UNIVERSITY OF OSLO

HEALTH ECONOMICS RESEARCH PROGRAMME

Errors in Survey Based Quality Evaluation Variables in Efficiency Models of Primary Care Physicians

Sverre A.C. Kittelsen

Guri Galtung Kjæserud

The Ragnar Frisch Centre for Economic Research

Odd Jarle Kvamme

Department of General Practice, University of Oslo

Working Paper 2001: 12

Errors in Survey Based Quality Evaluation Variables in Efficiency Models of Primary Care Physicians[†]

Sverre A.C. Kittelsen, Dr.Polit. [††] , **Guri Galtung Kjæserud, Cand.Oecon.,**
Frisch Centre, and HERO – Health Economic Research Programme at the University of Oslo,

Odd Jarle Kvamme, Dr.Med, G.P.
Department of General Practice, University of Oslo

Health Economics Research Programme at the University of Oslo

HERO 2001

JEL Classification: C61, D24, I12

Keywords: DEA, Health economics, Quality, Patient evaluation, Efficiency, Errors in variables, Resampling, Bootstrap, Selection bias, Sampling error.

[†] This work is supported by the Norwegian Research Council through "*HERO – Health economics research programme at the University of Oslo"*, and the project "*Quality and efficiency analysis in the public sector*". Thanks to Per Hjortdahl and Leif Sandvik for letting us use data from EUROPEP (European patient expectations and priorities). The use of data and interpretation of results are the responsibility of the authors alone.

[††] Corresponding author: Sverre A.C. Kittelsen, Frisch Centre, Gaustadalléen 21, N-0349 Oslo, Norway. Email: s.a.c.kittelsen@frisch.uio.no. Tel +47-22958815, Fax +47-22958815.

HERO - Health Economics Research Programme at the University of Oslo
Financial support from The Research Council of Norway is acknowledged.
ISSN 1501-9071, ISBN 82-7756-075-3

Abstract

Efficiency analyses in the health care sector are often criticised for not incorporating quality variables. The definition of quality of primary health care has many aspects, and it is inevitably also a question of the patients' perception of the services received. This paper uses variables derived from patient evaluation surveys as measures of the quality of the production of health care services. It uses statistical tests to judge if such measures have a significant impact on the use of resources in various Data Envelopment Analysis (DEA) models. As the use of survey data implies that the quality variables are measured with error, the assumptions underlying a DEA model are not strictly fulfilled. This paper focuses on ways of correcting for biases that might result from the violation of selected assumptions. Firstly, any selection bias in the patient mix of each physician is controlled for by regressing the patient evaluation responses on the patient characteristics.

The corrected quality evaluation variables are entered as outputs in the DEA model, and model specification tests indicate that out of 25 different quality variables, only waiting time has a systematic impact on the efficiency results. Secondly, the effect on the efficiency estimates of the remaining sampling error in the patient sample for each physician is accounted for by constructing confidence intervals based on resampling. Finally, as an alternative approach to including the quality variables in the DEA model, a regression model finds different variables significant, but not always with a trade-of between quality and quantity.

1. Introduction

Primary care physicians or general practitioners (GPs) provide most of the basic services within the Norwegian health care system. The treat larger part of illnesses and diseases in the society, advice on preventive care in general and verifies the need for sick leave and disability payments. Only ten percent of the patient-contacts in primary care are referred to specialist care. In addition, GPs act as gatekeepers for specialised care (hospitals and most private specialist practices). The efficiency of GPs is therefore of primary importance for the efficiency of the health care system as a whole.

As for the rest of the health care sector the heterogeneous products and asymmetric information have encouraged public control with the provision and organisation of primary health care. The lack of specific product evaluations and prices also implies that efficiency evaluations to a large extent need to be done by comparing the output quantities rather than their value to the amount or value of the resources used, by estimating technical or cost efficiency.

Many health care researchers have recognised that the non-parametric efficiency measurement methods are well suited to estimating efficiency within this framework, since these methods can easily handle multiple inputs and multiple outputs simultaneously without reference to prices, and do not need restrictive assumptions on the functional form of the production possibility set or on the distribution of efficiency. The most common non-parametric method known as data envelopment analysis (DEA) was suggested by Farrell (1957) and developed in a large body of literature following Charnes, Cooper & Rhodes (1981) who gave the method its name. The method is shown by Banker (1984) to be the minimum extrapolation estimate that satisfies a) convexity, b) free disposal and c) feasibility.

There are however some important drawbacks to the DEA method. In its basic form DEA is deterministic in the sense that actual behaviour is assumed to be observable without error. Recent developments initiated in a series of articles by Banker (1996, 1993), Kneip, Park & Simar (1996) and Simar (1996), have however given DEA a statistical foundation by assuming that the observations are drawn from an underlying possibility set whose properties can then be estimated. Asymptotic tests of model specification are thus available, and bootstrap methods have been developed by Simar & Wilson (1998) that can give confidence bands for the frontier of the possibility set. The possibility of measurement error in the observed variables in the

model is still not adequately taken account of in the DEA method, which implicitly assumes that all deviation from the frontier is inefficiency.

An ongoing research topic is also how to model quality within the DEA framework, which requires that all variables are either inputs or outputs, and are available as cardinal quantitative measures. Guiffrida defines the "final outcome of the primary care provided" as the "health improvement of the population served, measured for instance, in quality adjusted life years (QALYs)" (A. Guiffrida, 1998, p. 17). In most empirical cases, the change in QALYs caused by the care provided by the GPs will not be available. To approximate the outcomes by the GPs one can focus on the intermediate products such as the number of consultations. To express the quality of the service provided one can use supplementary outcome measures. In DEA the quality aspects must be formulated as equivalent with products, but in such a way as to retain the plausibility of the convexity assumptions in the method. Petersen & Olesen (1995) discusses ways ordinal quality measures can be incorporated in DEA.

The survey data used in this study makes available variables for a sample of GPs, together with patient quality evaluation responses from a sample of each physician's patients. The perceived quality of service is measured on a scale from one to five, facilitating a cardinal interpretation of the results. The approach used in this paper is therefore to use as quality variables the total quality evaluation score defined as the product of the average quality evaluation score and the number of consultations. These quality variables are then entered as outputs in the DEA model.

While recognising that patient quality evaluation score does not measure the quality or health outcomes of the GPs production directly, we would argue that it captures important quality aspects for two reasons. Firstly, some of the questions pertain to health outcomes, as the patients perceive them. Secondly, patient satisfaction is a quality aspect in itself; for the same health outcome a satisfied patient is of greater value than a dissatisfied one. While the patients are not directly asked about their satisfaction, their evaluation of quality will to some extent be coloured by the importance they attribute to the quality aspect in question.

Unlike most DEA studies, we have in this study direct information on the error structure of the quality variables. The scalar quality measures for each GP is an estimate derived from a sample of patients for that GP. The main methodological contribution of this paper is therefore to show how this error structure can be accounted for. Firstly, there is selection bias in average quality evaluation score stemming from the fact that GPs have a different mix of patients. Secondly,

there is a sampling error due to the fact that only a subsample of each GP's patients has been surveyed.

Section 2 presents the available data. In section 3 the assumed data generating process is described and the DEA method briefly presented, and in section 4 the method for correcting for selection bias is discussed. Section 5 uses the available statistical hypothesis tests to determine model specification and presents the results of the basic DEA model. Section 6 shows the resampling method used to estimate the confidence intervals for the efficiency estimates and the results of this analysis. Section 7 seeks to establish some determinants of efficiency by TOBIT estimation. In section 8 an alternative way of taking account of the quality variables through regression analysis is presented, while section 9 concludes.

2. Data

The data has been collected from a stratified sample of 60 GPs in Norway in 1998 as part of the European Patient Expectations and Priorities (EUROPEP) research programme (O.J. Kvamme and P. Hjortdahl, 1997, O.J. Kvamme et al., 2000). Table A.3 in the appendix lists the questions that each GP was asked (Q1-Q13), and includes information on the number of hours used by the GPs on consultations and other tasks, the number of consultations, the size of the practice and the number of support staff, as well as questions relating to organisation and finance.

In addition, each GP were to distribute a patient questionnaire to 40 of their patients. These were asked to return the questionnaire directly to the research group, and were reminded once. The patient questionnaire consisted of two parts, a set of evaluation questions and a list of personal characteristics.

The response rate for the GP's was 100%, but on average only 30 patients were given questionnaires. Of these less than 90% responded. After elimination of outliers, only 52 GPs remain in the sample with an average of 27.6 patients each. The sample is clearly to small to give any exact estimate of the level of GPs efficiency, but should be large enough to uncover structural features of their production function, including what quality variables significantly influences resource usage.

Table A1 in the appendix lists the quality evaluation questions (A1-A25), where the patient was asked to respond on a scale from one (worst) to five (best).

For the answers to reveal true differences between the GPs average quality level, the average patient evaluation should not just be random noise but be significantly different across GPs. The last two columns of the table show the results of an ANOVA analysis which show that for all variables the difference between the GPs is significant at the 5% level, and for all except one at the 1% level.

Table A.2 in the appendix shows the mean values of the personal characteristics of the patients. Again there is a significant difference in the patient mix across GPs, implying a potential selection bias if patient evaluation is influenced by these personal characteristics.

3. DEA model

The individual GP is assumed to face a given technology or production possibility set P , in the sense that

$$P = \left\{ (X, Y, S) \middle| Y, S \text{ can be produced with } X \right\} \tag{1}$$

where X is a vector of inputs, Y is a scalar of the quantitative output (number of consultation) and S is a vector of total qualitative outputs or product aspects. In parallel with the data generating process suggested by Kneip, Park & Simar (1996), each GP chooses independently and from the same distributions, the input levels and an output mix, which in this case can be formulated as an average quality evaluation level $s^l = S^l / Y$ in each real quality dimension. Conditional on this choice of inputs and quality (X,s) the GP chooses an output efficiency level, E_2 that is the ratio of actual to maximal feasible output of the quantitative product $Y = E_2 \overline{Y}$. Since each qualitative variable $S^l = s^l Y$ is proportional to Y, this implies that the resulting maximal output levels in all must be feasible in all dimensions $(X, \overline{Y}, \overline{S}) = (X, Y/E_2, S/E_2) \in P$. It can be seen that the chosen efficiency level can be defined by the Farrell (1957) technical output efficiency

$$E_2 = Min\left\{\theta \left| \left(X, \frac{Y}{\theta}, \frac{S}{\theta} \right) \in P \right.\right\} \tag{2}$$

With the additional assumption that the density of efficiency is such that one will observe points arbitrarily close to the frontier when the number of observations is sufficiently large, the DEA estimate of the technology

$$\hat{P}^{DEA} = \left\{ (X,Y,S) \left| X \geq \sum_{j \in N} \lambda_j X_j, Y \leq \sum_{j \in N} \lambda_j Y_j, S \leq \sum_{j \in N} \lambda_j S_j, \sum_{j \in N} \lambda_j = 1 \right.\right\} \tag{3}$$

can be shown to be consistent. The DEA estimate of the Farrell (1957) output technical efficiency is then simply

$$\hat{E}_2 = Min\left\{\theta \left| \left(X, \frac{Y}{\theta}, \frac{S}{\theta} \right) \in \hat{P}^{DEA} \right.\right\} \tag{4}$$

which can be solved by linear programming. The formulation in (3) and (4) is equivalent to the variable returns to scale (VRS) DEA model suggested by Banker, Charnes & Cooper (1984). Banker (1993) suggests several asymptotic tests for model specification based on comparing the distributions of the estimated efficiencies in the different models, with the null hypothesis that these are equal. Kittelsen (1999) evaluates these in Monte Carlo simulations and finds that they give crude but usable approximations of the true significance levels and power functions. In this paper we use the Kolmogorov-Smirnov D^+ test of one-sided hypothesis (N.J Johnson et al., 1994) that is conservative but usable in small samples, as well as the ordinary T-test for comparisons of group means (G.K. Bhattacharyya and R.A. Johnson, 1977) that has more power but tends to over reject the null in small samples. The tests are used to choose the scale assumption (VRS or CRS), and to select the variables to include in the DEA-model. In particular it is necessary to restrict the set of quality evaluation variables to those that have a significant impact on efficiency and resource use.

4. Selection bias

In the ANOVA analysis shown in table A.2 it was demonstrated that the patient mix was significantly different across GPs. To the extent that these characteristics in part determine how

satisfied a patient is, the average quality evaluation score of the patients $\bar{a}_i^l = \sum_{j=1}^{n_i} a_{ij}^l \Big/ n_i$ of a given GP *i* for each of the quality variables *l=1..25,* will be biased by his/her patient mix. For example, if female patients are more easily satisfied than male patients, then the average quality evaluation for GP's with many female patients will be higher than for GPs with few female patients, even though the target average quality level s_i^l is the same. Given a target average quality level, we will specify a linear relationship between the reported quality evaluation level of a patient and the personal characteristics of that patient,

$$a_{ij}^l = s_i^l + \sum_{k=1}^{K} \beta_k^l (b_{ij}^k - \bar{b}^k) + u_{ij}^l \tag{5}$$

where b_{ij}^k is the characteristic *k* for patient *j* with GP *i,* and u_{ij}^l is a random error term. As formulated the second term in (5) captures the characteristics deviation from the mean characteristic in the sample $\bar{b}^k$, so that s_i^l is the target quality level for an "average person". We estimate (5) by 25 OLS regressions with GP dummies d_i (leaving out *i=1*),

$$a_{ij}^l = \hat{\alpha}_1^l + \sum_{i=2}^{N} \hat{\alpha}_i^l d_i + \sum_{k=1}^{K} \hat{\beta}_k^l (b_{ij}^k - \bar{b}^k) + \hat{u}_{ij}^l \tag{6}$$

The estimate of the target quality levels and the total quality for each GP and type of quality is then given by

$$\hat{s}_i^l = \hat{\alpha}_1^l + \hat{\alpha}_i^l, \quad \hat{S}_i^l = \hat{s}_i^l Y_i \tag{7}$$

which is used as outputs for the GPs in the DEA model.

Since the focus of the analysis is on the correction of the quality evaluation score averages rather than on the patient characteristics coefficients themselves, the specification of each of the 25 regressions was not evaluated individually. Nevertheless it is worth noting that the patient characteristics coefficients were significant as a block in all regressions, as were most of the individual coefficients. In this sense the quality level used for the GPs were significantly corrected by this procedure.

Table 1: Regression of dependent variable a22 *"patient evaluation of physician waiting time"* on N-1=51 physician dummies and K=6 individual characteristics for 1361 patients. Introducing patient characteristics in the regression in addition to physician dummies raises R^2 from 0.135 to 0.282 and adjusted R^2 from 0.131 to 0.251. The number of stars** and *** corresponds to the significance levels 5% and 1%.

Code	Variable names	B	Std. Error	Sig.
const		3.067	0.161	***0.000
BD1D	Gender (0=female, 1=male)	-0.184	0.064	***0.004
BD2	Year of birth	-0.024	0.002	***0.000
BD3	Highest completed education	0.001	0.032	0.964
BD4	Number of physicians visits last 12 months	0.018	0.007	***0.007
BD5	Evaluation of own health status	-0.062	0.031	**0.046
BD6D	Presence of serious disease (0=no, 1=yes)	-0.032	0.069	0.645

As an example, table 1 gives the patient coefficient estimates for one of the 25 regressions of the form (5). These coefficients can be interpreted as the marginal effect on the average quality evaluation of waiting time on the scale from one to five of a patient characteristic, controlling for which GP the patient has seen. Male patients and younger patients tend to be less satisfied with waiting time, while the frequency of visits tends to increase evaluation levels.

5. Basic DEA results

While we have available only one non-quality output, there are 25 potential quality outputs and three potential inputs. To include them all in a DEA model with 52 units would not be informational, since it is a feature of the DEA methodology that a large number of variables combined with a small number of observations biases efficiency estimates upward, including estimating all those units that have the largest level on any output variable fully output efficient. Instead we test the model specification by means of the statistical tests described in section 3. Since we at the outset have a small sample we will want to reject a small model easily, and choose to accept all alternative variable and scale assumptions if the significance level is less than 10% on either the D^+ test or the T test.

Table 2 summarises the hypothesis tree. As a basic model we choose to include the number of hours worked treating patients as an input and the number of consultations as an output. In step 1 the possible inclusion of additional inputs is tested, but neither the number of other employees nor the GPs use of time for other purposes than treating patients were significant.

In step 2, all 25 quality variables were candidates, but only the two most significant are listed in the table. None of the quality variables have a very strong influence on resource usage, but the evaluation of the waiting time is significant at the 10% level with the T-test. This variable is therefore included as an output in the DEA model. In the next step the remaining 24 quality variables were candidates to be included in addition to the waiting time variable, but all were clearly insignificant. Finally, the scale assumption was tested, with strong support for variable returns to scale. The resulting DEA model has two outputs, one input and variable returns to scale.

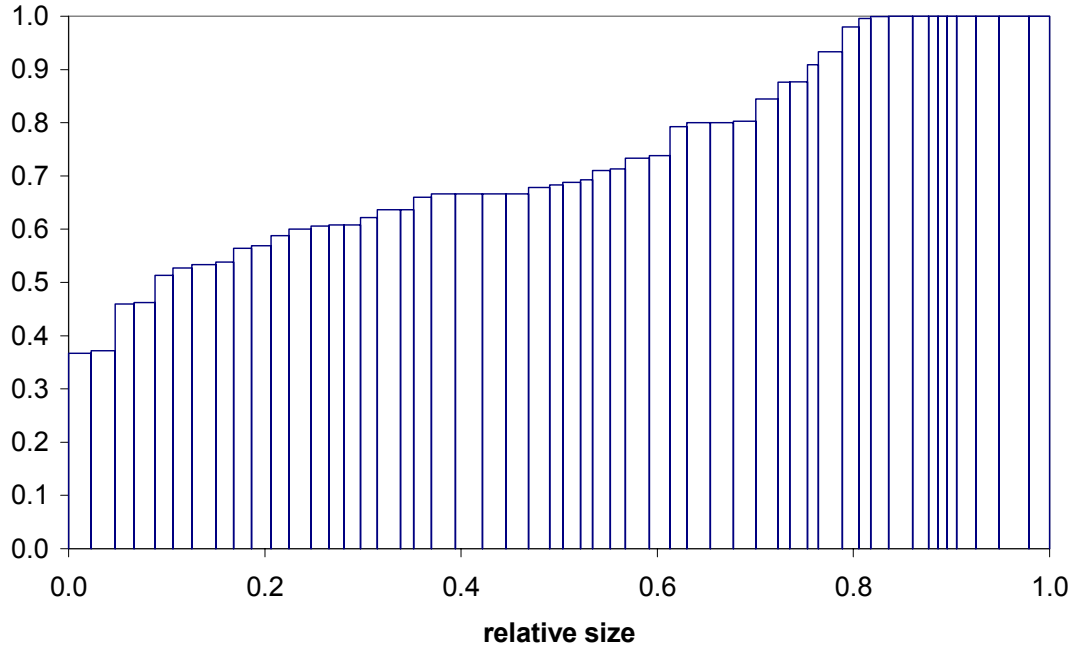
Table 2: Stepwise specification hypothesis tree. In steps 2 and 3 all the corrected patient evaluation variables s1-s25 are entered, but only the first of the insignificant results is shown. The number of stars * and ** corresponds to the significance levels 10% and 5%.

				Critical level	**D+**	**T**
90 %					0.210	1.290
95 %					0.240	1.660
Step		**Alternative**		**Variables/Scale assumptions**		
Basic model	Input		q4	Number of hours per week treating patients		
		Output	q10	Number of consultations per week		
		Scale	CRS	Constant returns to scale		
Step 1 H0: q4,q10,CRS						
		Additional input	q65	Other employees per physician in practice (q6/q5)	0.135	0.150
			q11	Number of hours per week *not* treating patients	0.096	0.391
Step 2 H0: q4,q10,CRS						
		Additional output	s22	The patients evaluation of the physicians waiting time in his practice	0.192	*1.412
			s23	The patients evaluation of the physicians ability to perform help in emergencies	0.173	1.133
Step 3 H0: q4,q10,s22 ,CRS						
		Additional output	s23	The patients evaluation of the physicians ability to perform help in emergencies	0.077	0.261
Step 4 H0: q4,q10,s22, CRS						
		Scale	VRS	Variable returns to scale	*0.212	**1.884
Result (Model I): q4,q10,s22, VRS						

The inference from the hypothesis tests is that only the provision of quality with regards to waiting time has an influence on resource usage that is strong enough to be measurable in a sample of this size. This implies that reducing waiting time is costly, in the sense that to do so requires either an increase in the time spent on patients or a decrease in the number of

consultations. It must be emphasised that the waiting time variable is only measured as important in the production process of the GPs, and does not directly reveal the social evaluation of this quality aspect. Conversely, the insignificance of the other evaluation variables does not imply that they have low social value, only that providing these quality aspects is not measured as costly.

Figure 1: Salter diagram of technical output efficiency Model I.



The main efficiency and productivity estimates are presented in table 3. Figure 1 is a Salter diagram showing the distribution of technical output efficiency as estimated in the preferred DEA model I, where the width of each column is proportionate to the only input. Output efficiency shows a large dispersion, with an average of 74% and a minimum of 37%. While most GPs have an efficiency between 60% and 90%, there is a clear tail of very inefficient GPs. These have fewer consultations per hour of patient contact than other GPs with a comparable quality level as measured by the evaluation of waiting time.

In addition to the preferred model I, results for two other models are also presented in table 3. Model II is equivalent to model I except that the corrected quality measure S22 is replaced by the original sample evaluation measure A22. Results are in fact remarkably similar in the two models, suggesting that while the correction was deemed necessary for the individual GPs'

measure of quality, it had little impact on the average estimates. Model III is the VRS version of the basic specification that did not include any quality measures. The efficiency estimates are on average 3% lower in this model, indicating that including patient evaluation has a noticeable, but not great influence.

Table 3: Main efficiency and productivity results from DEA models. Model I is the model preferred by the specification tests, while results for models II and III are shown for comparison.

	Average	Min	Stdev
Model I: with corrected quality (q4, q10, s22)			
E2 - Output increasing Efficiency	0.738	0.367	0.186
E3 - Productivity	0.673	0.341	0.168
E5 - Pure Scale Efficiency	0.921	0.428	0.109
Model II: with uncorrected quality (q4, q10, a22)			
E2 - Output increasing Efficiency	0.737	0.367	0.186
E3 - Productivity	0.671	0.340	0.168
E5 - Pure Scale Efficiency	0.920	0.428	0.111
Model III: without quality (q4, q10)			
E2 - Output increasing Efficiency	0.708	0.333	0.181
E3 - Productivity	0.628	0.286	0.156
E5 - Pure Scale Efficiency	0.898	0.396	0.107

6. Sampling error

One of the principal drawbacks of the DEA and other nonparametric methods is the inability to take account of measurement error in the variables. In its basic form, DEA is unable to give standard errors or confidence intervals for the efficiency estimates. In addition, it has been claimed that even symmetrically distributed measurement errors in the data may give biased estimates for efficiency, since "good" outlier will potentially affect the position of the frontier, and therefore biased efficiency estimates for the units that are referenced, while "bad" outliers will to a greater extent only affect the efficiency estimate of the outlier itself.

As DEA is a linear programming model, some researchers (e.g. A. Charnes et al., 1985) have used sensitivity analysis from mathematical programming theory to ascertain ranges within which data may be varied without changing the frontier estimate. This is a non-statistical

procedure since it does not use empirical information on the actual variability in the data. In recent developments, Simar & Wilson (1998) have developed bootstrap methods to resample from the distribution of the data in order to bias correct the efficiency measures, to calculate standard errors and confidence intervals, and to use in hypothesis testing. What their method in essence does is to generate distributions of the frontier of the production set, based on the sampling error in the sample of observations. It does not take account of any measurement error in the variables themselves, but uses the original observed variable values for each unit for calculating the efficiencies. In this section we suggest a procedure that is both more ambitious in that it takes account of measurement error in one GP variable, but clearly less ambitious in that it does not simulate the sampling error in the sample of GPs.

In addition to the selection error induced by the fact that GPs have a different mix of patients along the characteristics we have information on, there will be remaining measurement error in the quality evaluation variable stemming from the fact that only a sample of each GPs' patients were questioned. If a different set of patients had been asked, one would have received different responses and calculated a different GP quality evaluation level. While the ordinary average response $\overline{a}_i^{22}$ is an unbiased estimator for the mean evaluation of waiting time for all the patients of GP *i*, this estimate has a standard error. Similarly the correction regression (6) gives us a standard error for each GPs corrected quality level $\hat{s}_i^{22}$. This is a sampling error in the patient samples, but a measurement error in the quality level variable of the GPs. We have therefore additional information on the extent of error in the quality variable $\hat{S}^l$ that is used in the DEA model, and this information should be used to evaluate the extent of error in the resulting efficiency estimates $\hat{E}_2$ from (4).

In parametric analysis, the standard error of derived estimates can be calculated analytically using information about the functional form. In nonparametric methods such as DEA this is not an option, and resampling suggests itself as the natural way of tackling the problem. By taking a large number *B* of draws from the distribution of the evaluation variable $\hat{S}_i^l$ for each GP, and using these in *B* DEA runs, we can generate a resulting distribution of the efficiency estimates that can be used to calculate confidence intervals and standard errors resulting from the measurement error of GPs' quality evaluation level.

Since we will want to retain the correction for selection bias obtained in section 4, the resampling must be done from the distribution of the corrected $\hat{s}_i^{22}$ rather than the original average $\bar{a}_i^{22}$. This could be done by drawing from the t-distribution parameterised by the standard error of $\hat{s}_i^{22}$ estimated in (6) and (7), but the we know that the underlying distribution of the quality evaluation answers with mean 3.4 is very skewed towards the high end of the domain (1,5). We choose instead to redraw with replacement from the empirical distribution of patients' answers, corrected for any effect of patient characteristics. Each patient is assigned a response that is their estimated response had they been an average patient, which is equivalent to their GPs' estimated quality evaluation level plus the patient's individual residual from (6).

$$\tilde{a}_{ij}^{l} = a_{ij}^{l} - \sum_{k=1}^{K} \hat{\beta}_k^{l}(b_{ij}^{k} - \bar{b}^{k}) = \hat{s}_i^{l} + \hat{u}_{ij}^{l}, \quad l = 22 \tag{8}$$

Redrawing with replacement from $\tilde{a}_{ij}^{22}$ for each GP *i* with the same sample size as in the original, gives us *B* new estimates $\hat{\hat{s}}_{ib}^{22}$. These will have an expected value equal to $\hat{s}_i^{22}$ since $E(\hat{u}_{ij}^{22}) = 0$. Rerunning the DEA model *B* times gives us a distribution of *B* efficiency estimates for each GP.

Figure 2: Efficiency estimates with resampled quality evaluation variable s22.

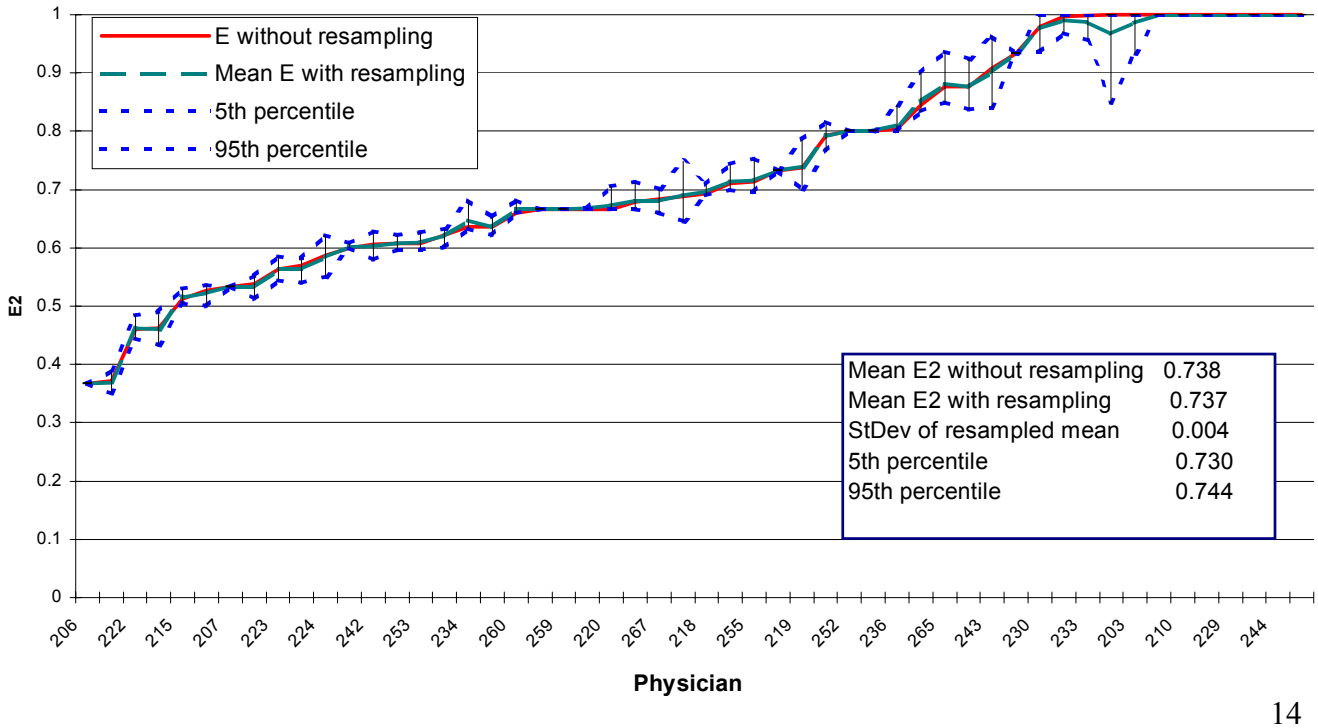


Figure 2 shows the results of the resampled efficiency estimates with B=1000 samples. The mean of the resampled mean efficiencies 73.7% is in fact very close to the original mean estimate 0.738, and the curve shows the efficiency estimates to be close for all the individual GPs. There is therefore no clear bias in the original efficiency estimates due to the measurement error in the quality variable.

There is, however, some uncertainty. The mean efficiency has a standard error of 0.004 as estimated by the standard deviation of the resampled mean efficiencies. Numerically this is small, but is still a noticeable fraction of the 3% impact of the s22 variable found by comparing models I and III in table 3. Graphically, the uncertainty is shown as confidence intervals for each individual GP in the figure. The range of the individual confidence intervals varies from 0% to 15%, but on the whole the ranking of the GPs by efficiency would be little affected even if some were at their upper limit and others at their lower. While the results show errors that are quite small, it must again be emphasised that these confidence intervals are partial since they only take account of measurement error in one of the three variables in the DEA model.

7. Determinants of efficiency

It has become common in the DEA literature to perform a second stage analysis of the efficiency estimates, by regressing these on explanatory variables that do not enter the DEA model. Lovell (1993) recommends that variables that are in the control of the unit itself should be included in the DEA model, and that variables that are exogenous to the unit should be used as independent variables in the second stage to try to explain efficiency. It might be argued that an alternative reason for not including variables in the DEA production model is that some variables are not part of the technological constraints that a production model should capture, but may nevertheless influence realised efficiency through their role in the objective functions of the agents or in the incentives that they face. A third view might be that some variables are to stochastic in their nature (have to much measurement error) to be included in a deterministic model such as DEA, in which case a regression could be seen as an attempt to correct the DEA efficiency estimates.

We will here not attempt to argue that we can explain the efficiency in a causal sense, since we do not specify a behavioural model. Rather we try to use regression methods to estimate the

correlation of estimated efficiency with variables that might be part of an explanation, but could also be the result of the efficiency level or have of common causes as efficiency.

As the distribution of estimated efficiencies is censored from above at the value one, a TOBIT regression model (J. Tobin, 1958) is specified

$$\begin{aligned} E_{2i}^{*} &= E_{2i} = \gamma_0 + \sum_{m=1}^{M} \gamma_m Z_{mi} + u_i \qquad \text{if} \quad \mathrm{E}_{2\mathrm{i}} < 1 \\ E_{2i}^{*} &= 1 \qquad\qquad\qquad\qquad\qquad\quad \text{if} \quad \mathrm{E}_{2\mathrm{i}} \geq 1 \\ u_i &\sim \mathrm{IN}(0, \sigma^2) \end{aligned} \tag{9}$$

where Z_m is shorthand for the available independent variables. The results of the regression are presented in table 4.

Table 4: Result of Tobit regression of technical output efficiency from model I. The number of stars *, ** and *** corresponds to the significance levels 10%, 5% and 1%.

Code	Variable name	Coeff.	Std.Err	Sig.
cons		0.008	0.373	**0.021
Physician personal characteristics				
q1	Physicians age	-0.012	0.006	*0.063
q2	Physicians gender (0=female, 1=male)	0.016	0.042	0.713
q3	Total number of years in practice	0.007	0.007	0.345
q8	Specialist (0=no/under training, 1=yes)	-0.030	0.046	0.512
q9	Number of years in current practice	-0.002	0.003	0.516
q13	Physicians evaluation of his work (1=good,… 5=bad)	-0.028	0.016	*0.083
Physician organizational characteristics				
q4	Number of hours direct patient contact	-0.023	0.003	***0.000
q10	Weekly number of consultations	0.008	0.001	***0.000
q11	Number of hours other tasks	0.004	0.002	*0.097
q12	Practice type GP (0=fully or partly fixed pay, 1= other financing)	0.044	0.032	0.181
Patients personal characteristics				
q7	Patients residence	0.023	0.030	0.752
b1	Patients gender	0.067	0.103	0.317
b2	Patients year of birth	0.003	0.002	0.143
b3	Patients education	0.001	0.044	0.963
b4	Patients number of consultations on yearly basis	0.010	0.013	0.184
b5	Patients judgment of their own health	0.090	0.054	0.257
b6	Patients suffering from severe disease (0=no, 1=yes)	0.236	0.133	*0.070

Few of the personal characteristics of the GPs are significant in the regression. Only the age of the GP influences efficiency negatively. Note that the GPs evaluation of his/her work as good, where the scale is reversed from the patient evaluation responses, is correlated with a high efficiency.

Of the organisational characteristics, the two variables that enter the DEA model are highly significant, and have effect in the same direction as in the production model itself, i.e. that increased use of inputs and decreased output reduces the efficiency estimate. This implies that these variables are in some sense heteroskedastic, in that there is an increasing spread away from the frontier with increased inputs and reduced outputs that come in addition to the marginal effect on the frontier. The financing of the GP has, perhaps surprisingly, no significant impact on efficiency, although the point estimate indicates slightly higher efficiency with incentive-based financing rather than fixed pay.

The patient mix does not seem to influence efficiency estimates, except for the variable reflecting whether the patient was suffering from a severe disease. They do not appear however to be more or less satisfied than others with the waiting time, as seen from table 1. Any remaining patient mix effect could be already captured through the correction of the quality evaluation variable.

8. Alternate treatment of quality evaluation

Quality of care is, at least to a large extent, under the control of the GP, and is definitely part of the production technology in the sense that there is a trade-off between quantity and quality or alternatively that the provision of high quality requires resource usage. This excludes the two first reasons discussed in the previous section for not including quality evaluation variables in the first stage DEA model. However, in the third rationale for including variables in a second stage analysis, it was argued that some variables are too stochastic in their nature to use in a deterministic method such as DEA. We have in this article demonstrated that it is possible to account for the stochastic nature of the quality evaluation variables within a DEA framework, but for the sake of comparison we have also investigated a two-stage analysis for model III. In this case the quality evaluation variables are not included in the DEA model, but analysed in a second stage regression.

Table 5 presents the results of an OLS regression attempting to correct or explain the biased efficiency scores of the model III. TOBIT is not required in this case, since there is only one censored unit in the DEA results from this model. The model specification is determined by stepwise exclusion of insignificant variables at the 5% level, which results in a model with five significant explanatory variables. That the number of significant variables is higher than in the DEA specification tests of section 5 is probably because the tests developed found usable in DEA models have much less power than those used in regression analysis.

More surprisingly, the evaluation of waiting time is not among these significant variables, which clearly demonstrates that the effect on average model III efficiency, which is measured without concern for quality, is different from the effect of increasing quality for GPs on the best practice frontier as in model I. It is worth noting that model I implicitly takes into account all the quality variables, but that only s22 was found to be significant. This means that for a given number of hours as input, in model I there is a trade-off on the frontier between the number consultations and s22, while in the regression below there is a trade-off for the average linear relationship between the number of consultations and e.g. s1, but not for s22. As for the case of q10 and q4 in the previous section, this emphasises that frontier effects and average effects are not to be confounded.

Table 5: Stepwise inclusion of quality evaluation variables s1-s25 in regression model on Technical output efficiency E2 from DEA model III (without quality in DEA model). The number of stars *, ** and *** corresponds to the significance levels 10%, 5% and 1%.

code	variable	B	Std. Err	Sig.
const		-0.206	0.375	0.586
S1	...spent sufficient time during the consultation	-0.351	0.109	***0.002
S3	...was easy to tell the GP about the patient's situation.	-0.377	0.150	**0.016
S11	…offered preventive action	0.488	0.130	***0.000
S14	…also handled emotional problems related to the health conditions	0.409	0.129	***0.003
S19	…was able to see you at a time suitable for you	0.111	0.052	**0.037
Model Fit		R	R2	R2adj
Quality evaluation variables corrected for selection bias (shown above)		0.389	0.322	0.149
Quality evaluation variables not corrected for selection bias		0.394	0.328	0.149

It is perhaps not surprising that evaluation of the time spent during the consultation should be negatively related to the efficiency, since more input usage reduces efficiency. Spending time also facilitates the patient's ability to explain her/his situation.

Three of the quality variables in table 5 have in fact a positive impact on efficiency, showing that on average there is a positive correlation between quantity and quality along these dimensions. In the DEA model, as in all production models, there is a trade-off on the frontier between any pair output variables, by assumption. This does not preclude a positive correlation of quality and quantity for the inefficient units, but in DEA this must stem from the behaviour of the GPs rather than from the technical constraints on health service production.

Other explanations are possible, among them errors in variables or specification. If the evaluation of the preventive action taken by the GP reflects that consultations could be shorter, then implicitly the number of consultations is not a homogenous variable. Similarly, if a positive relationship between quantity and quality reflects the inherent ability of the GP, and a good GP is both fast and apt at handling emotional problems, then ability is an omitted variable.

9. Conclusion

On the substantive questions of the shape of the production frontier and the level of efficiency for Norwegian GPs when taking account of the quality of their services, this article does not claim to give strong answers. For this the sample is too small and the availability of quality measures too restrictive. It does however, show that providing patient with quality of care can be costly, and that among the available variables only the evaluation of waiting time is significantly costly on the frontier. Again it must be emphasised that this does not show the relative social value of the different quality aspects.

On the methodological side, this work shows that it is possible to tackle the major problem of errors in variables in a non-parametric setting if there is information on the error structure of a variable. In this case the survey nature of the patient quality evaluation questions was used to correct for selection bias or case mix in the quality variable, and a resampling method was used to evaluate the extent of error in the efficiency estimates resulting from the fact that only a sample of patients were available for each GP.

This suggests several avenues for further research. Firstly, is it possible to develop methods to correct the other variables in the DEA model for case mix? There may e.g. be a tendency that GPs with many old patients can have fewer consultations. Secondly, could one use other

sources of information on the extent of error in variables to estimate the sampling error? In this dataset there is a tendency for responses to some questions to be lumpy, with many GPs having a round number of consultations per week. Possibly a smoothing mechanism could be used to estimate the error in such a distribution, again increasing the validity of results that one can obtain from nonparametric production analysis.

References

Banker, R.D. "Estimating Most Productive Scale Size Using Data Envelopment Analysis." *European Journal of Operational Research*, 1984, *17*, pp. 35-44.

Banker, R.D. "Hypothesis Testing Using Data Envelopment Analysis." *Journal of Productivity Analysis*, 1996, *7*, pp. 139-59.

____. "Maximum Likelihood, Consistency and Data Envelopment Analysis: A Statistical Foundation." *Management Science*, 1993, *39*(10), pp. 1265-73.

Banker, R.D.; Charnes , A. and Cooper, W.W. "Some Models for Estimating Technical and Scale Inefficiencies." *Management Science*, 1984, *30*, pp. 1078-92.

Bhattacharyya, G.K. and Johnson, R.A. *Statistical Concepts and Methods*. New York: John Wiley & Sons, 1977.

Charnes, A; Cooper, W.W. and Rhodes, E. "Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through." *Management Science*, 1981, *6*, pp. 668-97.

Charnes, A.; Cooper, W.W.; Lewin, A.Y. ; Morey, R.C. and Rousseau, J. "Sensitivity and Stability Analysis in Dea." *Annals of Operations Research*, 1985, *2*, pp. 139-56.

Farrell, M.J. "The Measurement of Productive Efficiency." *Journal of the Royal Statistical Society*, 1957, *120*, pp. 449-60.

Guiffrida, A. "Productivity and Efficiency Changes in Primary Care: A Malmquist Index Approach." *Health Care Management Science*, 1998, *2*, pp. 11-26.

Johnson, N.J; Kotz, S. and Balakrishnan, N. *Continuous Univariate Distributions*. New York: John Wiley and Sons, 1994.

Kittelsen, S.A.C. "Monte Carlo Simulations of Dea Efficiency Measures and Hypothesis Tests," *Memorandum.* Department of Economics, University of Oslo, 1999.

Kneip, A.; Park, B.U. and Simar, L. "A Note on the Convergence of Nonparametric Dea Efficiency Measures," U. C. d. L. Institut de Statistique, *Discussion Paper 9603.* 1996.

Kvamme, O.J. and Hjortdahl, P. "Den Gode Praksisen - Norske Pasientar Sine Vurderringar Og Prioriteringer." *Tidskrift for Den norske lægeforeningen*, 1997, *18*(117), pp. 2607-9.

Kvamme, O.J.; Sandvik, L. and Hjortdahl, P. "Pasientopplevd Kvalitet I Allmennpraksis (Patients' Evaluation of Quality in General Practice)." *Tidskrift for Den norske lægeforeningen*, 2000, *21*(120), pp. 2503-6.

Lovell, C.A.K. "Production Frontiers and Productive Efficiency," H. O. Fried, C. A. K. Lovell and S. S. Schmidt, *The Measurement of Productive Efficiency: Techniques and Applications.* Oxford: Oxford University press, 1993,

Petersen, N.C. and Olesen, O.B. "Incorporating Quality into Data Envelopment Analysis: A Stochastic Dominance Approach." *International Production Economics*, 1995, *39*, pp. 117-35.

Simar, L. "Aspects of Statistical Analysis in Dea-Type Frontier Models." *Journal of Productivity Analysis*, 1996, *7*, pp. 177-85.

Simar, L. and Wilson, P.W. "Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models." *Management Science*, 1998, *44*, pp. 49-61.

Tobin, J. "Estimation of Relationships for Limited Dependent Variables." *Econometrica*, 1958, *26*, pp. 24-36.

Appendix: Variables in the data

Table A.1: Variables in Group A, for each of 1435 patients.

Variable	The patient's evaluation of the physicians…	Mean value	ANOVA test of difference between physicians	
			F	Sign.
A1	…use of time during the consultation	4.121	4.493	***0.000
A2	…interest in the personal situation of the patient	4.283	3.759	***0.000
A3	…was easy to tell the about the patients situation	4.231	3.114	***0.000
A4	…way of let the patient participate on the decisions made	4.152	3.033	***0.000
A5	…listened	4.359	3.791	***0.000
A6	…was well informed about the patients condition	4.536	2.115	***0.000
A7	…ability in healing the patients symptoms quickly	4.221	1.591	***0.005
A8	…way of helping the patient feeling sufficiently well to continue everyday life	4.191	1.347	*0.053
A9	…thoroughness	4.220	3.646	***0.000
A10	…performance of the physical checks	4.113	2.888	***0.000
A11	…offer of preventive action	3.849	1.705	***0.002
A12	…explanations of the purpose of tests and treatment	4.155	2.601	***0.000
A13	…explanations to the patient on questions related to tests and treatment	4.163	2.912	***0.000
A14	…handling of the patients emotional problems related to the health conditions	3.919	2.359	***0.000
A15	…help to make the patient understand the importance of compliance	4.114	2.054	***0.000
A16	…knowledge regarding matters the patient had told him in previous occasions	4.007	2.410	***0.000
A17	…way of preparing the patient on what to expect if referred on to specialist or	3.955	2.380	***0.000
A18	…staff and its helpfulness	4.222	5.642	***0.000
A19	…ability to see you at a time suitable for you	4.104	5.053	***0.000
A20	…s office availability on the phone	3.378		***0.000
A21	…availability on the phone	3.320	3.516	***0.000
A22	…s waiting time in his practice	3.435	5.596	***0.000
A23	... ability to perform quick help in emergencies	4.245	2.607	***0.000
A24	…and if he is recommendable to the patients friends	4.518	3.597	***0.000
A25	…and if the patient is considering to change physician into another one	4.595	1.457	**0.020

Table A.2: Variables in Group B, for each of 1435 patients.

Variable	The patient characteristics	Mean value	ANOVA test of difference between physicians	
			F	Sign.
B1D	Gender (0=female, 1=male)	0.302	2.804	***0.000
B2	Year of birth	47.497	4.150	***0.000
B3	Highest completed education	1.576	4.048	***0.000
B4	Number of physicians visits last 12 months	5.574	1.715	***0.001
B5	Evaluation of own health status	2.945	2.048	***0.000
B6D	Presence of serious disease (0=no, 1=yes)	0.376	1.101	0.292

Table A.3: Variables in Group Q, for each of 52 physicians.

Variable	The physician characteristics	Mean value	St.Dev
Q1	Year of birth	52.38	6.06
Q2	Gender (1=male, 2=female)	1.31	0.47
Q3	Number of years in Primary care	15.50	5.96
Q4	Number of hours with direct patient contact	31.65	62.43
Q5	Number of physicians in current practice	3.20	2.04
Q6	Number of other employees in current practice	6.39	8.48
Q6/5	Ratio of other employees pr physician	2.08	0.27
Q 7	The degree of urbanisation of patients (1=city, 2=large town, 3=small town, 4=rural)	2.08	1.00
Q8	The physician a specialist (1=yes, 2=in training, 3=no)	1.12	0.32
Q9	Number of years in current Primary care	11.56	6.91
Q10	Weekly number of consultations	86.94	962.45
Q11	Weekly number of hours other tasks	7.89	44.22
Q12	Financing of practice (1=fixed subsidy independent of number of consultations, 2= fixed salary, 3= fixed subsidy and partly coverage pr consultation, 4 = only partly coverage pr consultation, 5= capitation, 6= other means of financing)	2.51	1.91
Q13	Satisfaction with own work (1=best, …, 5=worst)	1.94	0.86