UNIVERSITY OF OSLO

HEALTH ECONOMICS RESEARCH PROGRAMME

The effect of activity-based financing on hospital efficiency:

A panel data analysis of DEA efficiency scores 1992-2000

Erik Biørn
Department of Economics

Terje P. Hagen & Tor Iversen
Center for Health Administration

University of Oslo

Jon Magnussen
SINTEF Unimed Health Services Research

Working Paper 2002: 8

SINTEF Unimed

UNIVERSITAS OSLOENSIS • MDCCCXI •

The effect of activity-based financing on hospital efficiency:

A panel data analysis of DEA efficiency scores 1992-2000[1]

Erik Biørn[a,e], Terje P. Hagen[b,e,*], Tor Iversen[c,e] and Jon Magnussen[d,e]

30 April 2002

Health Economics Research programme at the University of Oslo

HERO 2002

[a] Department of Economics, University of Oslo, PO Box 1095 Blindern, NO-0317 Oslo, Norway.
E-mail: erik.biorn@econ.uio.no

[b] Center for Health Administration, University of Oslo, Rikshospitalet, NO-0027 Oslo, Norway.
E-mail: t.p.hagen@samfunnsmed.uio.no

[c] Center for Health Administration, University of Oslo, Rikshospitalet, NO-0027 Oslo, Norway.
E-mail: tor.iversen@samfunnsmed.uio.no

[d] SINTEF Unimed Health Services Research, NO-7465 Trondheim, Norway.
E-mail: jon.magnussen@unimed.sintef.no

[e] Health Economics Research Programme at the University of Oslo (HERO)

[*] Corresponding author

[1] The paper presents results from an evaluation project initiated by the Norwegian Ministry of Health and Social Affairs. Financial support from the Ministry and from the Norwegian Research Council to the Health Economics Research Programme at the University of Oslo is acknowledged. We wish to thank participants at the Nordic Health Economists' Study Group Meeting in Lund 2000, discussant Peter Smith and other participants at the Health Economics Workshop at Universitat Autonoma de Barcelona 26-27 January 2001, and participants at various seminars at the University of Oslo for constructive comments on previous versions of the paper. The usual disclaimer applies.

Financial support from The Research Council of Norway is acknowledged.
ISSN 1501-9071, ISBN 82-7756-088-5

Abstract:

Activity-based financing (ABF) was implemented in the Norwegian hospital sector from 1 July 1997. A fraction (30 to 50 per cent) of the block grant from the state to the county councils has been replaced by a matching grant depending upon the number and composition of hospital treatments. As a result of the reform, the majority of county councils have introduced activity-based contracts with their hospitals. This paper studies the effect of activity-based funding on hospital efficiency. We predict that hospital efficiency will increase because the benefit from cost-reducing efforts in terms of number of treated patients is increased under ABF compared with global budgets. The prediction is tested using a panel data set from the period 1992-2000. Efficiency indicators are estimated by means of data envelopment analysis (DEA) with multiple inputs and outputs. Using a variety of econometric methods, we find that the introduction of ABF has improved efficiency when measured as technical efficiency according to DEA analysis. Contrary to our prediction, the result is less uniform with respect to the effect on cost-efficiency. We suggest several reasons why this prediction fails. Keywords are poor information of costs, production-oriented drive, tight factor markets and soft budget constraints.

JEL Classification: I11, I18, C23, L32

Keywords: Public hospitals, financing, efficiency, DEA-scores, panel data, Norway

1. Introduction

The question of optimal hospital reimbursement schemes has been widely discussed in the literature (see e.g. Newhouse, 1996, for an overview). The main trade-off is generally believed to be between providing incentives for efficiency in the production of hospital services and avoiding adverse patient selection. Put simply: high powered prospective payment systems are generally believed to increase efficiency, but may generate problems due to creaming, skimping or dumping (Ellis, 1998). Fee-for-service systems, on the other hand, may give rise to serious inefficiencies in the hospital system.

When hospital reimbursement schemes have received attention in the literature, the main distinction has often been made between retrospective (e.g. fee-for-service) and prospective (e.g. fixed price per DRG) systems. Hence, much of the empirical literature deals with the US transition from a fee-for-service system to a prospective DRG-based system for its Medicare population in 1983 (Hadley et al., 1989; Hodgkin et al., 1994; Newhouse, 1989). Recently, Yip and Eggleston (2001) have also published a similar study of the change from retrospective to prospective reimbursement with Chinese data. In many European countries, however, the policy question has been (and is) whether to finance hospitals by global budgets or introduce activity-based financing systems. Hence, the choice is between two different forms of prospective payment. In this respect the insight gained from the US studies is of limited interest.

The empirical evidence of the effects of reforming systems based on global budgets is scarce. The Thatcher reforms in Great Britain in the early 1990s aimed at improving efficiency both by introducing competition between hospitals and by changing contracts based on costs to contracts based on costs and volume. Unfortunately, there is little published evidence on the results of this reform. Le Grand (1999) reports an annual increase in efficiency post-reform of 2 per cent compared with 1.5 per cent prior to the reform. Koen (2000) is, however, skeptical about these results. In a summary of the evidence of the effects of increased competition, Propper (1997) is unable to find any effects. In a study of a reform with certain similarities in Sweden, Gerdtham, Rehnberg and Tambour (1999) find that a switch from budget-based allocations to output-based allocations leads to a 13 per cent decrease in costs among Swedish hospitals. The study utilizes data from two years, 1993 and 1994. Later analyses, in particular

Charpentier and Samuelsson (1999) have studied productivity changes in the county of Stockholm in the period from 1992 to 1997. They find productivity gains in 1993 and 1994 and productivity reductions in the following years.

In Norway, there are three levels of government: the state or central government, the counties and the municipalities.[1] In the period we analyze, hospitals are owned and financed by the county councils.[2] Before 1980 hospital costs were reimbursed on a per diem basis. This system was costly, and from 1980 hospitals were given annual global budgets. This led to a period of cost containment; annual expenditures increased by an average of around 1.1 per cent per year, but questions were raised about the efficiency of the hospital sector. Activity-based financing (ABF)[3] was implemented in the Norwegian hospital sector from 1 July 1997. A fraction of the block grant from the state to the county councils has been replaced by a matching grant depending on the number and composition of hospital treatments. At first, 30 per cent of the DRG-based cost of a treatment was refunded from the state. From 1 January 1998, the percentage was increased to 40 and from 1 January 1999, to 50.

The government's arguments for introducing ABF were put forward in a white paper from the Ministry of Health and Social Affairs (1995). An increase in the number of elective treatments was considered to be needed in order to fulfil the waiting list guarantee adopted by the parliament. Furthermore, an increase in the block grant to the county councils was assumed to be insufficient because of the leakage to other sectors for which the county councils are responsible, in particular secondary schools and transportation. A reform of the financing mechanism was therefore sought. By introducing a matching grant to the county councils the government intended to influence the county councils' cost of hospital treatment relative to other services, and hence, shift the county councils' priorities in the direction of hospitals. The government's and the parliament's intention was that the activity-based financing should also be implemented as activity-based contracts between a county council and its hospitals. The county councils were, however, free to decide the kind of funding mechanism they would use. It turned out that 15 of Norway's 19 county governments

[1] For a general description of the Norwegian health care system, see van den Noord et al. (1998) and European Observatory on Health Care Systems (2000).
[2] The central government has taken over both ownership and financing from January 2002. See http://www.dep.no/shd/sykehusreformen/aktuelt/rapport/030071-990126/index-dok000-b-n-a.html for a brief description.
3 The term used in Norwegian is ´Innsatsstyrt finansiering´ or the abbreviation ´ISF´.

introduced activity-based financing (ABF) of their hospitals when the matching grant was implemented. Another two county governments introduced activity-based financing of their hospitals from 1 January 1998, another one from 1 January 1999 and the last one from 1 January 2000.

In this paper we study the effect of this reform of the Norwegian financing system on hospital efficiency. The study may provide a valuable supplement to the literature on financing reforms. The present study comprises data for 48 somatic hospitals over a period of nine years, five prior to the reform and four after the reform. Thus, compared with other European studies (cf. Gerdtham, Rehnberg and Tambour, 1999; Le Grand, 1999; Sommersguter-Reichmann, 2000), we are able to analyze effects of the reform over a longer period of time.

Our paper is organized as follows: Section 2 presents our main hypothesis that activity-based funding of hospitals will improve efficiency relative to a situation where hospitals are funded by global budgets. This hypothesis is derived from a stylized model of hospital decisions. In Section 3, our data on hospital inputs and outputs are described. The estimation of efficiency is made using Data Envelopment Analysis (DEA) with data from the period 1992 - 2000. On average, technical efficiency is higher at the end of the period than at the beginning, while the average level of cost-efficiency is lower. Section 4 contains the empirical analysis of the hypothesis. We find that the introduction of ABF has improved efficiency when measured as technical efficiency according to the DEA analysis. The results are less uniform with respect to the effect on cost-efficiency. In some cases the estimated ABF effect is insignificant, in other cases it is significantly negative. According to our model predictions, we would have expected an increase in cost-efficiency as a result of the introduction of ABF. In the concluding remarks we suggest several reasons why this prediction fails. Keywords are poor information of costs, production-oriented drive, tight factor markets and soft budget constraints.

2. An economic model of the effect of activity-based financing on hospital efficiency

Inspired by the development of hospital financing in the US, the replacement of cost-based (retrospective) reimbursement by output-based (prospective) financing is studied in several works (see Dranove and Satterthwaite (2000) for a review). Under retrospective reimbursement the insurer covers hospital costs irrespective of their magnitude. Hence, there are no incentives to make efforts that aim at reducing production costs. With output-based financing a hospital's revenue is independent of previous costs (prospective financing). The hospital now reaps the gains from cost-reducing efforts, and an incentive for undertaking these efforts is created.

The effect of replacing a fixed, global budget with output-based financing is less obvious. Since both systems are prospective, the agent keeps the results of cost-reducing efforts in both systems[4]. In this section we study the economic mechanisms behind the hypothesis that:

Hospital efficiency is expected to be greater with activity-based financing (ABF) of hospitals than with fixed budgets.

The model we present summarizes the economic logic behind activity-based financing. Hospitals are complex organizations performing such multiple tasks as treatment of patients, education and research. A tractable formal model of hospital behavior may easily miss points of importance for actual decision-making and results. On the other hand, less formal reasoning may lead to logical inconsistencies between assumptions and conclusions. We shall comment on this further in the concluding remarks.

As already noted, in the Norwegian hospital sector there are three levels of decision-makers: the state, the county council and the hospital management. The county council, as hospital owner, receives revenue from the state as insurer. The county council is free to decide the type of financing system for its own hospitals. When ABF is introduced, a fraction of the block grant from the state to the county councils is replaced by a matching grant depending on the number and type of hospital treatments. Accordingly, for a county council the cost of

hospital treatment is reduced relative to other activities under the county council's responsibility. Due to the familiar substitution effect, this change in relative cost encourages the county council to increase hospital budgets relative to other budgets under its control. The effect of this budget increase on hospital efficiency is likely to depend on whether a fixed budget or ABF is in operation. On the other hand, hospital efficiency is likely to influence the county council's willingness to pay for hospital treatment. Hence, there is an interaction between the county council's and the hospital's decisions. In Appendix 1 this interaction is modeled as Nash equilibrium with decisions at the state level regarded as exogenous. In this section we concentrate on the economic mechanism that determines a hospital's composition of cost-reducing efforts and number of treatments, leaving the county council's decisions as exogenous. This analysis creates the necessary link between the theory and the empirical analysis that follows in Sections 3 and 4.

We assume that the hospital management has an additively separable utility function[5]:

$$U = u(n) - \gamma(e) \qquad (1)$$

where $u'(n)>0$, $u''(n)<0$, $\gamma'(e)>0$, $\gamma''(e)>0$, $\gamma'(e)>0$, $\gamma''(e)>0$, and the superscript ' (") denotes first (second) order derivative, n is the number of treated patients, and e is the level of cost-reducing efforts[6]. Cost reductions often require change in tasks and organization that involves discomfort and hence, have a negative impact on the utility of managers and employees. The marginal disutility of cost-reducing efforts is assumed to increase with the level of effort.

Since hospitals in this paper are non-profit institutions, profit is not an argument in (1). A hard budget constraint is assumed:

$$B + wn \geq c(n,e) + g \qquad (2)$$

---

[4] However, even in a formal system of global budgeting, cost-compensation may occur in practice.

[5] Teaching and research are omitted from the model. Costs that are driven by these activities are also excluded from the empirical model.

[6] Since cost-reducing efforts often involve resources (for instance time for meeting and discussion) with an alternative use in treating patients, the effect of cost-reducing efforts on costs should be interpreted as a net effect.

where the left-hand side of the inequality sign is the hospital's revenue and the right-hand side is the cost. B is fixed revenue, w is revenue per treated patient, g is a cost component exogenous to the hospital. This cost component depends upon the age and composition of hospital buildings, the geographical location of the hospital, etc. The effect of number of treatments and level of effort on cost is described by the cost function c(n,e). We assume positive and non-decreasing marginal costs of treating patients; i.e. $c_n^{'}(n,e) > 0$ and $c_{nn}^{''}(n,e) \geq 0$. An increasing marginal cost is likely to exist when some resources are fixed in the short run and capacity utilization is high. For instance, when waiting times occur at the x-ray department and the laboratories, an increase in a patient's length of stay is likely to occur.

The level of cost-reducing effort may have an impact on costs through many sources. For instance, a reorganization of personnel on call may release resources now available for elective treatments. Improved planning and utilization of operating theatres and other measures may increase the flow of patients. The cost function is assumed to be decreasing in effort at an increasing rate; i.e. $c_e^{'}(n,e) < 0$ and $c_{ee}^{''}(n,e) > 0$.

If the level of effort only influences fixed costs (for instance costs related to personnel on call, and regular staffing of operating theatres), the effect of an increase in effort on marginal cost is obviously zero. If the level of effort also influences variable costs, we assume the marginal effect of effort on marginal cost to be greater (in absolute value) the higher the level of capacity utilization is. Hence, we assume the interaction term between number of treatments and effort to be $c_{ne}^{''}(n,e) \leq 0$.

Since profit is not an argument in Eq (1), Eq (2) is obviously fulfilled with equality. Maximizing the objective function Eq (1) constrained by the budget Eq (2) gives from the first-order conditions of an interior solution of the Lagrangian[7]:

$$\begin{aligned} &u'(n)c_e^{'}(n,e) + \gamma'(e)[c_n^{'}(n,e) - w] = 0 \\ &c(n,e) + g - B - wn = 0 \end{aligned} \qquad (3)$$

[7] That n is considered as decision variable implies that emergency cases that the hospital cannot control, are ignored. These emergency cases account for a considerable proportion of a general hospital's patients. Our argument is still valid, because the main purpose of the reform of the financing system was to encourage an increase in the number of elective admissions.

and the second-order condition for a constrained maximum of the objective function is :

$$D \equiv [c_n'(n,e) - w]^2 \gamma''(e) + [c_n'(n,e) - w]u'(n)c_{ee}''(n,e)$$
$$-2c_e'(n,e)c_{ne}''(n,e)u'(n) - [c_e'(n,e)]^2 u''(n) - \qquad (4)$$
$$\gamma'(e)c_e'(n,e)c_{nn}''(n,e) > 0$$

Sufficient conditions for D>0 are:

(i) $[c_n'(n,e) - w] \geq 0$; i.e. the revenue per treated patient is less than the marginal cost.

(ii) $c_{ne}''(n,e) = 0$; i.e. the effect of e on the marginal cost of treating patients is zero.

(i) is fulfilled since we have assumed $c_{nn}''(n,e) \geq 0$, i.e. $c_n'(n,e) \geq \frac{c\ (n,e)}{n}$, and we consider a system with per case payment covering only a proportion of the average cost. According to our assumptions regarding the cost function, (ii) may not be fulfilled, and then pulls in the direction of a convex Lagrangian. The second-order condition states that this effect is small enough to ensure a concave Lagrangian.

Eq (3) determines n and e as functions of w and B:

$$n = n(\underset{?}{w}, \underset{?}{B})$$

$$e = e(\underset{?}{w}, \underset{?}{B})$$

where the sign under a functional argument shows the sign of the impact of an increase in the variable. Due to the interaction effect, $c_{ne}''(n,e)$, the sign of all effects are in general indeterminate. With a small absolute value of interaction effect, an increase in B has an income effect that leads to an increase in the number of treatments (n) and a decline in cost-reducing efforts (e). Similarly, an increase in the revenue per treatment (w) has an income effect that pulls in the direction of increased n and reduced e, while the substitution effect pulls in the direction of an increase in both n and e. Hence, the total effect on n is positive, while the total effect on e is indeterminate, even if the interaction effect is small.

We model the change from a global budget to a mixed system as an increase in w and a reduction in B of a magnitude allowing the hospital to choose the same n and e after the change as chosen before the change. The reduction in the fixed budget is then assumed to be $-n^o\Delta w$, where n^o is the optimal number of patients treated under a global budget and Δw is the increase in revenue per treatment. By means of differentiating (3) we find:

$$\frac{\partial n}{\partial w} - n^o \frac{\partial n}{\partial B} = -\frac{1}{D}[\gamma'(e)c_e^{'}(n,e)] > 0 \quad (5)$$

and

$$\frac{\partial e}{\partial w} - n^o \frac{\partial e}{\partial B} = \frac{1}{D}[c_n^{'}(n,e) - w] > 0 \quad (6)$$

Hence, the hospital's optimal n is expected to increase when revenue per treatment replaces a part of the fixed budget in the financing system. Accordingly, e is also expected to increase. Hence, we have:

$$e = h(w_{|dB=n^0 dw}^{+}, \underset{?}{B}) \quad (7)$$

We are now able to sum up the predictions:

- an increase in the budget is in general predicted to have an indeterminate effect on effort and hence, on hospital efficiency
- a change from a fixed budget to a combination of fixed budget and revenue per treatment is predicted to result in an increase in the level of effort and hence, an increase in hospital efficiency.

These predictions are tested in the Section 4.

3. Measures of efficiency

In order to analyse the effects of ABF on hospital efficiency, we need to establish measures of efficiency. This again raises two questions: the measurement of hospital production, and the choice of method when establishing efficiency measures.

Input and output of hospital production

Hospitals are multi-product firms, treating a variety of patients with a variety of inputs. There is no established consensus as to how one should most accurately measure the outputs of hospital production. Since the conceptual output, relative change in health, is unobservable,

we go on by measuring health services, rather than health. We have chosen the following outputs:[8]

Inpatient care: Inpatient care is measured as number of discharges adjusted for case-mix by weighting discharges by diagnosis related groups (DRGs). Day care is included in the measure of inpatient care.

Outpatient care: Outpatient care is measured as number of outpatient visits weighed by the fee paid by the state for each visit. Thus a hospital´s revenue from outpatient care is an approximation of the volume of outpatient care adjusted for case-mix. Outpatient revenue measured in NOK 1000 (Norwegian Kroner)is deflated to 2000 prices.

Hospital inputs are measured as:

Physician FTEs (full-time equivalents): The physician input is measured as number of FTEs per year. This is only an approximation of the number of hours actually worked, and may distort the efficiency measures if use of overtime varies substantially between hospitals and over time. Evidence suggests that the number of hours worked per FTE is fairly constant over the period studied here.[9]

Other labour FTEs: All other types of labour than physician labour are merged in one category. A more detailed specification of labour input did not alter the results.

Medical expenses: Medical expenses are measured in NOK 1000, and deflated to 2000 prices.

Total running expenses: Total running expenses are used as alternative input in one model, where the purpose is to provide a measure of cost-efficiency. Running expenses measured in NOK 1000 are deflated to 2000 prices.

---

[8] A variety of other specifications have been chosen as well. None present a picture that substantially differs from the one chosen here.

[9] A survey among 2100 hospital physicians (Hagen and Nerland 2001) indicates that approximately 78 per cent of the respondents spend an equal number of working hours on patient related work in 2001 as they did before the introduction of ABF, 10 percent reported an increase and 12 percent reported a reduction in the number of patient related work hours. Approximately 35 percent indicated that the number of work hours spent on administrative work has increased a bit.

Norwegian hospital cost data are imperfect in the sense that capital costs are not included. If the use of high-cost medical equipment has increased over this time period, the results given here are likely to overstate the growth in hospital efficiency. Summary statistics are given in Table 1.

(Table 1)

Efficiency concepts

The basic efficiency concept used in this paper is that of technical efficiency. A hospital is said to be technically efficient if an increase in an output requires a decrease in at least one other output, or an increase in at least one input. Alternatively, a reduction in any input must require an increase in at least one other input or a decrease in at least one output. This is the usual Pareto-Koopmans notion of efficiency. The measures used in this paper originated with Farrell (1957) and were further developed for piecewise linear technologies in Fare and Lovell (1978), Charnes, Cooper and Rhodes (1978) and Banker, Charnes and Cooper (1984). The non-parametric mathematical programming approach used in this paper has come to be known as Data Envelopment Analysis (DEA).

One advantage of DEA is that it accommodates a setting with multiple inputs and multiple outputs more easily than parametric models. Moreover, this approach does not require a specific functional form for the technology or specific distributional assumptions about the efficiency measure. A deterministic approach is susceptible to measurement errors. In this case we have used data that were collected and checked for errors by the Statistics Norway and the Norwegian Patient Register. Thus, we believe that we have taken sufficient steps in securing the quality of the data.

Formally, the efficiency measures are derived by first defining the reference technology relative to which efficiency is measured. Let $y=(y_1,.....,y_m) \in \Re_+^m$ denote a vector of outputs and $x=(x_1,.....,x_n) \in \Re_+^n$ denote a vector of inputs. Assuming constant returns to scale we can obtain a measure of input-saving technical efficiency (for unit 0), TE_{CRS}, by solving the following LP problem:

$$MIN_{TE,\lambda} TE_{CRS}$$
$$subject\ to:$$
$$\lambda Y \geq y_0$$
$$\lambda X \leq TE_{crs} x^0$$
$$\lambda_i \geq 0, i = 1, ..., k$$

Here *k* is the number of hospitals, *Y* is the *k*m* matrix of observed outputs, *X* is the *k*n* matrix of observed inputs, and λ is the intensity vector.

The efficiency frontier is based on a pooled set of observations, i.e. we calculate an intertemporal efficiency frontier (Harris et al., 2000, Tulkens & Vanden Eeckauout, 1993). This is done in order to be able to compare efficiency between years.

We also provide a measure of "cost-efficiency" by measuring inputs in costs. The measure of cost-efficiency will be equal to Farrell's (1957) measure of total efficiency, i.e. the product of technical and allocative efficiency. When we measure the development in cost efficiency we note, however, that this may change due to a wage and price increase that deviates from the price deflator, and not necessarily due to suboptimal combinations of inputs.

Results

Average levels of efficiency are presented in Table 2. Best practice implies a level of 100; thus an average technical efficiency of around 82 in 2000 implies that hospitals on average are 18% below best practice.

(Table 2)

Technical efficiency increases over this period. Thus hospitals seem to improve their utilization of resources, and in particular to increase patient throughput. There is a large positive shift in efficiency the first year after the reform of the funding system. We return in Section 4 to the question of whether this can be attributed to the reform.

The trend in *cost-efficiency* is roughly equal to the trend in technical efficiency until 1996 when we observe a substantial decline in cost-efficiency. This is believed to be due to a particularly large increase in physician wages at that time. We also note that there is a decline in cost-efficiency between 1998 and 1999, while technical efficiency is constant. This is believed to be related to an expensive increase in activity between these two years.

4. Empirical specifications and results

As pointed out in the introduction, 15 of the country's 19 county governments introduced activity-based financing (ABF) of their hospitals at the same time as the central government introduced the matching grant (1 July 1997). Another two county governments introduced activity-based financing from 1 January 1998, another one from 1 January 1999 and the last from 1 January 2000. The main question to be answered is whether the introduction of ABF of hospitals has affected hospital efficiency as stated in Eq. (7).

Operationalization of the models

Based on the theoretical arguments in Section 3, we assume that hospital efficiency (E), measured as technical efficiency (TE) and cost-efficiency (CE), is affected by the six variables defined in Table 3.

(Table 3)

BUD is standardized per hospital bed to correct for differences in hospital size. As discussed in Section 3, outpatient revenues are included in the output vector in the efficiency analysis (DEA) to account for numbers of outpatients. We are forced to do this since data on the number of outpatients are lackin for many of the large hospitals in the period we are analyzing. However, outpatient revenues have both a price and a volume component. Since fees for outpatient services have increased in the period, we may overestimate the change in efficiency. To take account of this, we include a variable measuring outpatient revenues as a share of total hospital revenues (OUT). Furthermore, we include a variable representing the share of patient-days with irregularly long lengths of stay (LONG) to capture possible effects

of this on hospital efficiency. There are reasons to believe that LONG to a large extent is beyond the hospital's control and probably affected by the volume and composition of formal care for the elderly in surrounding local governments. The number of beds (BEDS) is included to represent scale effects not captured by the DEA-measures. Hospitals are of different types, ranging from local hospitals with few or no specialties to university clinics. Introductory analyses indicate that one dummy variable, TYPE, is sufficient to capture these differences. Table 4 presents descriptive statistics for explanatory variables.

(Table 4)

Efficiency is measured both as technical efficiency (TE) and as cost-efficiency (CE), and parallel sets of estimates are reported for the two measures. In order to examine the robustness of the results, different model specifications and estimation methods are considered. Table 5 relates to static models, Table 6 dynamic models. The dynamic models are intended to represent lagged response of efficiency to introduction of ABF. All the estimation results reported below are obtained by PC programs constructed in the Gauss software code by the first author.

Results from static models

The Ordinary Least Squares (OLS) estimates obtained from a static regression model are reported in Table 5 – Model 1. If the true disturbance covariance matrix is not a scalar matrix, it is known that the usual standard error estimates are biased. Since a non-scalar disturbance covariance matrix is indicated by the results in Model 2, we report in Model 1 the unbiased OLS standard error estimates, assuming an error components specification following from random hospital-specific heterogeneity, cf. Baltagi (2001, section 2.3).

(Table 5)

The hospital budget comes out with a negative coefficient (-11.91 for TE, t-value=-3.72 and -14.58 for CE, t-value=-6.86), indicating that the budget level affects efficiency negatively. Hospital size (BEDS) has no significant effect, which we take as supportive of the use of the DEA CRS-framework. Both outpatient revenues as a share of total hospital revenues (OUT) and "long-term days as a share of total numbers of patient days" (LONG), both of which are treated as control variables, have significant effects. Our main interest is the ABF dummy and

we observe that it comes out with a significantly positive effect on TE (coefficient estimate = 5.87, t-value=4.01), but also that the effect on CE is insignificant (coefficient estimate = -0.72, t-value = -0.74).

Models 2, 3 and 4 in Table 5 give estimation results when hospital-specific heterogeneity is allowed for. Various parameterizations are considered. Underlying Model 2 is an assumption that the heterogeneity can be captured as a hospital-specific effect, α_i, in the intercept, with zero mean, constant variance, zero correlation across hospitals and zero correlation with the specified regressors and the genuine disturbance. The relationship has the form

$$y_{it} = x_{it}\beta + \alpha_i + u_{it}\ , \qquad i = 1,....,N;\ t = 1,.....,T, \tag{8}$$

where y is efficiency, x is the vector of explanatory variables, β is the coefficient vector, and u_{it} is a genuine disturbance. Model 2 gives the Generalized Least Squares (GLS) estimates based on variance components estimated from residuals (OLS in the first step, GLS in later steps) when the process has been iterated until convergence (100 iterations). The iterative GLS estimates will, under certain regularity conditions, coincide with the Maximum Likelihood estimates when the disturbance components, α_i and u_{it}, are normally distributed, cf. Breusch (1987) and Baltagi and Li (1992). We note that the GLS estimates are superior to the OLS estimates, as they have considerably smaller standard errors. Again, the effect of ABF on TE is significantly positive, with a coefficient estimate of 2.8 (t-value=3.13), while the effect on CE now is significantly negative (estimate=-1.2, t-value=-2.20). The effect of BUD turns out insignificant for TE but significant for CE (estimate -10.90, t-value=-10.45) and the effect of BEDS is negative, but not significant. The parameter ρ, given at the bottom of the table, measures the degree of latent, hospital-specific heterogeneity on efficiency. It is defined as the ratio of the variance of the hospital-specific effect to the variance of the `gross disturbance' $\alpha_i + u_{it}$, i.e., $\rho = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_u^2)$. This parameter has the alternative interpretation as the coefficient of correlation between two `gross disturbances' from the same hospital in different years. The heterogeneity is sizeable, as the ρ estimate is 0.58 for TE and 0.61 for CE. Note, however, that these estimates measure unobservable heterogeneity, but may also, to some extent, represent the effect of unspecified hospital-specific regressors which are not recorded in our data set, e.g. building year and the technical status of the hospital.

An alternative, and less restrictive, way of modeling heterogeneity may be to treat it as fixed, represented by hospital-specific shifts in the intercept term of the regression equation and using OLS, which is equivalent to including hospital-specific dummies in the regression equation; cf. e.g. Baltagi (2001, section 2.2). Still more heterogeneity can be modelled by including year-specific dummies in addition to the hospital-specific ones in order to capture intercept differences in the efficiency equations across years. The results are given in Table 5 – Model 3 and Model 4 - respectively. The former estimates are denoted as `within hospital' estimates, since they utilize the data variation in the panel data set `within' the hospitals only. The latter are denoted as `within hospital and years' estimates (or residual estimates), since the only data variation they utilize in the estimation of the regression coefficients, including the ABF coefficient, is the variation which remains when both the variation between hospitals and between years have been `accounted for', see e.g. Baltagi (2001, section 3.2). Note that this fixed effects-OLS approach leads to more robust inference than the random effects-GLS approach, since it relies on no distributional assumptions with respect to the hospital-specific effects. For example, the estimates will remain consistent even if the latent heterogeneity is correlated with some of the specified regressors, which is not the case for the estimates from the random effects model. On the other hand, they may lose some estimation efficiency, and we note that the standard error estimates in Model 4 exceed those in Model 3, which again exceed those in Model 2. The conclusion that ABF improves TE still remains, however; its within hospital coefficient estimate is 1.8 (t-value=2.0), the estimated effect on CE is barely significantly negative (coefficient estimate -1.1, t-value = -1.98). Somewhat larger estimates are obtained when including both fixed hospital-specific and fixed year-specific shifts in the intercept. The estimated effect of ABF on both TE and CE are significantly positive, 3.4 (t-value=2.53) and 2.0 (t-value=2.58), respectively. A possible explanation of the positive effect on CE is that the yearly dummies in the latter model capture the negative trend in CE in the period after 1996 (c.f. Table 2). The effects of BUD are negative on CE in both Model 3 and 4. However, its effect on TE is positive and significant according to both models.

The results described so far are based on separate estimation of the two equations of the form (8) for TE and CE. Owing to the potential correlation between the genuine disturbances and between the latent heterogeneity in the two equations, an efficiency gain may be obtained by estimating the two equations jointly as a system of regression equations by means of GLS, see Baltagi (1980), (2001, chapter 6). Results corresponding to those in Model 2 are given in Model 5. The qualitative results are not substantially changed, but the standard error estimates

are smaller, which reflects the improved efficiency. Comparing Model 5 with Model 2 we see that the iterative GLS estimate of the coefficient of the ABF dummy is increased from 2.81 to 2.94 (t-value=3.30) for TE and reduced from -1.23 to -1.28 (t-value=2.29) for CE. The estimated variances of the genuine disturbances in the two equations are 30.6 and 11.8, respectively, and their covariance is 10.4. The corresponding estimates for the hospital-specific effects are 40.5, 18.6, and 19.1. Overall, this indicates, not unexpectedly, rather strong positive cross equation correlation between the genuine disturbances and between the hospital-specific effects. There is thus latent variation in efficiency between hospitals, which affects both measures we consider. The implied ρ parameter estimates are approximately the same as in the case with separate GLS estimation of the two equations, 0.57 for TE and 0.61 for CE, again indicating a high degree of latent heterogeneity in hospital efficiency.

Results from dynamic models

The econometric versions of the theory models we have considered so far are static. A sensible hypothesis may be the hospitals may adjust with a lag to the ABF reform. A simple and (in terms of the number of parameters) economic way of modelling this is to extend (8) to the autoregressive form

$$y_{it} = x_{it}\beta + y_{i,t-1}\lambda + \alpha_i + u_{it}\ , \qquad i = 1,....,N;\ t=1,.....,T, \qquad (9)$$

where λ is between zero and one. If this model is appropriate, estimating it by OLS or GLS is inconsistent, unlike model (8), because the lagged regressor $y_{i,t-1}$ and the latent effect α_i are correlated. The intuition is that the effect of the latent heterogeneity is included in all the observations on hospital efficiency. Attempts to do this gave estimates of λ and the coefficient of ABF equal to 0.70 and 2.68 for OLS and 0.70 and 3.21 for GLS (both "significant"), respectively, and a very low value of ρ for the latter. The high λ estimate indicates a substantial delay in the hospitals' response. However, the within estimator, which might be an answer to this inconsistency problem because it eliminates α_i, is also inconsistent when T is finite and N goes to infinity (see Sevestre and Trognon (1985)) and has not been computed. Another way of eliminating the hospital-specific heterogeneity from (9) is to transform it to differences as follows:

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta y_{i,t-1}\lambda + \Delta u_{it}\ , \qquad i = 1,....,N;\ t=1,.....,T, \qquad (10)$$

where Δ is the one period backward difference operator. Table 6 – Model 6 contains OLS estimates of (10). We find low, and indeed negative, estimates of the autoregressive coefficient: -0.15 (t-value=-2.78) for TE and -0.14 (t-value=-3.30) for CE. This indicates that there is hardly any lag in the response when the unobserved heterogeneity is eliminated by taking differences, and the estimated effect of ABF is still positive and significant for TE (coefficient estimate=2.20, t-value=2.66), but insignificant for CE (coefficient estimate=-0.06, t-value=-0.10). However, when estimated within Eq. (10), the effect of BUD on TE comes out with a positive estimate (4.50, t-value 2.17). Again, this OLS estimator is inconsistent since $\Delta y_{i,t-1}$ is correlated with $\Delta u_{i,t}$ (both `containing' $u_{i,t-1}$). Instrumenting the lagged differenced regressor $\Delta y_{i,t-1}$, by lagged level values of the x vector and of y (which are correlated with $\Delta y_{i,t-1}$ but uncorrelated with $\Delta u_{i,t}$) may be an answer to this problem, see Arellano and Bond (1991) and Baltagi (2001, section 8.2). Table 6 – Model 7 and Model 8 - report results based on (10), using the Generalized Method of Moments (GMM), in which all lagged level values of x (excluding the lagged values of the ABF dummy) and of y are included in the instrument set in an `asymptotically optimal' way. Model 7 gives the `one step' GMM estimator, which is efficient in the case of disturbance homoskedasticity. Model 8 gives the `two step' GMM estimator, which utilizes the residuals from the one step estimation to improve the efficiency when heteroskedasticity of unspecified form is taken into account. The qualitative results from the OLS estimates in Model 6 remain, but the coefficient estimates differ. The estimated effect of ABF is still positive and significant for TE (coefficient estimate=1.95, t-value=2.62), but now negative and significant for CE (coefficient estimate=-2.60, t-value=-4.22). Again the effect of BUD is positive for TE (estimate=3.53, t-value=3.04) negative for CE (estimate =-13.01, t-value =-15.79). The Hansen-Sargan J-test statistics indicate that the orthogonality conditions underlying the GMM estimators are valid (p-value=0.78 for TE and 0.17 for CE).

A conclusion we can draw from these different models and estimation procedures is that the introduction of ABF has improved efficiency when measured as technical efficiency according to DEA analysis. The results are less uniform when it comes to assessing the effect of the reform on cost-efficiency. In some cases the estimated effect of ABF is insignificant, in other cases it is significantly negative. A possible explanation of this difference is given in the concluding section. Among the other variables, we are particularly interested in the effect of the size of the hospital budget, BUD. This effect comes out with a negative and significant

effect on CE (c.f. Hagen 1997), while the sign of its estimated effect on TE depends on the estimation method we use.

5. Concluding remarks

Activity-based financing (ABF) was introduced in the Norwegian hospital sector from 1 July 1997. The system implies that a proportion of the block grant from the central government to the county councils is replaced by a matching grant depending upon the number and composition of hospital treatments. The parliament set the matching grant at 30 per cent of the DRG price in 1997, increasing it to 45 per cent in 1998 and 50 per cent in 1999 and 2000. A main objective behind the introduction of activity-based financing was to encourage counties and hospitals to increase the number of hospital treatments without reducing hospital efficiency. There was no obligation on part of county councils to introduce activity-based financing of the hospitals, although the central government encouraged the counties to do so. As pointed out in the introduction, 15 of the country's 19 county governments introduced activity-based financing (ABF) of their hospitals at the same time as the central government introduced the matching grant (1 July 1997). Another two county governments introduced activity-based financing from 1 January 1998, another one from 1 January 1999 and the last from 1 January 2000. In this paper the focus has been on whether we can detect any relationship between activity-based financing and efficiency at the hospital level.

In Section 2 we developed the hypothesis to be tested in the empirical section. The central mechanism is that the introduction of ABF increases the hospital’s benefit from cost-reducing efforts in terms of number of treated patients. The hospital is encouraged to increase the level of these efforts and hence, efficiency. In Section 3 we describe the hospital efficiency for the period 1992 – 2000 by means of data envelopment analysis (DEA). We find that average technical efficiency is higher at the end of the period than at the beginning, while cost-efficiency shows the opposite trend.

In Section 4 results from the empirical analyses are presented. We find that the introduction of ABF has improved efficiency when measured as technical efficiency according to DEA analysis. This result is rather robust, as it holds for several econometric models and estimation

methods. The results are less uniform with respect to the effect on cost-efficiency. In some cases the estimated ABF effect is insignificant, in other cases it is significantly negative.

With regard to the effect of a hospital's budget on efficiency, the prediction from our model was indeterminate, due to a possible interaction effect between number of patients and effort on marginal cost. If the interaction effect is small, we expect that an increase in hospital revenue leads to a decline in efficiency since higher revenue makes cost-reducing efforts less attractive. In the empirical analysis this effect comes out with a negative and significant effect on cost-efficiency, while the sign of its estimated effect on technical efficiency depends on the estimation method we use. Two factors may contribute to the explanation of these apparently conflicting results. A budget increase may generate an increase in overtime work. The cost of overtime is included in the measure of cost-efficiency, but not in the measure of technical efficiency. Hence, the volume of inputs may be underestimated in the measure of technical efficiency. Also, use of overtime increases the price of production factors. Again, this is taken account of in the measure of cost-efficiency, but not in technical efficiency.

According to the predictions of our economic model in Section 2, we would have expected an increase in cost-efficiency after the introduction of ABF. We suggest several reasons why this prediction fails. Keywords are poor information of costs, production-oriented drive, tight factor markets and soft budget constraints.

The introduction of ABF was accompanied by a strong signal to hospitals to increase the number of treated patients in order to reduce politically annoying waiting lists. As a means to increase production without reducing cost-efficiency, hospital data indicate that ABF has been successful. For inpatients measured in DRG-equivalents there was an average yearly increase in hospital activity of 3.2 per cent in the period from 1997 to 1999, compared with 2.0 per cent per year in the period from 1992 to 1996. Because of poor information systems the cost of increasing the number of treated patients is uncertain, and the reported information is probably downward biased because of incentives to underestimate costs at the hospital department level. Because of tight markets for physicians and nurses, marginal resources could only be mobilized by paying a high compensation. Our empirical finding of increased technical efficiency and constant or declining cost-efficiency is consistent with expensive factors of production at the margin. The politically initiated production-oriented drive was probably also interpreted by county councils and hospitals as a signal of softer budget

constraints. Parallel to the introduction of ABF a marked increase in hospital deficits was revealed. The deficits were to a large extent covered by supplementary funds from the state to county councils and hospitals.

We have found considerable heterogeneity between hospitals both regarding the level of efficiency and regarding the effect of ABF on hospital efficiency (not reported in the tables). Hence, a follow-up is to examine whether there is a relationship between the effect of ABF and the initial level of efficiency. Is it the relatively efficient or inefficient hospitals that respond to the reform of the financing system? Perhaps the inefficient hospitals turn out to be non-responders and hence, conduct their business as usual without much consideration for institutional reform.

Our study emphasizes the importance of factor markets for the result of reforms in the hospital sector. With an excess supply of health personnel, employment and production could have been increased with a roughly constant wage level, and hence with a more beneficial effect on cost-efficiency.

The theoretical prediction of increased cost-efficiency was based on a model that assumed full information of costs and hard budget constraints. The institutional structure of the hospital sector has probably more in common with the institutions surrounding the soft budget constraints described by Kornai (2001). In that case the effect of the formal financing system is expected to be small, because of the cost-compensating properties of soft budget constraints. Hence, the power of economic incentives, as described in our model, depends on the existence of reasonably hard budget constraints. Hence, measures that work in the direction of hardening budget constraints seem essential to achieve efficiency gains in the hospital sector. In this context, decentralization of economic responsibility and competition are potential measures that should be further explored.

References:

Arellano, M., and S. Bond (1991): Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58, 277-297.

Banker, R.D., A. Charnes, and W.W. Cooper (1984): Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Management Science*, 30, 1078-92.

Baltagi, B.H. (1980): On seemingly unrelated regressions with error components. *Econometrica*, 48, 1547-1551.

Baltagi, B.H. (2001): *Econometric analysis of panel data.* Second edition. Chichester: Wiley, 2001.

Baltagi, B.H., and Q. Li (1992): A monotonic property for iterative GLS in the two-way random effects model. *Journal of Econometrics*, 53, 45-51.

Biørn, E. (2001): The Efficiency of Panel Data Estimators: GLS versus Estimators Which Do Not Depend on Variance Components. Department of Economics, University of Oslo. Memorandum No. 28/2001.

Breusch, T.S. (1987): Maximum likelihood estimation of random effects models. *Journal of Econometrics*, 36, 383-389.

Charnes, A., W. W. Cooper, and E. Rhodes (1978): Measuring the Efficiency of Decision Making Units. *European Journal of Operational Research*, 2, 429-44.

Charpentier, C., and L. Samuelsson (1999): "Effekter av en sjukvårdsreform - En analys av Stockholmsmodellen". Stockholm: Nerenius och Santerus Förlag AB.

Dranove D., and M. A. Satterthwaite (2000): The industrial Organization of health care markets. In A.J. Culyer and J.P. Newhouse (ed.): *Handbook of Health Economics* vol. 1B, 1093-1140. Amsterdam: Elsevier Science.

Ellis, R. P. (1998): Creaming, skimping and dumping: provider competition on the intensive and extensive margins. *Journal of Health Economics*, 17, 537-555.

European Observatory on Health Care Systems (2000): *Health Care Systems in Transition - Norway*. Copenhagen: WHO Regional office for Europe.

Farrell, M.J. (1957): The measurement of productive efficiency. *Journal of the Royal Statistical Society*, Series A, General, 120 (3), 253-81.

Färe, R., and C. A. Knox Lovell (1978): Measuring the Technical Efficiency of Production. *Journal of Economic Theory*, 19, 150-62.

Gerdtham, U.-G., C. Rehnberg and M. Tambour (1999): The impact of internal markets on health care efficiency: Evidence from health care reforms in Sweden. *Applied Economics*, 31, 935-945.

Hadley, J., S. Zuckerman and J. Feder (1989): Profits and fiscal pressure in the Prospective Payment System: their impacts on hospitals. *Inquiry*, 26, 354-365

Hagen, T. P. (1997): Agenda-setting Power and Moral Hazard in Principal-Agent Relations: Evidence from Hospital Budgeting in Norway. *European Journal of Political Research*, 31, 287-314.

Hagen, T. P. and S. M. Nerland (2001): Sykehuslegenes oppfatning av ISF: Kartlegging basert på en spørreundersøkelse. Working paper 2001:3. Oslo: Center for Health Administration, University of Oslo.

Harris II, J., H. Ozgen and Y. Ozcan (2000): Do mergers enhance performance of hospital efficiency? *Journal of Operational Research Society*, 51, 801-811.

Hodgkin, D., and T. G. McGuire (1994): Payment levels and hospital response to prospective payment. *Journal of Health Economics*, 13, 1-29.

Hsiao, C. (1986): *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Koen, V. (2000): Public Expenditure Reform: The Health Care Sector in the United Kingdom; *OECD Economics Department Working Papers*, ECO/WKP (2000)29

Kornai, J. (2001): Hardening the budget constraint: The experience of the post-socialist countries. *European Economic Review*, 45, 1573-1599.

LeGrand, J. (1999): Competition, Cooperation or Control? Tales from the British National Health Service. *Health Affairs*, 18 (3), 27-39.

Ministry of Health and Social Affairs (1995): Ventetidsgarantien – kriterier og finansiering, *Stortingsmelding nr 44, 1995-96.* Oslo: Statens trykning.

Newey, W. K. (1985): Generalized Method of Moments Specification Testing. *Journal of Econometrics*, 29, 229-256.

Newhouse, J. P. (1989): Do unprofitable patients face access problems? *Health Care Financing Review*, 11, 33-42

Newhouse, J. P. (1996): Reimbursing Health Plans and Health Providers: Efficiency in Production Versus Selection. *Journal of Economic Literature,* XXXIV, 1236-1263

Propper, C., and N. Söderlund (1998): Competition in the NHS internal market: An overview of its effects on hospital prices and costs. *Health Economics*, 7, 187-197.

Sevestre, P., and A. Trognon (1985): A note on autoregressive error component models. *Journal of Econometrics*, 28, 231-245.

Sommersguter-Reichmann, M. (2000): The impact of the Austrian hospital financing reform of hospital productivity: empirical evidence on efficiency and technology changes using a non-parametric input-based Malmquist approach. *Health Care Management Science*, 3, 309-321.

Tulkens, H. and P. Van den Eeckaut (1993): Non-Parametric Efficiency, Progress and Regress Measures for Panel Data: Methodological Aspects. CORE Discussion Paper no 9316. Louvain-la-Neuve: Center for Operations Research and Econometrics, Universite Catholique de Louvain.

van den Noord, P., T. P. Hagen and T. Iversen (1998): The Norwegian Health Care System. *Economics Department Working Papers no. 198.* Paris: OECD.

White, H. (1984): *Asymptotic Theory for Econometricians*. Orlando: Academic Press.

Yip, C., and K. Eggleston (2001): Provider payment reform in China: The case of hospital reimbursement in Hainan Province. *Health Economics*, 10, 325-339.

Appendix 1

In section 2 we analyzed the effect of a change in the financing system on a hospital's optimal choice of number of treatments and level of effort. The main result is described by Eq (7):

$$e = h(w_{|dB=n^{0}dw}^{+}, \underset{?}{B}) \qquad (7)$$

The county council determines the hospital's revenue, and the hospital determines effort and hence, the cost of hospital production. In this appendix the interaction between these variables is further analyzed.

The county council is also assumed to have an additive objective function with the number of treated patients (n) and the level of an aggregate, m, of the other services the county council is responsible for as arguments. The county council's budget constraint is R + sn = B + wn + rm, where R is the block grant from the state, s is the matching grant from the state per treated patient and r is a parameter that shows the expenditure per unit of other services the county council provides. Inserting from the hospital's budget constraint (2) into the county council's budget constraint gives R+sn = c(n,e) + g + rm. The county council's decision problem can then be described as:

$$\underset{n,m}{Max}\ v(n)+t(m)$$
$$s.t.$$
$$R+sn=c(n,e)+g+rm$$

We make the Lagrange function and find the first-order conditions:

$$\begin{aligned} &rv'(n)-t'(m)[c_n^{'}(n,e)-s]=0 \\ &c(n,e)+g+rm-R-sn=0 \end{aligned} \qquad (A1)$$

The second-order condition is:

$$G\equiv -(c_n^{'}(n,e)-s)^2 t''(n)-r^2 v''(n)+rt'(m)c_{nn}^{''}(n,e)>0$$

which is certainly fulfilled for $c_n^{'}(n,e)-s>0$.

Eq (A1) determines the county council's optimal n as a function of the exogenous variables:

$$n^* = n(\underset{+}{s}, \underset{+}{R}, \underset{+}{e})$$

$$m^* = m(\underset{?}{s}, \underset{+}{R}, \underset{?}{e}) \qquad \text{(A2)}$$

The effect of changes in the independent variables is found by differentiation of Eq (A1), and is denoted by the sign under the functional arguments in Eq (A2).

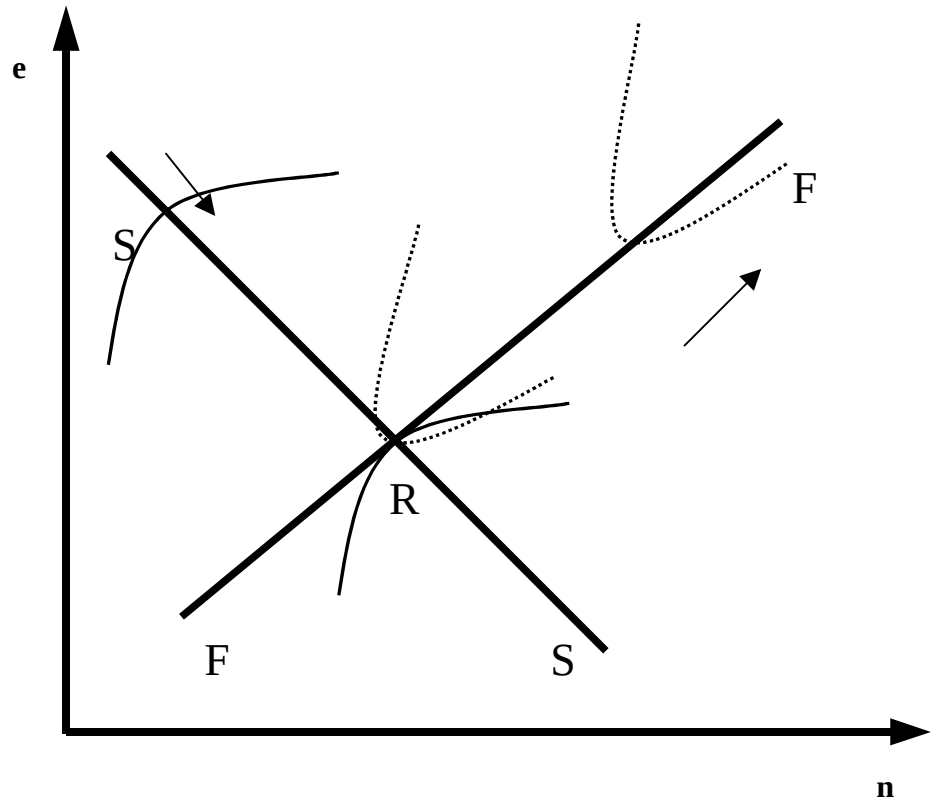
The change from block grant to a mixed system of block grant and matching grant is modelled on the assumption that the optimal allocation under the block grant system should also the feasible under the mixed system. Hence, we have that the reduction in the block grant is equal to $-n^*\Delta s$, where n^* is the optimal number of treated patients under the block grant system and Δs is the matching grant from the state. From the differentiation of Eq (A1) we find:

$$\frac{\partial n}{\partial s} - n^* \frac{\partial n}{\partial R} = \frac{1}{G}[rt'(m)] > 0 \qquad \text{(A3)}$$

Hence, due to the substitution effect, the county council's optimal n is expected to increase after the change from a block grant system to a mixed system of block grant and matching grant.

We now move on to the interaction between hospital decisions and county council decisions. The county council determines the hospital's revenue, while the hospital determines the volume and composition of hospital services.

Figure A1: The number of treated patients and the level of cost-reducing efforts under a block grant to the county council and fixed revenue for the hospital



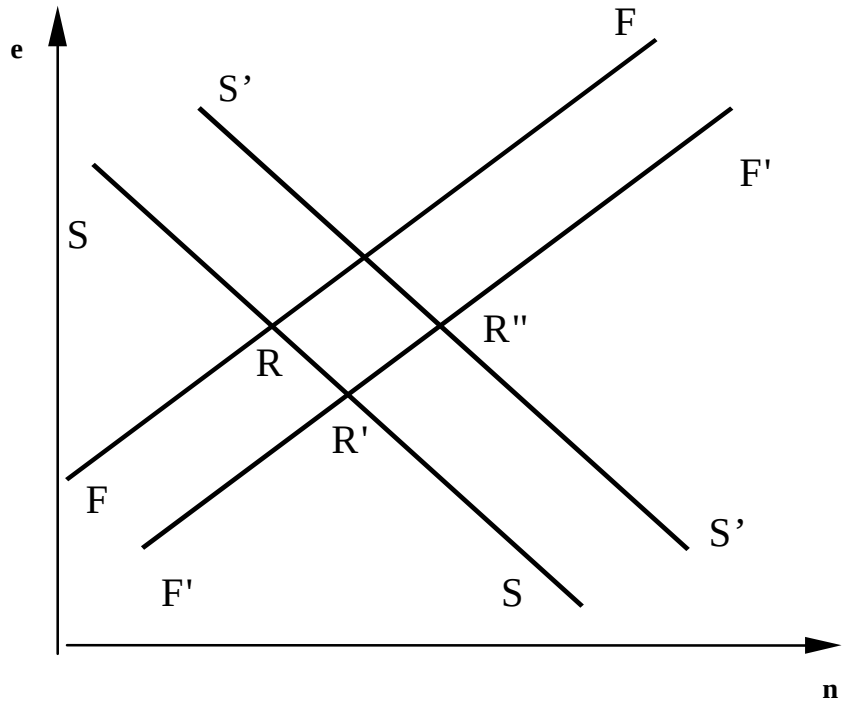
In Figure A1 the number of treated patients (n) is measured along the horizontal axis and the level of cost-reducing efforts (e) is measured along the vertical axis. The dotted curves show the county council's indifference curves in the (n,e) diagram. The line F-F shows the number of treatments that maximizes the county council's utility given the level of cost-reducing efforts (the county council's reaction curve). An increase in e results in an increase in the optimal n, since the cost per treatment declines. Hence, F-F is increasing in the diagram. The solid curves are the hospital's indifference curves in the (n,e) diagram. S-S expresses the hospital's optimal mix of e and n for various levels of revenue[10]. Referring to Section 2, we now assume that the indirect effect ($c''_{ne}(n,e)$) is not dominating, so that an increase in revenue leads to both an increase in n and a reduction in e.

[10] The tangential points between indifference curves and budget curves (not depicted) determine SS in the diagram.

We assume Nash equilibrium. In Nash equilibrium each party chooses his optimal decision given the other party's optimum. Since S-S is the hospital's reaction curve and F-F is the county council's reaction curve, Nash equilibrium is located at R, where the curves intersect. We are now ready to examine the effect of changing the financing system.

The introduction of a mixed system of block grant and matching grant results in, according to (A3), an increase in the county council's optimal n. Hence, F-F shifts to F'-F', as depicted in Figure A2. The new equilibrium depends upon the price elasticity of the county council's demand for treatments. In Figure A2 the new equilibrium is denoted R´.

Figure A2: The number of treated patients and the level of cost-reducing efforts under a mixed system



If the county council introduces a revenue component for the hospital that depends on the number of treated patients, n and e will both increase. In Figure A2, S-S then shifts to S´- S´ and the equilibrium shifts to R´´. From Figure A2 we now see that the number of treated patients is greater with a matching grant than with a block grant only (R´ compared to R) and even greater if a treatment-dependent component is introduced into the hospital financing system (R´´ compared to R´). We also see that the level of cost-reducing efforts and hence, hospital efficiency is greater with a treatment-dependent revenue component than without (R´´ compared to R´), in accordance with the hypothesis stated at the beginning of Section 2.

TABLE 1 Descriptive statistics, input and output-variables in DEA analyses. Mean (standard deviation) per year.

	1992	1993	1994	1995	1996	1997	1998	1999	2000
Physician FTEs	81.24 (89.34)	84.66 (96.16)	87.54 (101.07)	93.03 (106.68)	100.21 (116.74)	107.76 (131.67)	117.70 (145.22)	123.23 (151.26)	129.63 (157.24)
Other labour FTEs	706.82 (736.54)	720.07 (759.07)	733.38 (772.07)	762.92 (805.50)	810.56 (883.15)	837.05 (930.52)	869.29 (967.47)	901.10 (1005.89)	934.48 (1038.48)
Medical expenses	476.07 (594.37)	526.85 (651.68)	532.66 (700.56)	563.32 (734.33)	578.57 (785.35)	615.29 (856.02)	611.87 (807.67)	718.42 (951.63)	690.53 (887.20)
Total running expenses	317842.40 (314269.63)	325689.68 (325242.21)	329178.26 (329293.44)	343135.01 (343403.02)	373204.95 (388226.67)	404782.56 (423550.63)	429906.33 (452308.21)	470952.29 (491487.46)	482693.03 (497351.30)
Inpatient care	12609.48 (12589.84)	13075.27 (13016.61)	13084.92 (13058.78)	13780.56 (13814.47)	13950.94 (13959.23)	14303.46 (14269.07)	15317.78 (15535.97)	15917.20 (16095.12)	16356.41 (16285.23)
Outpatient care	32345.75 (36728.28)	33224.22 (38482.90)	34255.22 (38846.29)	36165.02 (41428.87)	38473.84 (45428.85)	46143.58 (53815.81)	48788.19 (58043.90)	52437.20 (64275.60)	52880.80 (62142.38)

TABLE 2: Average (standard deviation) levels of efficiency 1992-2000.
Two different input/output specifications.

	1992	1993	1994	1995	1996	1997	1998	1999	2000
Technical efficiency	78.95 (10.89)	78.97 (9.31)	77.68 (7.81)	78.17 (8.25)	77.15 (8.83)	80.38 (9.59)	80.90 (9.10)	81.68 (9.47)	82.12 (10.31)
Cost-efficiency	79.61 (9.04)	80.65 (8.75)	80.46 (7.79)	81.24 (9.03)	77.34 (7.56)	77.76 (8.35)	78.16 (8.28)	74.32 (8.33)	74.58 (8.84)

TABLE 3 Definitions of explanatory variables

Variable	Operationalization	Data source
BUD	Hospital's total revenues (accounting data)/BEDS	SINTEF Unimed, Statistics Norway
OUT	(Outpatient revenues/Total hospital revenues)*100	Statistics Norway
ABF	Dummy variable that takes the value of 1 if the hospital has an activity-based contract with the county council the current year, 0 otherwise	Center for Health Administration
LONG	(Number of days with irregularly long length of stay/Total number of inhospital days)*100	Norwegian Patient Register (NPR)
BEDS	Number of hospital beds	Statistics Norway
TYPE	Dummy variable that takes the value of 1 if the hospital is a university clinic or a central hospital, otherwise 0.	SINTEF Unimed

TABLE 4: Descriptive statistics, explanatory variables. Mean (standard deviation) per year.

	1992	1993	1994	1995	1996	1997	1998	1999	2000
BEDS	226.44 (213.04)	226.10 (215.02)	232.73 (224.63)	234.20 (225.96)	235.26 (226.75)	238.09 (234.99)	240.27 (233.18)	235.63 (230.49)	231.10 (221.47)
BUD	1352.06 (219.16)	1395.62 (228.68)	1368.94 (214.779	1422.07 (193.44)	1507.74 (219.36)	1628.68 (234.65)	1692.51 (268.65)	1902.21 (310.35)	1967.91 (321.14)
OUT	8.78 (2.71)	8.77 (2.73)	9.01 (2.70)	9.14 (2.91)	8.89 (2.63)	10.29 (2.43)	10.21 (2.51)	9.86 (2.42)	9.81 (2.50)
ABF	0	0	0	0	0	0.65	0.90	0.9	1.00
LONG	11.05 (4.42)	9.32 (4.43)	9.12 (4.18)	8.34 (3.92)	7.92 (3.60)	7.41 (3.41)	7.45 (3.68)	7.90 (3.61)	6.17 (2.63)
TYPE	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.35
N	48	48	48	48	48	48	48	48	48

TABLE 5: Static models. Estimates (t-values)

	Model 1		Model 2		Model 3		Model 4		Model 5	
	TE	CE	TE	CE	TE	CE	TE	CE	TE	CE
C	87.391	79.230	74.059	73.340	-	-	-	-	74.439	73.153
	(14.629)	(19.945)	(20.302)	(31.566)					(20.611)	(31.635)
BEDS	-1.146	0.412	-0.913	-0.444	1.111	-1.168	2.230	-0.093	-1.150	-0.371
	(-1.589)	(0.858)	(-1.602)	(-1.199)	(1.083)	(-1.822)	(2.095)	(-0.149)	(-2.055)	(-1.010)
BUD	-11.911	-14.584	-0.106	-10.902	3.009	-10.659	6.017	-6.556	-0.425	-10.752
	(-3.722)	(-6.860)	(-0.064)	(-10.449)	(1.761)	(-9.983)	(2.659)	(-4.933)	(-0.257)	(-10.323)
OUT	1.618	2.551	0.802	2.342	0.692	2.253	0.998	2.657	0.823	2.330
	(3.944)	(9.344)	(3.300)	(15.260)	(2.569)	(13.380)	(3.422)	(15.510)	(3.420)	(15.238)
ABF	5.869	-0.717	2.805	-1.231	1.834	-1.124	3.418	2.041	2.942	-1.283
	(4.012)	(-0.744)	(3.134)	(-2.198)	(2.013)	(-1.975)	(2.534)	(2.577)	(3.297)	(-2.294)
LONG	-44.860	-28.999	-3.474	12.350	9.623	18.446	-5.348	18.086	-2.027	12.276
	(-1.825)	(-1.775)	(-0.272)	(1.544)	(0.744)	(2.281)	(-0.378)	(2.178)	(-0.159)	(1.537)
TYPE	-0.301	-0.667	-1.639	1.721	-3.492	6.519	-6.336	3.512	-0.743	1.458
	(-0.093)	(-0.311)	(-0.587)	(0.943)	(-0.600)	(1.792)	(-1.087)	(1.026)	(-0.272)	(0.807)
$\hat{\sigma}_u$	7.823	5.409	8.543	5.555	5.198	3.248	5.121	3.007	8.434	5.512
ρ	-	-	0.582	0.618	-	-	-	-	0.570	0.612

Model 1: OLS estimates. Correct (unbiased) formulae for standard error estimates under random effects specification.

Model 2: Static random effects model. Iterative GLS (ML) estimates. Convergence achieved after 100 iterations

Model 3: Static model. Within hospital estimation. Corresponds to OLS with hospital-specific dummies added

Model 4: Within hospital and within year estimation. Corresponds to OLS with hospital-specific and year-specific dummies added

Model 5: Random effects model. Joint estimation of equations for technical efficiency and cost-efficiency. Iterative GLS (ML) estimates. Convergence achieved after 100 iterations

TABLE 6: Dynamic models. Estimates (t-values)

	Model 6		Model 7		Model 8	
	TE	CE	TE	CE	TE	CE
BEDS	2.821	-1.909	2.953	-1.654	2.234	-1.839
	(2.496)	(-2.562)	(2.242)	(-1.527)	(4.287)	(-4.355)
BUD	4.504	-13.060	5.266	-11.868	3.529	-13.012
	(2.166)	(-9.493)	(2.534)	(-7.484)	(3.035)	(-15.788)
OUT	1.125	2.393	1.235	2.582	1.102	2.678
	(3.822)	(12.263)	(3.422)	(7.949)	(6.236)	(23.027)
ABF	2.203	-0.058	1.071	-2.505	1.951	-2.599
	(2.661)	(-0.104)	(0.808)	(-2.038)	(2.620)	(-4.224)
LONG	13.663	9.064	12.816	10.647	12.256	10.532
	(0.960)	(0.969)	(0.825)	(1.061)	(1.314)	(2.475)
DY(-1)	-0.146	-0.137	-0.198	-0.238	-0.228	-0.257
	(-2.784)	(-3.304)	(-1.387)	(-2.491)	(-3.384)	(-5.682)

Model 6: Autoregressive model in differences. OLS estimates

Model 7: Autoregressive model in differences. GMM estimates, with level variables as instruments. One step GMM

Model 8: Autoregressive model in differences. GMM estimates, with level variables as instruments. Two step GMM. Jtest for validity of orthogonality conditions

Appendix 2. Estimation procedures, details

The purpose of this appendix is to give a condensed description, although more detailed than in the main text, of the estimation procedures for the static single-equation model, the static multi-equation model, and the dynamic single-equation model for which estimation results are reported in Section 4.

1. The static, single-equation random effects model

The static single-equation model we consider can be written compactly as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i = \boldsymbol{e}_T\alpha_i + \boldsymbol{u}_i \ \sim \ \mathsf{IID}(\boldsymbol{0}, \boldsymbol{\Omega}), \quad i = 1, \ldots, N, \tag{A.1}$$

$$\boldsymbol{\Omega} = \sigma_\alpha^2 \boldsymbol{e}_T\boldsymbol{e}_T' + \sigma_u^2\boldsymbol{I}_T = \sigma_u^2\boldsymbol{K}_T + (\sigma_u^2 + T\sigma_\alpha^2)\boldsymbol{J}_T, \tag{A.2}$$

where $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})'$ is the $(T\times 1)$ vector of observations on the endogenous variable, $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}', \ldots, \boldsymbol{x}_{iT}')'$ is the $(T \times K)$ vector of observations on the K exogenous variables, α_i is the latent individual effect (with variance σ_α^2), $\boldsymbol{u}_i$ and $\boldsymbol{\epsilon}_i$ are, respectively, the genuine and 'gross disturbance' $(T \times 1)$ vectors (with covariance matrices $\sigma_u^2\boldsymbol{I}_T$ and $\boldsymbol{\Omega}$, respectively), $\boldsymbol{\beta}$ is the $(K \times 1)$ coefficient vector, $\boldsymbol{e}_T$ is the $(T \times 1)$ vector of ones, $\boldsymbol{I}_T$ is the T dimensional identity matrix, $\boldsymbol{J}_T = (\boldsymbol{e}_T\boldsymbol{e}_T')/T$, and $\boldsymbol{K}_T = \boldsymbol{I}_T - \boldsymbol{J}_T$. The two latter matrices are orthogonal and idempotent, so that $\boldsymbol{\Omega}^{-1} = \boldsymbol{K}_T/\sigma_u^2 + \boldsymbol{J}_T/(\sigma_u^2 + T\sigma_\alpha^2)$. The GLS estimator of $\boldsymbol{\beta}$ is obtained by minimizing $\sum_{i=1}^N(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$, giving [see Hsiao (1986, section 3.3.2)]

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{GLS} &= \left(\textstyle\sum_{i=1}^N \boldsymbol{X}_i'\boldsymbol{\Omega}^{-1}\boldsymbol{X}_i\right)^{-1}\left(\textstyle\sum_{i=1}^N \boldsymbol{X}_i'\boldsymbol{\Omega}^{-1}\boldsymbol{y}_i\right) \\ &= (\boldsymbol{W}_{XX} + \theta\boldsymbol{B}_{XX})^{-1}(\boldsymbol{W}_{XY} + \theta\boldsymbol{B}_{XY}),\end{aligned} \tag{A.3}$$

where $\theta = \sigma_u^2/(\sigma_u^2 + T\sigma_\alpha^2)$, when we use the following notation for the within individual and between individual covariance matrices of two arbitrary panel data vectors $\boldsymbol{z}_{it}$ and $\boldsymbol{q}_{it}$

$$\boldsymbol{W}_{ZQ} = \textstyle\sum_{i=1}^N\sum_{t=1}^T(\boldsymbol{z}_{it} - \bar{\boldsymbol{z}}_{i\cdot})'(\boldsymbol{q}_{it} - \bar{\boldsymbol{q}}_{i\cdot}), \qquad \boldsymbol{B}_{ZQ} = T\textstyle\sum_{i=1}^N(\bar{\boldsymbol{z}}_{i\cdot} - \bar{\boldsymbol{z}})'(\bar{\boldsymbol{q}}_{i\cdot} - \bar{\boldsymbol{q}}),$$

$\bar{\boldsymbol{z}}_{i\cdot}$ and $\bar{\boldsymbol{q}}_{i\cdot}$ being individual mean vectors and $\bar{\boldsymbol{z}}$ and $\bar{\boldsymbol{q}}$ corresponding global mean vectors. The covariance matrix of $\widehat{\boldsymbol{\beta}}_{GLS}$ is

$$\mathsf{V}(\widehat{\boldsymbol{\beta}}_{GLS}) = \left(\textstyle\sum_{i=1}^N \boldsymbol{X}_i'\boldsymbol{\Omega}^{-1}\boldsymbol{X}_i\right)^{-1} = \sigma_u^2(\boldsymbol{W}_{XX} + \theta\boldsymbol{B}_{XX})^{-1}. \tag{A.4}$$

'Model 2' in Table 5 is based on this estimator.

Consider the following more general estimator of $\boldsymbol{\beta}$ [see Biørn (2001, section 2.2)]:

$$\widehat{\boldsymbol{\beta}} = [\lambda_W \boldsymbol{W}_{XX} + \lambda_B \boldsymbol{B}_{XX}]^{-1} [\lambda_W \boldsymbol{W}_{XY} + \lambda_B \boldsymbol{B}_{XY}], \tag{A.5}$$

where λ_W and λ_B are non-negative weights. The GLS, the OLS, and the within individual estimators are obtained for (λ_W, λ_B) equal to $(1, \theta)$, $(1, 1)$, and $(1, 0)$, respectively. Its covariance matrix is

$$\begin{aligned}\mathsf{V}(\widehat{\boldsymbol{\beta}}) = [\lambda_W \boldsymbol{W}_{XX} + \lambda_B \boldsymbol{B}_{XX}]^{-1} [\lambda_W^2 \sigma^2 \boldsymbol{W}_{XX} + \lambda_B^2 (\sigma^2 + T\sigma_\alpha^2) \boldsymbol{B}_{XX}] \\ \times [\lambda_W \boldsymbol{W}_{XX} + \lambda_B \boldsymbol{B}_{XX}]^{-1}.\end{aligned} \tag{A.6}$$

This expression with $\lambda_W = \lambda_B = 1$ is used in calculating the 'corrected' standard error estimates of the OLS estimates denoted as 'Model 1' in Table 5.

2. The static, multi-equation random effects model

The static multi-equation model with G equations (in our case, $G = 2$) can be written compactly as

$$\boldsymbol{y}_i = \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i = \boldsymbol{e}_T \otimes \boldsymbol{\alpha}_i + \boldsymbol{u}_i \ \sim \ \mathsf{IID}(\boldsymbol{0}, \boldsymbol{\Omega}), \quad i = 1, \ldots, N, \tag{A.7}$$

$$\boldsymbol{\Omega} = (\boldsymbol{e}_T \boldsymbol{e}_T') \otimes \boldsymbol{\Sigma}_\alpha + \boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u = \boldsymbol{K}_T \otimes \boldsymbol{\Sigma}_u + \boldsymbol{J}_T \otimes (\boldsymbol{\Sigma}_u + T\boldsymbol{\Sigma}_\alpha), \tag{A.8}$$

where $\otimes$ is the Kronecker product operator, $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}', \ldots, \boldsymbol{y}_{iT}')'$ is the $(TG \times 1)$ vector of observations on the G endogenous variables [$\boldsymbol{y}_{it}$ denoting the $(G \times 1)$ vector of the endogenous variables from individual i in period t], $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}', \ldots, \boldsymbol{X}_{iT}')'$ is the $(TG \times K)$ matrix of observations on the K exogenous variables [$\boldsymbol{X}_{it}$ denoting the $(G \times K)$ matrix of the exogenous variables from individual i in period t], $\boldsymbol{\alpha}_i$ is the $(G \times 1)$ vector of latent individual effects (with covariance matrix $\boldsymbol{\Sigma}_\alpha$), $\boldsymbol{u}_i$ and $\boldsymbol{\epsilon}_i$ are, respectively, the genuine and 'gross disturbance' $(TG \times 1)$ vectors (with covariance matrices $\boldsymbol{I}_T \otimes \boldsymbol{\Sigma}_u$ and $\boldsymbol{\Omega}$, respectively), $\boldsymbol{\beta}$ is the $(KG \times 1)$ coefficient vector and the other symbols are defined as above. The inverse of $\boldsymbol{\Omega}$ can, by exploiting the properties of $\boldsymbol{J}_T$ and $\boldsymbol{K}_T$, be written as $\boldsymbol{\Omega}^{-1} = \boldsymbol{K}_T \otimes \boldsymbol{\Sigma}_u^{-1} + \boldsymbol{J}_T \otimes (\boldsymbol{\Sigma}_u + T\boldsymbol{\Sigma}_\alpha)^{-1}$. The GLS estimator of $\boldsymbol{\beta}$ is obtained by minimizing $\sum_{i=1}^N (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta})' \boldsymbol{\Omega}^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta})$, giving $\widehat{\boldsymbol{\beta}}_{GLS} = \left(\sum_{i=1}^N \boldsymbol{X}_i' \boldsymbol{\Omega}^{-1} \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^N \boldsymbol{X}_i' \boldsymbol{\Omega}^{-1} \boldsymbol{y}_i \right)$,

with covariance matrix $\mathsf{V}(\widehat{\boldsymbol{\beta}}_{GLS}) = \left(\sum_{i=1}^{N} \boldsymbol{X}_i'\boldsymbol{\Omega}^{-1}\boldsymbol{X}_i\right)^{-1}$. These expressions can be rewritten as

$$\begin{aligned}(\text{A.9}) \qquad \widehat{\boldsymbol{\beta}}_{GLS} = &\left[\sum_{i=1}^{N} \boldsymbol{X}_i'[\boldsymbol{K}_T \otimes \boldsymbol{\Sigma}_u^{-1}]\boldsymbol{X}_i + \sum_{i=1}^{N} \boldsymbol{X}_i'[\boldsymbol{J}_T \otimes (\boldsymbol{\Sigma}_u + T\boldsymbol{\Sigma}_\alpha)^{-1}]\boldsymbol{X}_i\right]^{-1} \\ &\times \left[\sum_{i=1}^{N} \boldsymbol{X}_i'[\boldsymbol{K}_T \otimes \boldsymbol{\Sigma}_u^{-1}]\boldsymbol{y}_i + \sum_{i=1}^{N} \boldsymbol{X}_i'[\boldsymbol{J}_T \otimes (\boldsymbol{\Sigma}_u + T\boldsymbol{\Sigma}_\alpha)^{-1}]\boldsymbol{y}_i\right],\end{aligned}$$

$$(\text{A.10}) \qquad \mathsf{V}(\widehat{\boldsymbol{\beta}}_{GLS}) = \left[\sum_{i=1}^{N} \boldsymbol{X}_i'[\boldsymbol{K}_T \otimes \boldsymbol{\Sigma}_u^{-1}]\boldsymbol{X}_i + \sum_{i=1}^{N} \boldsymbol{X}_i'[\boldsymbol{J}_T \otimes (\boldsymbol{\Sigma}_u + T\boldsymbol{\Sigma}_\alpha)^{-1}]\boldsymbol{X}_i\right]^{-1}.$$

'Model 5' in Table 5 is based on this estimator.

3. The dynamic, single-equation model

The autoregressive single equation model has the form

$$(\text{A.11}) \quad y_{it} = \alpha_i^* + y_{i,t-1}\lambda + \boldsymbol{x}_{it}\boldsymbol{\beta} + u_{it}, \quad |\lambda| < 1, \quad u_{it} \sim \mathsf{IID}(0, \sigma_u^2), \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 2, \ldots, T, \end{array}$$

where α_i^* is an individual effect (fixed or random), including the intercept. Taking one period differences in order to eliminate α_i^*, we get

$$(\text{A.12}) \qquad \Delta y_{it} = \Delta y_{i,t-1}\lambda + \Delta \boldsymbol{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, \qquad \begin{array}{l} i = 1, \ldots, N, \\ t = 3, \ldots, T. \end{array}$$

Solving this equation recursively back to time $-\infty$ we get, since $|\lambda| < 1$,

$$\Delta y_{it} = \textstyle\sum_{s=0}^{\infty} \lambda^s(\Delta \boldsymbol{x}_{i,t-s}\,\boldsymbol{\beta} + \Delta u_{i,t-s}),$$

which shows that Δy_{it} is correlated with $\Delta \boldsymbol{x}_{i,t-s}$ and $\Delta u_{i,t-s}$, $s = 0, 1, 2, \ldots$.

OLS applied on on (A.12) is inconsistent, since $\Delta y_{i,t-1}$ and Δu_{it} are correlated. However, the following orthogonality conditions are valid:

$$\begin{aligned} &\mathsf{E}(y_{i,t-\tau}\Delta u_{it}) = \mathsf{E}(\Delta y_{i,t-\tau}\Delta u_{it}) = 0, \quad \text{for all } i \text{ and } t \text{ and } \tau \geq 2, \\ &\mathsf{E}(\boldsymbol{x}_{i\theta}'\Delta u_{it}) = \mathsf{E}(\Delta \boldsymbol{x}_{i\theta}'\Delta u_{it}) = \boldsymbol{0}_{K,1}, \quad \text{for all } i, t \text{ and } \theta. \end{aligned}$$

This suggests a large number of potential instruments for $(\Delta y_{i,t-1}, \Delta \boldsymbol{x}_{it})$ which can be used within the framework of GMM procedures. We formally consider (A.12) for $t = 3, \ldots, T$ as a system of $T-2$ equations, each having N observations, and write it as

$$(\text{A.13}) \qquad \begin{array}{l} \Delta y_{i3} = \Delta y_{i2}\lambda + \Delta \boldsymbol{x}_{i3}\boldsymbol{\beta} + \Delta u_{i3} = (\Delta y_{i2}, \Delta \boldsymbol{x}_{i3})\boldsymbol{\delta} + \Delta u_{i3}, \\ \Delta y_{i4} = \Delta y_{i3}\lambda + \Delta \boldsymbol{x}_{i4}\boldsymbol{\beta} + \Delta u_{i4} = (\Delta y_{i3}, \Delta \boldsymbol{x}_{i4})\boldsymbol{\delta} + \Delta u_{i4}, \\ \vdots \\ \Delta y_{iT} = \Delta y_{i,T-1}\lambda + \Delta \boldsymbol{x}_{iT}\boldsymbol{\beta} + \Delta u_{iT} = (\Delta y_{i,T-1}, \Delta \boldsymbol{x}_{iT})\boldsymbol{\delta} + \Delta u_{iT}, \end{array} \qquad i = 1, \ldots, N,$$

where $\boldsymbol{\delta} = (\lambda, \boldsymbol{\beta}')'$, or compactly, in obvious notation, as

$$\Delta \boldsymbol{q}_i = (\Delta \boldsymbol{W}_i)\boldsymbol{\delta} + \boldsymbol{u}_i, \qquad i = 1, \ldots, N. \tag{A.14}$$

Our GMM procedure, in two alternatives, can be described as follows:

Alternative A: Limited IV set. In the system (A.13) we use

$\boldsymbol{z}_{i1} = (y_{i1}, \Delta \boldsymbol{x}_{i3})$ as IV matrix for $(\Delta y_{i2}, \Delta \boldsymbol{x}_{i3})$ in the first equation,
$\boldsymbol{z}_{i2} = (y_{i2}, \Delta \boldsymbol{x}_{i4})$ as IV matrix for $(\Delta y_{i3}, \Delta \boldsymbol{x}_{i4})$ in the second equation,
$\vdots$
$\boldsymbol{z}_{i,T-2} = (y_{i,T-2}, \Delta \boldsymbol{x}_{iT})$ as IV matrix for $(\Delta y_{i,T-1}, \Delta \boldsymbol{x}_{iT})$ in the $(T-2)$'th equation,

Alternative B: Extended IV set. In the system (A.13) we use

$\boldsymbol{z}_{i1} = (y_{i1}, \Delta \boldsymbol{x}_{i3})$ as IV matrix for $(\Delta y_{i2}, \Delta \boldsymbol{x}_{i3})$ in the first equation,
$\boldsymbol{z}_{i2} = (\boldsymbol{z}_{i1}, y_{i2}, \Delta \boldsymbol{x}_{i4})$ as IV matrix for $(\Delta y_{i3}, \Delta \boldsymbol{x}_{i4})$ in the second equation,
$\vdots$
$\boldsymbol{z}_{i,T-2} = (\boldsymbol{z}_{i,T-3}, y_{i,T-2}, \Delta \boldsymbol{x}_{iT})$ as IV matrix for $(\Delta y_{i,T-1}, \Delta \boldsymbol{x}_{iT})$ in the $(T-2)$'th equation.

Both alternatives, of which B has a substantially larger number of orthogonality conditions than A, were considered. They gave fairly similar coefficient and standard error estimates (although the latter are somewhat lower for B), indicating redundance (or near redundance) of several orthogonality conditions. In view of this, and some problems of numerical instability in Alternative B, we decided to stick to Alternative A, although with lagged differences in the ABF dummy excluded from the IV set.

We define the composite IV matrix for (A.14) as

$$\boldsymbol{Z}_i = \begin{bmatrix} \boldsymbol{z}_{i1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{z}_{i2} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{z}_{i,T-2} \end{bmatrix}, \qquad i = 1, \ldots, N. \tag{A.15}$$

The first stage GMM estimator of $\boldsymbol{\delta} = (\lambda, \boldsymbol{\beta}')'$ can then be written as

$$\begin{aligned} \widehat{\boldsymbol{\delta}} \;=\; & \left[\left(\textstyle\sum_{i=1}^N (\Delta \boldsymbol{W}_i)' \boldsymbol{Z}_i \right) \left(\textstyle\sum_{i=1}^N \boldsymbol{Z}_i' \boldsymbol{Z}_i \right)^{-1} \left(\textstyle\sum_{i=1}^N \boldsymbol{Z}_i' (\Delta \boldsymbol{W}_i) \right) \right]^{-1} \\ & \times \left[\left(\textstyle\sum_{i=1}^N (\Delta \boldsymbol{W}_i)' \boldsymbol{Z}_i \right) \left(\textstyle\sum_{i=1}^N \boldsymbol{Z}_i' \boldsymbol{Z}_i \right)^{-1} \left(\textstyle\sum_{i=1}^N \boldsymbol{Z}_i' (\Delta \boldsymbol{q}_i) \right) \right]. \end{aligned} \tag{A.16}$$

'Model 7' in Table 6 is based on this estimator. This estimator represents an optimal way of using the IV set only if the disturbances in (A.12) are homoskedastic. It may be unlikely that this is the case for our application. The second stage GMM estimator can be used to increase efficiency by utilizing the residuals obtained from the first stage estimator to take account of disturbance heteroskedasticity of an unspecified form [see White (1984, sections IV.3 and VI.2)]. Briefly, it can be described as follows: Let $\widehat{\boldsymbol{u}}_i = \Delta \boldsymbol{q}_i - (\Delta \boldsymbol{W}_i)\widehat{\boldsymbol{\delta}}$ be the first stage GMM residuals. The second stage GMM estimator then reads:

$$\text{(A.17)} \qquad \widetilde{\boldsymbol{\delta}} = \left[\left(\sum_{i=1}^{N}(\Delta \boldsymbol{W}_i)'\boldsymbol{Z}_i\right)\left(\sum_{i=1}^{N}\boldsymbol{Z}_i'\widehat{\boldsymbol{u}}_i\widehat{\boldsymbol{u}}_i'\boldsymbol{Z}_i\right)^{-1}\left(\sum_{i=1}^{N}\boldsymbol{Z}_i'(\Delta \boldsymbol{W}_i)\right)\right]^{-1} \times \left[\left(\sum_{i=1}^{N}(\Delta \boldsymbol{W}_i)'\boldsymbol{Z}_i\right)\left(\sum_{i=1}^{N}\boldsymbol{Z}_i'\widehat{\boldsymbol{u}}_i\widehat{\boldsymbol{u}}_i'\boldsymbol{Z}_i\right)^{-1}\left(\sum_{i=1}^{N}\boldsymbol{Z}_i'(\Delta \boldsymbol{q}_i)\right)\right].$$

'Model 8' in Table 6 is based on this estimator.

The estimated covariance matrix of $\widetilde{\boldsymbol{\delta}}$ is

$$\text{(A.18)} \qquad \widehat{\mathsf{V}(\widetilde{\boldsymbol{\delta}})} = \left[\left(\sum_{i=1}^{N}(\Delta \boldsymbol{W}_i)'\boldsymbol{Z}_i\right)\left(\sum_{i=1}^{N}\boldsymbol{Z}_i'\widehat{\boldsymbol{u}}_i\widehat{\boldsymbol{u}}_i'\boldsymbol{Z}_i\right)^{-1}\left(\sum_{i=1}^{N}\boldsymbol{Z}_i'(\Delta \boldsymbol{W}_i)\right)\right]^{-1}$$

(the expression for the covariance matrix of $\widehat{\boldsymbol{\delta}}$ is more complicated and is not reported here) and the test statistic for the Sargan-Hansen test for the validity of the orthogonality conditions is [cf. Newey (1985) and Arellano and Bond (1991)]

$$\text{(A.19)} \qquad J = \left[\left(\sum_{i=1}^{N}\widehat{\boldsymbol{u}}_i'\boldsymbol{Z}_i\right)\left(\sum_{i=1}^{N}\boldsymbol{Z}_i'\widehat{\boldsymbol{u}}_i\widehat{\boldsymbol{u}}_i'\boldsymbol{Z}_i\right)^{-1}\left(\sum_{i=1}^{N}\boldsymbol{Z}_i'\widehat{\boldsymbol{u}}_i\right)\right]^{-1}.$$

Under the null hypothesis, it is asymptotically distributed as χ^2 with a number of degrees of freedom equal to the number of orthogonality conditions. The orthogonality conditions are rejected when J exceeds an appropriate quantile in this χ^2 distribution.