e.g. trends in the numbers of medical consultations, surgeries and medicines used. The expectations can vary during the year since the information available to the HSDM also changes over time. For simplicity, we assume that the HSDM adjusts its expectation of q_i^t at the end of each month. Then, the expectation of the HSDMs of location i at the end of month j of what could be the total value of the output at the end of the year t is called $E_j(q_i^t)$.

C_i^t represents the amount of expenditure that health location i makes during the year t to produce q_i^t . Since there exists a series of output expectations $E_j(q_i^t)$, the health location i also has at month j an expectation of the present year expenditures: $E_j(C_i^t)$.

As the basis for the model we have adapted a stochastic frontier analysis (SFA) methodology. Special attention has been paid elsewhere to analysis of C_i^t through the estimation of cost frontiers using SFA (Førsund and Jansen, 1977). SFA is ideal for our purposes as it allows analysis of two of the three factors that determine HSDMs' decisions: efficiency level and health service demand reflected in the minimum level of expenditure needed. SFA suggests that the expenditures can be divided into three components, as follows:

$$C_i^t = C_i^{Mt}(P_l^t, q_i^t) + v_i^t + u_i^t \qquad (1)$$

where the first component, $C_i^{Mt}(P_l^t, q_i^t)$, corresponds to a cost function, which describes the minimum cost of producing output q_i^t when firm i (in our case the health location i) faces input prices P_l^t $(l \in 1,2,\dots,m)$ and has the same functional form for every i. In our case $C_i^{Mt}(.)$ corresponds to the minimum expenditure needed to satisfy the demand. The second element (v_i^t) is a systematic term which reflects inefficiency in production. The last term (u_i^t) is a standard normally distributed random noise.

It is important to highlight that, for the purpose of this model, efficiency is what in economics is termed *productive efficiency* or *technical efficiency*. When the health location is producing a given level of q_i^t with the least-cost methods of production available it is possible to say that the location has reach 'productive efficiency' (Griffiths

and Wall, 2004). Then v_i^t is a measure of the difference between the minimum cost achievable and the point of productive efficiency. Moreover, other factors, apart from inefficiency, can temporarily separate the cost from its minimum value; these factors are expressed in equation (1) as u_i^t. Even though our intention is to represent behaviour within a year, it is worth noting that the level of maximum productive efficiency can change through time. This could happen if lower cost technologies of production are developed. A consequence is that as future time periods unfold, the level of inefficiency (v_i^t) could rise even if the method of production in a location did not change.

Even though SFA can be used to examine allocative efficiency, our model does not do so as it is assumed that the production of each k health service is driven by the demand. Additionally, our objective is to analyse the variation in the health location decisions given the three mentioned factors (efficiency, budget and demand); therefore, the opportunity costs outside the health sector are not captured in this model. Given that $C_i^{Mt}(.)$ depends on q_i^t, it is also possible to estimate a minimum cost based on the output expectations at month j of year t: $E_j(C_i^{Mt}\left(P_l^t, E_j(q_i^t)\right))$. For the purposes of this model, it is necessary to establish some assumptions relating to equation (1):

> ***Assumption 1***: The functional form of $C_i^{Mt}(.)$ does not change during the year, meaning that $C_i^{Mt}(.)$ has the same distribution as $E_j(C_i^{Mt}(.))$ for every j.
>
> ***Assumption 2*:** The status quo is always preferred (Samuelson and Zeckhauser, 1988). This means that even if a health location is inefficient ($v_i^t > 0$), the location will look for an efficiency improvement only when forced to do so.

Based on Assumption 1 and equation (1), we can define:

$$E_j(C_i^t) = E_j(C_i^{Mt}\left(P_l^t, E_j(q_i^t)\right)) + E(v_i^t) + E(u_i^t) \quad (2)$$

If no unexpected change has occurred since the end of the last year $(t-1)$ until period j that forces location i to change the status quo, Assumption 2 implies that: $E(v_i^t) = v_i^{t-1}$. As u_i^t is a standard normally distributed random noise, its expected value is equal to zero.

$$E_j(C_i^t) = E_j(C_i^{Mt}\left(P_l^t, E_j(q_i^t)\right)) + v_i^{t-1} \quad (3)$$

At this point we can introduce the budget constraint. Location i faces a budget constraint equal to B_i^t which is the maximum amount of expenditure that the national government allows location i to have during the present year. As mentioned above, in reality there are deficit and surplus health locations, meaning that C_i^t could be greater than, less than or equal to B_i^t. However, government tolerance of deficit is not unlimited. The decision makers know that when the deficit is *too high* they will suffer penalties, which could include losing their jobs. Nevertheless, the exact meaning of "*too high*" is unknown by the health locations. Based on this, we can establish a third assumption:

> ***Assumption 3:*** There is a δ_g^t such that when $C_i^t > B_i^t \delta_g^t$ the health location decision maker will suffer a penalty, where the value of δ_g^t is unknown.

It is reasonable to assume that δ_g^t is greater than 1, i.e. the government would accept a small deficit. Otherwise, if δ_g^t was less than 1, the government would be requiring savings relative to budget from the locations.

As noted above, C_i^t can be above, below or equal to B_i^t. This means that there exists a θ_i^t such that:

$$C_i^t = B_i^t \theta_i^t \quad (4)$$

where: $\theta_i^t > 0$ and $\theta_i^t \leq \delta_i^t$

$\theta_i^t < 1$ means that there is a surplus and $\theta_i^t > 1$ indicates a deficit. δ_i^t is the upper limit of θ_i^t and represent the maximum willingness to reach a deficit (overspend) of health location i. δ_i^t reflects two characteristics: 1) the estimation that the decision maker of health location i has about the value of δ_g^t and 2) how risk averse they are. The smaller δ_i^t, the more risk averse the local decision maker is.

Only the decision maker knows the precise value of δ_i^t. Nevertheless, δ_i^t can be approximated by observing θ_i^t. Even though the value of θ_i^t can be in a range between 0 and δ_i^t, it is realistic to accept that the actual relationship between expenditure and budget, reflected in θ_i^t, will be as close as possible to the decision makers' willingness to reach a deficit. We name the approximation of this willingness to reach a deficit, the θ_i^t, the *margin of tolerance*. This can be easily estimated as:

$$\theta_i^t = C_i^t / B_i^t \quad (5)$$

The value of θ_i^t will be known only at the end of the year. But once again there exists an expected value at month j of year t:

$$E_j(\theta_i^t) = E_j(C_i^t(.))/B_i^t \quad (6)$$

In order to analyse the reaction of the HSDM an unexpected shock is introduced. We assume that a new medicine has been approved and that from month $j+1$ the health location i must cover the cost of providing patients with this medicine.

The first effect is an increase in the total quantity of expected demand, and therefore in the production of the related health services, such that $E_j(q_i^t) < E_{j+1}(q_i^t)$. And so, $E_j(C_i^{Mt}\left(P_l^t, E_j(q_i^t)\right)) < E_{j+1}(C_i^{Mt}\left(P_l^t, E_{j+1}(q_i^t)\right))$ because the distributional function of the minimum cost does not change. Finally, from equation (6) we also know that $E_j(\theta_i^t) < E_{j+1}(\theta_i^t)$. A last assumption is required:

> ***Assumption 4:*** The least desirable option is to produce at a level where some demand cannot be satisfied.

For simplicity, we assume that this is the only unexpected shock during the year. However, this assumption is easy to relax by introducing sums to the effects. There are two possibilities:

1. $E_{j+1}(\theta_i^t) < \delta_i^t$

   The HSDM believes that the location's total expenditure is in their comfort zone. Here, without adjustments, the location can meet the increased demand with a cost of production below the cost at which they expect there would be a penalty.

2. $E_{j+1}(\theta_i^t) \geq \delta_i^t$

   The HSDM believes that the location's total expenditure is outside their comfort zone. The HSDM then faces three options which can be adopted separately or together in any combination:

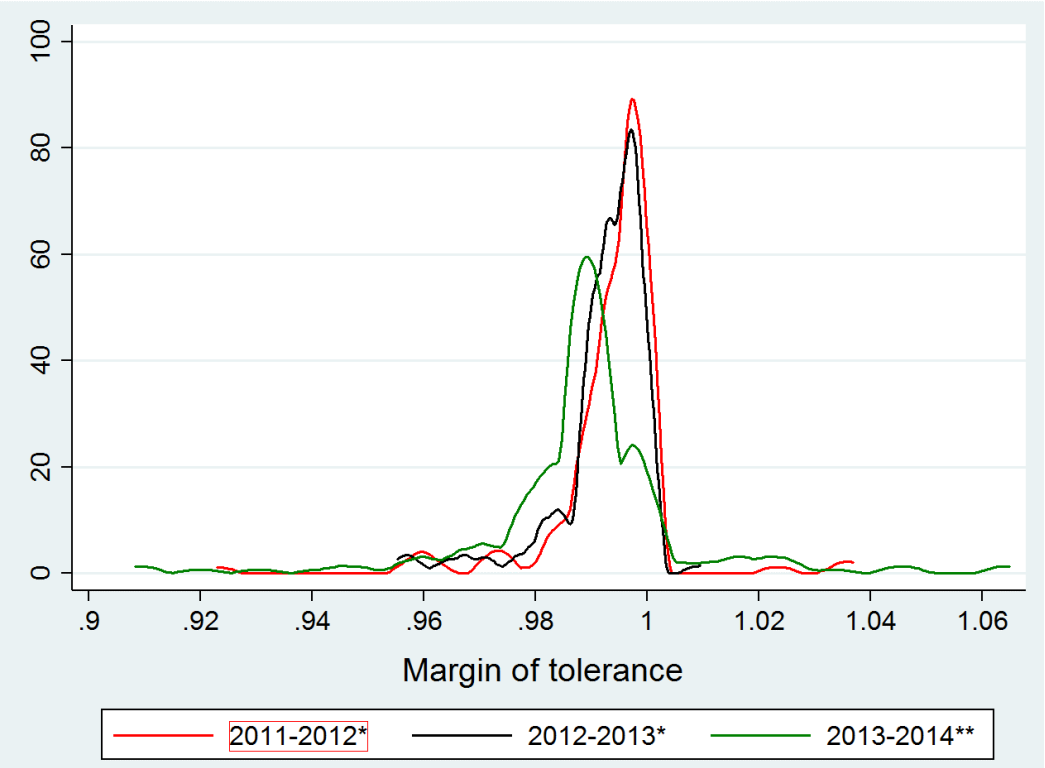
   - Improve efficiency. This will be reflected in a change in the inefficiency coefficient: $v_i^t - v_i^{t-1} < 0$.

- Increase expenditure by obtaining increased funding – a 'bail-out' – from the government. In this case the opportunity cost falls on other sectors outside the health system. This is the only case in which $E_{j+1}(\theta_i^t)$ could exceed δ_i^t.
- Reduce production so that $E_{j+1}(\theta_i^t) \leq \delta_i^t$. The least preferred option of location i is the non-satisfaction of demand. However, when the capacity of HSDMs is not enough to lead the location into the comfort zone, it is necessary to decrease production. It is in this situation that the mandating of a new medicine has an opportunity cost within the health service in terms of the reduction (displacement) of other health services.

3. Analysing the margin of tolerance

The model suggests hypotheses to test in future research. For example if stronger penalties are announced, this would be expected to reduce θ_i^t. Also, other things being equal, we could expect steadier long term behaviour of θ_i^t for those health locations with unchanged management compared to those with changed HSDMs. Based on time-series data of locations' expenditure relative to budget, we can estimate θ_i^t, based on equation (5). Figure 2 shows the distributions of the margin of tolerance for local health care purchasers in the NHS in England in the three financial years, 2011/12-2013/14. Even though a more robust analysis of the margin of tolerance would require data from a longer period (which are not readily available), Figure 2 indicates that the distribution of the margin of tolerance was very similar in 2011/12 and 2012/13. The period 2013/14 shows a slightly different figure, which may be explained by the major restructuring of the NHS in England in that year.

Figure 2. Distribution of the margin of tolerance in the NHS in England 2011/12 to 2013/14

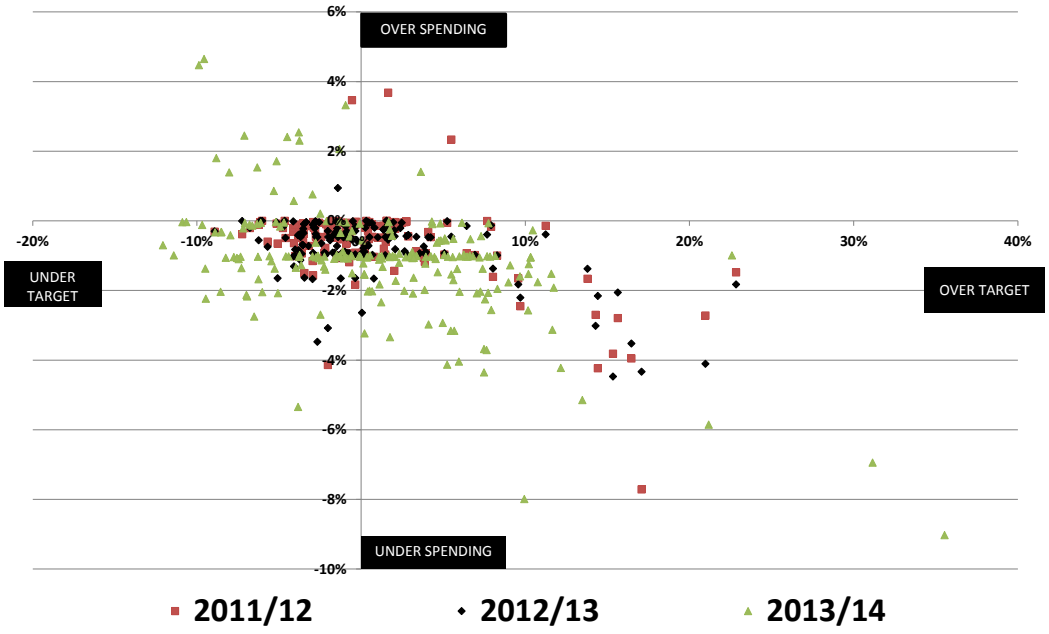


* 152 Primary Care Trusts (PCTs)
** 211 Clinical Commissioning Groups (CCGs)
Source: Authors' calculations based on published funding allocations (budgets) and outturn expenditures for local NHS health care purchasing organisations in England

Additionally, the NHS classifies the health location according to the budget allocated. Those locations in which the budget allocated is higher than the predicted budget required are grouped as 'over target' locations. We expect that over target locations have a lower δ_i^t since they presume that a lower δ_g^t applies in their case. Therefore, these locations will have a higher tendency to underspend. Figure 3 shows a negative relationship between the true spending level and the target level.

A second group of hypotheses relating to the efficiency level also emerges. When HSDMs announce, in response to a mandate to pay for a new health care technology, that they are increasing efficiency and not displacing any existing services, the outcome may nevertheless be some unheralded displacement of services. A hypothesis that might be tested is that the most inefficient health locations would have the lowest probability of displacing other health services as a response to the mandated inclusion of a new health technology.

Figure 3. Level of underspending or overspending according to distance from the opening target allocation by NHS health location



* 152 PCTs
** 211 CCGs
Given the availability of the target level information, expenditure from 2011/12 and 2012/13 is compared to target levels for 2011/12, and expenditure of 2013/14 is compared to target levels for 2014/15

4. Conclusion

The main objective of this paper is to describe a simple model that reflects the observed practice of HSDMs. Displacement of existing services is not the only option when HSDMs respond to a mandate to fund a new health care technology, despite the ostensibly fixed nature of their budgets. One option for a location facing an unexpected increase in expenditure is contraction of health services production, which implies the less-than-total satisfaction of demand. This is the option that has an opportunity cost for the health sector: the displaced health care services. But it is not the only option. Evidence from qualitative studies of expenditure decisions by local HSDMs reveals reluctance to displace services and the deployment of other responses as well as displacement:

namely efficiency improvements and increased expenditure. We are developing a model that better reflects a reality where displacement is not the only response to pressure for increased expenditure on a new or expanded service.

References

Audit Commission NHS, 2012. NHS financial year 2011/12. London: Audit Commission.

Bojke, C., Castelli, A., Street, A., Ward, P. and Laudicella, M., 2013. Regional variation in the productivity of the English National Health Service. Health Economics 22 (2), pp. 194-211.

Cyert, R., March, J.G., 1992. A behavioural theory of the firm. 2nd edition. Oxford: Blackwell Publishers.

Department of Health, 2008. Department of Health annual report and accounts 2007-08. HC1042. London: The Stationery Office.

Department of Health, 2009. Department of Health annual report and accounts 2008-09. HC456. London: The Stationery Office.

Department of Health, 2010. Department of Health annual report and accounts 2009-10. HC208. London: The Stationery Office.

Department of Health, 2013. Department of Health annual report and accounts 2013-14. HC14. London: The Stationery Office.

Førsund, F. and Jansen, E. 1977. On estimating average and best practice homothetic production functions via cost functions. International Economic Review. 18 (2), pp. 463-476.

Karlsberg Schaffer, S., Sussex, J., Devlin, N., and Walker, A., 2015. Local health care expenditure plans and their opportunity costs. Health Policy. 119 (9), pp. 1237-1244.

Karlsberg Schaffer, S., Sussex, J., Hughes, D. and Devlin, N., 2014. Opportunity costs of implementing NICE decisions in NHS Wales. OHE Research Paper 14/02. June 2014. London: Office of Health Economics.

Samuelson, W. and Zeckhauser, R., 1988. Status quo bias in decision making. Journal of Risk and Uncertainty. 1 (1), pp. 7-59.

Sussex, J., 1998. Controlling NHS expenditure: the impact of Labour's NHS White Papers. London: Office of Health Economics.