The Journal of Mental Health Policy and Economics
J. Mental Health Policy Econ. **4**, 79-90 (2001)

Scale, Efficiency and Organization in Norwegian Psychiatric Outpatient Clinics for Children

Vidar Halsteinli,[1] **Sverre A.C. Kittelsen**[2]***** **and Jon Magnussen**[3]

[1]*Master of Sc., Research Scientist, SINTEF Unimed, Health Services Research, Trondheim, Norway*
[2]*Ph.D., Research Economist, Frisch Centre, and HERO - Health Economics Research Programme at the University of Oslo, Norway*
[3]*Ph.D., Research Director, SINTEF Unimed, Health Services Research, Trondheim*
and HERO - Health Economics Research Programme at the University of Oslo, Norway

Abstract

Background: It is generally believed that 5 percent of the population under 18 years is in need of specialist psychiatric care. In 1998, however, services were delivered to only 2.1 percent of the Norwegian population. Access to services can be improved by increasing capacity, but also by increasing the utilization of existing capacity. Changing financial incentives has so far not been considered. Based on a relatively low number of registered consultations per therapist (1.1 per therapist day) the ministry has stipulated that productivity should increase by as much as 50 percent.

Aims of the Study: Measuring productivity in psychiatric care is difficult, but we believe that studies of productivity should be an important input in policy making. The aim of this paper is to provide such an analysis of the productive efficiency of psychiatric outpatient clinics for children and youths, and in particular to focus on three issues: (i) is an increase in productivity of 50 percent a realistic goal, (ii) are there economies of scale in the sector, and (iii) to what extent can differences in productivity be explained by differences in staff-mix and patient-mix?

Methods: We utilize an approach termed Data Envelopment Analysis (DEA) to estimate a best-practice production frontier. The potential for efficiency improvement is measured as the difference between actual and best-practice performance, while allowing for trade-offs between different staff groups and different mixes of service production. The DEA method gives estimates of efficiency and productivity for each clinic without the need for prices, and thus avoids the pitfalls of partial productivity ratios. The Kolmogorov-Smirnov statistic is used to compare efficiency distributions, providing tests of variable specification and scale properties.

Results: Based on 135 observations for the years 1997 to 1999, the tests lead to a model with two inputs, two outputs and variable returns to scale. The outputs are number of hours spent on direct and indirect interventions, while neither the number of interventions nor the number of patients was found to be significant. The inputs are the number of university-educated staff and other staff, but disaggregation of the latter group was not significant. The average of estimated clinic efficiencies is 71%. The mean productivity is 64%, but many large clinics have considerably lower performance due mainly to scale inefficiency.

Discussion: There seems to be considerable room for improved performance in these clinics. It is interesting that the potential is not that far from the officially stipulated goal of 50% increased productivity. Staff composition does matter for clinic performance, but the different groups do not have significantly different marginal productivities, indicating a lack of ability to utilize specialized skills. It should be noted that these results to some extent depend on the assumptions that medical practice is efficient, and that the available data accurately captures the activities of the clinics.

Implications for Future Research and Health Policy: More appropriate outcome measures, e.g. global assessment of functioning scores (GAF), will soon be available and will improve the policy value of this type of analysis, as will a more refined data set with information about the number of personnel in training positions. The analyses in this paper indicate that a lack of consensus on the issues of who should be treated, how they should be treated and by whom results in large variations in productive efficiency. These issues are being debated in Norway, and it should be interesting to see whether this in itself leads to higher efficiency or whether a change in the incentive structure will be needed.

Received 25 June 2001; accepted 28 September 2001

Introduction

It is generally believed that 5 percent of the population under 18 years is in need of specialist psychiatric care.[1,2] Psychiatric care for children and youths (BUP*) is a relative new service in Norway, developed gradually since the 1960s. In 1998, however, services were delivered to only 2.1 percent of the Norwegian population.[3] There is also a substantial variation in capacity between different geographical regions. Consequently, an overall increase in capacity and a more even geographical distribution of services have both been political goals.[4]

Access to services can be improved by increasing capacity, but also by increasing the utilization of existing capacity. There are current plans to increase capacity both by opening more

***Correspondence to**: Sverre A.C. Kittelsen, Frisch Centre, Gaustadalléen 21, N-0349 Oslo, Norway
Phone +47-22-958 815
Fax +47-22-958 825
Email: s.a.c.kittelsen@frisch.uio.no

Source of Funding: Contract grant sponsors: The Norwegian Research Council, through HERO and The Norwegian Ministry of Health and Social Affairs, through SINTEF

* "BUP" is the Norwegian abbreviation for Children and Youth Psychiatry. We have chosen to keep this rather than use an English abbreviation.

outpatient clinics and by increasing the number of therapists. Based on a relatively low number of registered consultations per therapist (1.1 per therapist day) it is however stipulated that within the existing capacity productivity should increase by as much as 50 percent.[5]

Measuring productivity in psychiatric care is difficult, because there are inherent difficulties in measuring the outcome of service production and because there are few agreements as to what constitutes an efficient production process. We believe, nevertheless, that studies of productivity should serve as an important input in policy making. Thus, the aim of this paper is to provide such an analysis of the productive efficiency in psychiatric outpatient clinics for children and youths and in particular to focus on three issues:

(i) Is an increase in productivity of 50 percent a realistic goal?
(ii) Are there economies of scale in the provision of outpatient services?
(iii) To what extent can differences in productivity be explained by differences in staff-mix and patient-mix?

To answer these questions we utilize a methodological approach termed Data Envelopment Analysis to construct a best-practice production frontier for the years 1997 to 1999. The potential for efficiency improvement is measured as the difference between actual and best-practice performance.

The paper is organized as follows: in the next section we set the background for the analysis by providing a more thorough description of the production process in BUP outpatient clinics. The subsequent sections discuss measurement of inputs and outputs and the methodology used. The last two sections present data and results and provide a discussion.

*Organization of BUP - Outpatient Clinics**

Loosely formulated psychiatric services for children and youths are aimed at the treatment of emotional and mental disorders and at correcting an undesired behavioral pattern through the combined use of therapy and interaction with the patient's environment (relatives, school, etc). As much as 95% of all psychiatric care for children and youths in Norway are delivered in an outpatient setting, but it is not altogether clear what specific purpose the BUP-clinics will serve.[6] The patient's condition may not be easy to diagnose, and unlike somatic illnesses it is not at all obvious how one should proceed with treatment. Thus each outpatient clinic will to a large degree have discretion regarding the type of personnel needed to provide treatment, the type of services that are to be delivered to the patients and the duration of the treatment. In addition, the seriousness of the problem cannot always be assessed, making priority decisions difficult and also creating differences in patient mix from one clinic to the next. Thus we tend to observe differences among clinics in *priority decisions, staffing decisions and treatment patterns.*

Priority Decisions

Outpatient clinics in Norway are part of the secondary, specialized health care system. In this system clinics are responsible for serving the population of specific catchment areas. While epidemiological studies indicate that 5 percent of the population aged 18 and below will need specialized psychiatric services,[1,2] only 2.1 percent currently receive such care. Consequently there are waiting lists and a need to choose among different types of patients. Still we observe that few clinics explicitly recognize that they play any role when prioritizing among patient groups, or even feel that they should play such a role. Rather patients are often treated on a first come-first serve basis and waiting lists are regarded solely as a result of scarce resources, and not a result of the decisions made by the clinic. Additionally, there are no centrally stated rules of thumb for priority setting, and both local and central government implicitly expect the clinics to "do the right thing".

There is, within the clinics, no consensus as to when one should admit a patient into treatment. Furthermore, there is a marked difference among clinics on how the decision to admit is done and by whom. In one variant there is the equivalent of the admitting physician who reviews the applications and makes the decisions as to who will and will not be treated. In another variant the decision to admit is done after meetings involving several members (or even all) of the staff and thus is a much more time consuming procedure. Clearly, the type of admitting process will have implications for the productive efficiency of the outpatient clinics.

Staffing Decisions

Outpatient clinics are generally staffed with two types of personnel: university educated (mainly psychiatrists and psychologists) and college-educated (mainly in the fields of social work and education). How patients and tasks should be divided among these professions is an unresolved issue in the outpatient clinics. The conflict is partly about how patients should be treated (and is thus related to the discussion of treatment guidelines; see below), but it is also a struggle for authority within the clinics. This situation is not particular to Norway. Hagen & Hatling.[8] note that similar conflicts exist in all Nordic countries. Again, it is worth noting that this is a situation that is allowed to persist in part because local and central authorities choose not to interfere.

One particular effect of lack of treatment guidelines is that the allocation of patients among different professions tends to become more ad hoc. Thus in many cases the allocation of patients to therapists is based on the workload of the therapists rather than a principle of matching the patient's problem to the therapist's qualifications. This is so not only because there are many cases where it is unclear exactly what type of qualifications are needed, but also because a less clear division of tasks among professions will benefit those who (by no specific definition) are least qualified.

Treatment Guidelines

Services can be provided in many ways, and there are few established treatment standards or evidence-based guidelines as how to treat patients.[9] Thus the struggle among professions

* The discussion in this section builds on a more thorough discussion found in Hatling and Magnussen[7].

is carried over to the treatment process. This is most apparent on three levels: the decision to admit a patient into treatment, the choice between using individual therapy or family therapy, and the choice between using a single therapist or a team of therapists with different backgrounds.

The end result of this situation is a sector that offers a multitude of solutions, some founded in local beliefs and cultures and some the result of a professional impasse. To put it strongly, the professional and cultural environment may be a better predictor of treatment type than the diagnosis itself. In some ways this is to be expected, since it is difficult to assign an accurate diagnosis and there is no blueprint treatment for the majority of patients. On the other hand, this uncertainty makes it easier to adopt practice patterns that lead to lower levels of productivity. In any case it is beyond the scope of this paper to assess the usefulness of the different approaches that can be observed in the BUP-clinics.

Financial Incentives

Although it is commonly acknowledged that there probably is potential for improved efficiency, there has so far been little focus on the use of financial incentives. Global budgets from county councils account for 80 percent of the outpatient clinics' income. The additional 20 percent is financed by the National Insurance Scheme, and are related partly to the number of treated patients, partly to the number of opening hours available for patient-related activities and partly to the size of the treatment staff. In practice, then, only a minor fraction of the outpatient clinics' total income will be related to the actual treatment of patients. Obviously the financial incentives to be efficient are, at best, weak.

That said, it is not at all clear how one should construct a financing system that provides the appropriate incentives for efficiency. Health care financing systems are characterized by an inherent trade-off between efficiency and selection (see Newhouse.[10] for an overview of this literature). This trade-off is likely to be magnified in the financing of mental health services, due to a high degree of product heterogeneity. Thus, although a move to a high powered per-case financing system is likely to lead to an increase in the number of treated patients, as a side effect it may produce a bias towards simpler cases when it comes to patient selection.

It is not the purpose of this paper to discuss the relative merits of different financing systems. Thus at this point we merely acknowledge that there are few financial incentives for the clinics to perform efficiently. The main question here is how large the potential for efficiency improvement is. By using the concept of a best practice technology, however, we are able to assess the overall performance of the sector and thus to draw implications about the effect of these variations on production performance and thereby about the efficiency of resource allocation. To do this we have to provide a measure of productive efficiency that captures the essence of the activities and is recognizable to those working in the sector.

Measuring Inputs and Outputs

The treatment process will consist of a series of interventions related to each patient. The interventions will be of different forms depending on the type of disorder, the social setting, and - as we have argued - the outpatient clinic itself. Interventions may be aimed directly at the patient or also at the patient's surroundings (schools, relatives, primary health care, etc). They can take place in situations when the patient and therapist are alone, or in various forms of group setting.

Ideally one would like to model the input-output relationship using data on number of interventions by type and number of personnel full-time equivalents (FTEs) by category. While FTEs are available on a fairly detailed level, the number of interventions is not. In the BUP clinics the following figures are available:

Number of Cases/Patients (P). This measure approximates the number of clients in the system, but is limited to clients who are currently involved in a treatment program.

Number of direct patient-related interventions (I-dir). This measure will be closely related to number of visits by the patient, but may also include visits in the patient's home, in schools, etc.

Number of indirect patient-related interventions (I-ind). This measure will capture all activity related to the clients that is not direct treatment, e.g. consultations with schools and other community institutions.

Number of hours spent on direct patient related interventions (H-dir). Interventions may be of different length and may involve one or more therapists. Unfortunately we are not able to combine number of hours with number of therapists. This may have implications for our measures of efficiency. It should also be noted that when we include a measure of number of hours spent on interventions as an output in the analysis we assume that this is "time well spent".

Number of hours spent on indirect patient related interventions (H-ind). Depending of the type of problem, each patient will receive a number of interventions, each intervention implying a certain number of therapist hours. We also make a distinction between direct and indirect interventions. Including all five outputs allows us to compare efficiency in clinics where a small number of patients receive a large number of interventions with clinics where a large number of patients receive a small number of interventions. We can also compare efficiency among clinics with a relatively large or small share of indirect interventions. Note, however, that we do assume that there are no inefficiencies in the chosen treatments. Thus every hour of every intervention is assumed to be "necessary" and equally valuable to the patient.

Measures of input usage are available for three different types of personnel:

- University* educated staff (S1).
- College educated staff (S2)
- Administrative staff (S3)

College educated staff includes nurses, social workers and those with a college degree in education, while university educated staff includes psychologists, psychiatrists and physicians.

* The difference between university and college is similar to the notion used in the US.

Data were collected from the total population of all 49 Norwegian BUPs over a three-year period (1996-98). After removing outliers and missing observations we are left with 135 observations in the sample. **Table 1** summarizes the data on inputs and outputs and their aggregates. The size of BUPs varies widely, with staff size ranging from 3 to 82.9, and the number of patients from 23 to 715. Staff composition is very dispersed with university-educated proportions from zero to more than two thirds.

The simple measure of productivity discussed in the introduction, consultations per therapist day, likewise varies from 0.36 to 2.13 with a mean of 1.09. A major aim of this analysis is to see whether such differences in productivity carry over to a richer model of production in the BUP clinics, and whether differences in productivity among staff groups can explain some of the productivity dispersion.

Methods

DEA Efficiency Estimates

The idea of measuring technical efficiency by a radial measure representing the proportional input reduction possible for an observed unit while staying in the production possibility set stems from Debreu[11] and Farrell,[12] and has been extended in a series of papers by Färe, Lovell and others.[13,14] Farrell's specification of the production possibility set as a piecewise linear frontier has also been followed up using linear programming (LP) methods by Charnes, Cooper et al (e.g. Charnes, Cooper & Rhodes[15] who originated the name DEA. For an overview of the literature on DEA see e.g. Seiford).[16] The decomposition of Farrell's original measure relative to a constant returns to scale (CRS) technology into separate measures of scale efficiency and technical efficiency relative to a variable returns to scale (VRS) technology is due to Førsund & Hjalmarsson[17] and has been implemented for a piecewise linear technology by Banker, Charnes & Cooper.[18] Their Data Envelopment Analysis (DEA) formulation has served as the main model of most recent efficiency studies and is the basic model in this paper.

The DEA method estimates the frontier of the technical feasible production set as the piecewise linear envelopment of the best practice observed units. In parallel with the non-parametric DEA approach, an alternative parametric tradition has developed in which the frontier is given a specific functional form. While the original contribution of Aigner and Chu[19] was a deterministic frontier, which like DEA assumes the absence of measurement error, later development in stochastic frontier analysis (SFA) has been able to estimate a decomposition of the residuals into inefficiency and noise.[20] While previously SFA alone had the advantage of being able to test hypotheses, this has been changed in the establishment of a statistical basis for DEA by Banker[21] and Korostelev, Simar & Tsybakov,[22,23] as explained in the next subsection. Similarly, the advantage that DEA is able to model multiple outputs and multiple inputs at the same time has been challenged in recent work by Coelli & Perelman.[24] In this application we have chosen to use DEA primarily because it does not require the assumption of a specific functional form and is therefore a better fit with the data than SFA would have been.

Various measures of productive efficiency are possible, such as social efficiency and allocative and cost efficiency, which we are not able to estimate due to a lack of data on prices and/ or social evaluation of production. Instead we concentrate on technical measures of efficiency, in the sense that we compare actual behavior with some point on the frontier of the technically feasible set. This frontier point will in general not be the optimal behavior if values are applied, but if the model is correctly specified, the optimal behavior will be one of the points on the frontier.

Technical efficiency can be measured both in an input direction, as the proportion of inputs that are necessary to produce a given level of output, and in an output direction, as the ratio of actual production to the maximum production given the level of inputs. In the psychiatric outpatient clinics we have chosen to concentrate on the latter, implying a focus on how much more psychiatric treatment could be provided with existing levels of staffing, if clinics were technically efficient.

This paper reports the means and variation of three measures of efficiency, as well as a scale indicator and the shadow prices associated with each of the variables. Using the terminology of Førsund & Hjalmarsson,[17] the Farrell[12] radial estimate of technical output efficiency is reported as E_{2i}, which is the ratio of the actual production of the clinic i to the potential production if this clinic were producing the maximum feasible quantities given its level of input usage. Technical productivity E_{3i} is the ratio of actual production to the maximum feasible production had the clinic been operating at the optimal scale. Scale efficiency E_{5i} is the ratio of technical productivity to technical efficiency (E_{3i}/E_{2i}), and thus represents the productivity a clinic *i* would have had, if it had been technically efficient. A scale inefficient clinic could have become more productive if it had operated at the optimal scale, and the scale indicator λ_i is a measure of how large (>1) or small (<1) it is compared with the optimal size (=1). The shadow prices ω_{ij} are the marginal properties of the frontier as estimated in the DEA method. Only the relative values of two shadow prices are of interest here, as this represents rate of substitution between the two variables, i.e. how much more of an output could be produced had one produced less of another output, or used more of an input. The mathematical details of the DEA method and the various measures are given in **Appendix A**.

Data Analytic Procedures

Statistical tests have been few in the DEA literature. Valdmanis,[25] among others, has used the Mann-Whitney rank-order test to compare the efficiency of public vs. not-for-profit hospitals and found the public hospitals significantly more technically efficient in seven out of ten different input-output specifications. While her approach is fruitful in assessing the performance of separate groups and demonstrates the robustness of results across specifications, her method does not give an answer to the question of which specification is best.

Table 1. Summary statistics for the sample of 135 BUP clinics

			Mean	Median	Standard Deviation	Minimum	Maximum
Output	P	Cases/patients with interventions	209	185	118	23	715
	I-dir	Number of direct interventions	1566	1388	1233	82	7899
	I-ind	Number of indirect interventions	738	524	673	39	3964
	H-dir	Number of hours direct interventions	1744	1441	1509	120	9956
	H-ind	Number of hours indirect interventions	587	438	616	44	4399
	I = I-dir + I-ind	Sum number of interventions	2304	1913	1793	239	11863
	H = H-dir + H-ind	Sum number of hours	2331	1795	2072	271	14355
Input	S1	University educated staff	4.83	4.00	4.06	0.00	25.70
	S2	College educated staff	4.94	3.70	5.68	0.80	39.35
	S3	Administrative staff	2.24	2.00	2.49	0.00	18.10
	S12 = S1 + S2	University or college educated staff	9.77	7.70	9.46	2.00	64.80
	S23 = S2 + S3	College educated or administrative staff	7.18	5.37	8.10	2.00	57.20
	S = S1 + S2 + S3	Sum staff	12.01	9.20	11.87	3.00	82.90
	I / (S12*230)	Interventions per therapist day	1.09	1.06	0.36	0.36	2.13
	S1/S	University staff as share of sum staff	0.41	0.42	0.11	0.00	0.69

Data from 135 observations, 43 from 1996, 45 from 1997 and 47 from 1998

Farrell[12] recognized that statistical tests should be based on the frequency distribution of efficiencies. The problem is that when one assumes that all observations are feasible, i.e. no measurement error, any sampling error would bias the DEA efficiency estimators upward, since the true frontier generally lies outside the estimated frontier. However, recognizing that sampling error exists in DEA analysis also gives a basis for statistical analysis of "deterministic" frontiers.

While tests such as the Mann-Whitney rank-order tests have been used for subset comparisons,[25,26] the assumptions underlying most tests are not fulfilled when testing model specifications since such models generally will be nested. A model 0 will be nested within another model 1 if model 0 can be obtained from model 1 as a special case. This implies that a CRS model is nested within a VRS model, an aggregated model is nested within a disaggregated model, and a model without a specific variable is nested within a model that includes this variable. In nested models, the DEA estimates of efficiency will be ranked so that $\hat{E}^1 \geq \hat{E}^0$ for every observed unit, implying that the bias of the estimators will be at least as large for model 1 as for model 2, and usually larger. Any simple test based on the difference or ratio of such estimators will therefore also be distorted.

In recent developments, Banker[21] has proven the consistency of the DEA estimators under specific assumptions and suggested statistical tests of model specification, while Korostelev, Simar & Tsybakov [22,23] have been concerned with the rate of convergence of non-parametric frontier estimators. Kneip, Park & Simar[27] extend these results to a more general model. Simar & Wilson[28] suggest a bootstrap method for estimating the bias and confidence intervals of efficiency estimates, and Simar & Wilson[29] extend this to suggest a test of returns to scale.* Even though this approach seems feasible, it would be advantageous if simpler techniques were available.

So far, no tests have been suggested that can be shown analytically to be able to discriminate among competing models, especially in small samples. While suggesting among others the Kolmogorov-Smirnov test used below, Banker[21] warns that "... the results should be interpreted very cautiously, at least until systematic evidence is obtained from Monte Carlo experimentation with finite samples of varying sizes." Banker[30] has summarized a series of Monte Carlo runs, using 10-30 repetitions in each evaluation, while Kittelsen[31] has extended this to 1000 repetitions. The results indicate that some tests give crude but usable approximations of the true significance level and power functions, except in very small samples. Of the tests evaluated, the Kolmogorov-Smirnov test is the most conservative, while the ordinary T-test of the difference of means has more power, but tends to more easily overreject a true null hypothesis in small samples and high dimensionality. Banker[21] has also suggested two F-tests that yield similar results to the Kolmogorov-Smirnov test, but unlike the latter, these F-tests are based on specific assumptions on the distribution of inefficiency, and are not reported here. Details of the Kolmogorov-Smirnov and ordinary T-tests reported are given in **Appendix B**.

Results

The procedure chosen in this paper is to start out with a simple model and then proceed to test whether a more disaggregated approach will give a more accurate representation of the production technology. Thus we first specify a model with constant returns to scale and with only one output and one input. Next we include one variable at a time, and test whether the variable has a significant impact on the estimated efficiencies. The null hypothesis is in each case the conservative

* See Grosskopf[32] for a survey of statistical inference in nonparametric models.

Table 2. Hypothesis tree and test results for various DEA models

H0	HAlt	Change in E	KS-test	P-value	T-test	P-value	Result
(H,S,CRS)	Include interventions I	0.024	0.096	0.286	1.137	0.128	Accept H0
(H,S,CRS)	Include cases/patients P	0.032	0.111	0.189	1.569	0.059	Accept H0
(H,S,CRS)	Split hours in H-dir and H-ind	0.038	0.141	0.069	1.827*	0.034	Reject H0
(H-dir,H-ind,S,CRS)	Split personnel in S12 and S3	0.030	0.111	0.189	1.381	0.084	Accept H0
(H-dir,H-ind,S,CRS)	Split personnel in S1 and S23	0.041	0.170*	0.020	1.843*	0.033	Reject H0
(H-dir,H-ind,S1,S23,CRS)	Split S23 in S2 and S3	0.030	0.096	0.286	1.259	0.105	Accept H0
(H-dir,H-ind,S1,S23,CRS)	Variable return to scale	0.064	0.222**	0.001	2.657**	0.004	Reject H0
(H-dir,H-ind,S1,S23,VRS)	Accepted model						

Note: One * denotes a p-value less than 5% and two ** less than 1%. With 135 observations and 268 degrees of freedom, the T-test has critical values of 1.651 (5% level) and 2.340 (1% level), while the Kolmogorov-Smirnov test has critical values of 0.149 (5% level) and 0.185 (1% level).

choice that the variable has no significant impact. If the test statistic is less than the critical value, the null hypothesis is accepted, and the variable in question is excluded from the model. A similar procedure is used for testing for aggregation, where allowing aggregation is the null hypothesis, and for testing returns to scale, where constant returns to scale (CRS) is the null hypothesis. Since the sample size of 135 observations is larger than the threshold of about 100, below which the T-test tends to overreject, we use this as the decisive statistic, but report also the more conservative Kolmogorov-Smirnov statistic D+. On the other hand we do not want to accept the null too easily, so we will use a 5% rejection level.

In specifying our simple model we begin by noting that total number of hours (direct and indirect) serves as a measure of case-mix adjusted activity in the clinics. Thus:

Hours = Patients * (Interventions/Patient)*(Hours/Intervention)

The number of hours equals number of patients weighted with treatment intensity along the dimensions of interventions per patient and hours per intervention. As the only input we use total number of FTEs, assuming that there are no differences in marginal productivity between the different types of personnel. With reference to the previous discussion, this seems to be a reasonable starting point. The inclusion of variables in the disaggregated model will depend on the test results. Outputs are added with number of interventions first followed by number of patients (cases). If interventions are accepted we add the number of cases before we split hours into direct and indirect care. When extra outputs are accepted (or rejected) outputs are split in the order of hours followed by interventions. Inputs are disaggregated only when the full output model has been chosen. Then administrative personnel are defined as a separate input followed by university-educated personnel and finally all three types of personnel. Finally we test for VRS on the chosen input and output model. The results of the tests are summarized in **Table 2**.

Proceeding from the simple output/input ratio with constant returns to scale, we end up with a preferred model consisting of two outputs and two inputs and with variable returns to scale. This path warrants some comments.

First, we note that adding neither the number of interventions nor the number of cases to the number of hours provides extra information. Given that the number of hours per FTE is in the same range, number of interventions or number of cases does not seem to influence the operating environment.

Second, we note that splitting number of hours into time spent on direct care and indirect care does make a difference. Thus there are different operating environments between clinics using a high share of their total time on indirect care versus direct care. One explanation for these differences is that they are due to variations in patient population.

Third, we note that defining administrative labor as a separate input does not influence the efficiency distribution. University personnel, however, need to be separated from other personnel. This implies that there is a statistically significant difference between the marginal productivities of the university-educated staff and the rest, while there is no significant difference between the marginal productivities of

Table 3. Main efficiency, productivity and scale results

		Mean	Median	Standard Deviation	Minimum	Maximum	Weighted Mean
E2	Technical efficiency	0.709	0.734	0.205	0.197	1.000	0.734
E3	Productivity	0.645	0.640	0.189	0.197	1.000	0.623
E5	Scale efficiency	0.919	0.963	0.108	0.501	1.000	0.869
λ	Scale indicator	2.054	1.457	2.813	0.408	22.525	3.674

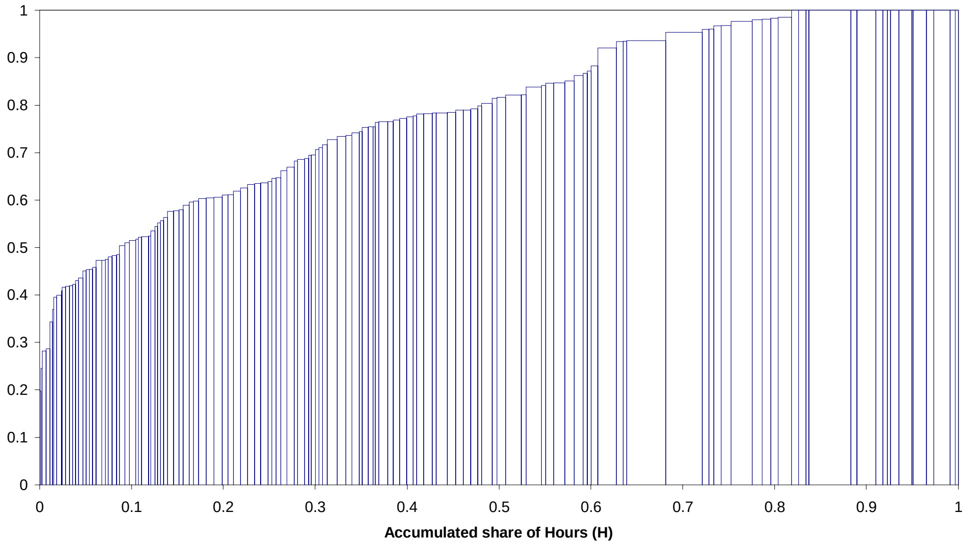


Figure 1. Hecksher-Salter diagram of technical output efficiency E_2

college-educated and administrative staff.

Finally, we note that a hypothesis of variable returns to scale is accepted, implying different productivities of efficient BUPs depending on their size.

The main efficiency results and other properties of the estimated technology are given in **Table 3**. The average of estimated clinic efficiencies is 71%, but the variability is still large. In addition to the mean and spread of clinic efficiencies, the "weighted means" are the measures weighted by the total number of hours, both direct and indirect. The weighted mean technical efficiency is slightly larger than the unweighted mean, a sign that larger BUPs are somewhat more efficient than smaller BUPs. This can be seen more clearly in **Figure 1**, which shows the efficiencies of the clinics in ascending order, and where the widths of the bars are proportionate to the number of hours produced by each clinic. There is a clear tendency for the larger BUPs to be at the efficient end of the chart, but with many smaller BUPs interspersed. From the diagram one can also see that the wholly efficient BUPs, which define the frontier or reference for the inefficient clinics, represent about 18% of the total production in the sample.

Considerably fewer clinics define the maximum productivity in the sector, representing only about 8% of total production, as can be seen from the Hecksher-Salter diagram in **Figure 2**. The larger units are well dispersed in the diagram, and the very largest BUPs have quite low productivity. The mean productivity is 65%, but many large clinics have considerably lower performance. The tail of worst performers in both diagrams consists, however, of very small BUPs, and some of these results may be due to circumstances not captured in the model.

The reason why large BUPs can have high efficiency and

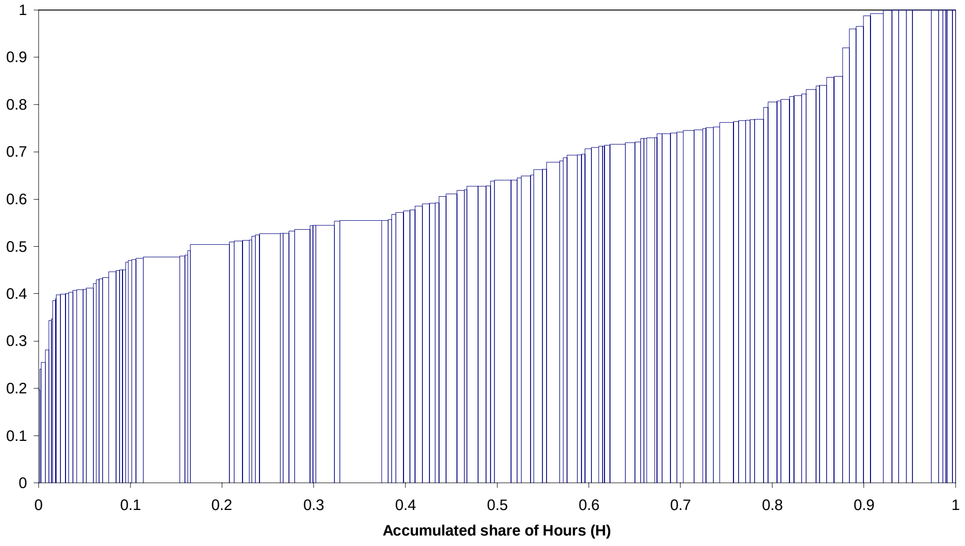


Figure 2. Hecksher-Salter diagram of technical productivity E_3

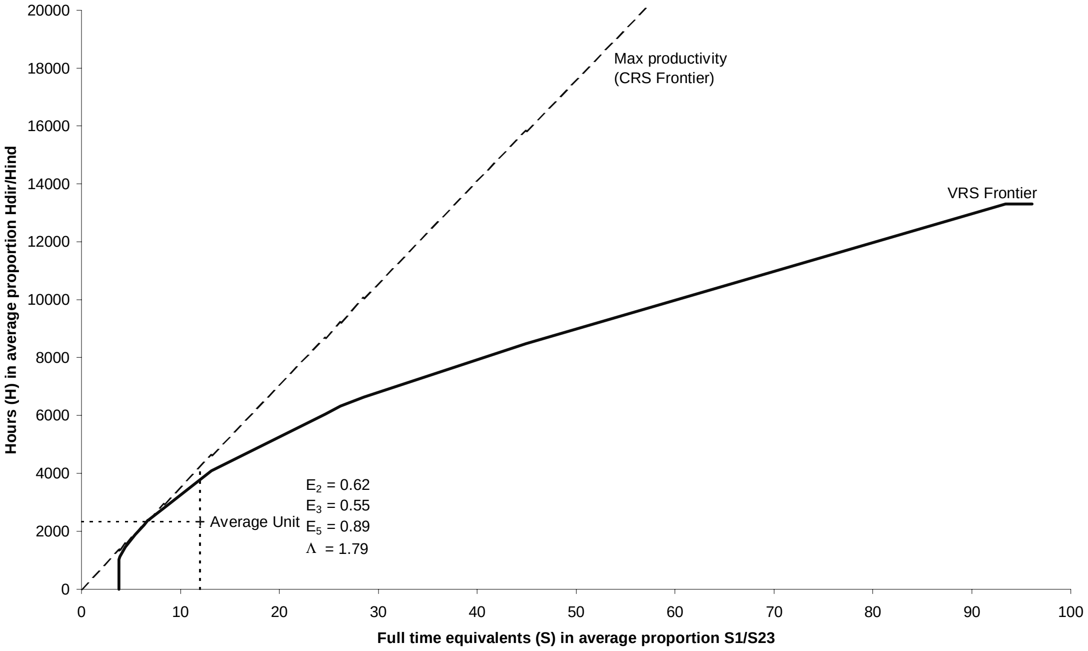


Figure 3. Hackman-Passy-Platzman diagram of estimated frontier in plane defined by the average unit

low productivity can best be seen in **Figure 3**. This diagram represents the intersection of the four-dimensional estimated production frontier and a two-dimensional plane defined by the average input and output proportions in the sample, and is calculated using an algorithm from Hackman, Passy & Platzman.[33] The average unit is defined by the total number of hours produced and the total number of FTEs used divided by the number of observations, and is a point on this plane. One sees that the maximal productivity is achieved at a point near the average output size, but there is a region of sizes from about four to fourteen FTEs where the estimated VRS frontier is quite close to the maximal productivity "CRS front". This range is at or near optimal size, but BUPs that are larger than about 20 FTEs are clearly larger than optimal. Large BUPs can therefore be technically efficient since they are on the efficient frontier, and doing the best they can given their size, but still be less productive than the smaller BUPs.

The scale efficiency E_5 reported in **Table 3** is the ratio of productivity E_3 to efficiency E_2, and on average it is about 92%. This measures the lack of productivity due to inoptimal scale, and can be interpreted as the productivity of a clinic had it been technically efficient. The decreasing returns to scale that the figure shows is strongly significant by the tests in **Table 2**. While the optimal scale in general varies with the mix of inputs and outputs, similar diagrams for different mixes (not shown here) give much the same range of near-optimal sizes.

The marginal product of each labor input on the frontier mapping of each clinic is reported in **Table 4**, revealing an estimate of how many more hours an efficient clinic could spend on direct patient interventions if it increased its staffing in that category by one position. Interestingly, this is on average greater for college-educated (339) than for university-educated personnel (270). Because of the piecewise linear structure of the DEA estimate of the frontier, the variability of these estimates is large, and they are not significantly different from each other. One should exercise care in interpreting marginal products for individual clinics, but average results are still of interest. On the output side, one hour spent on indirect patient interventions is 15% more costly in terms of resource usage than one direct hour, but again this is not a statistically significant difference. Multiplying the shadow prices by quantities, one can get an estimate of implied value shares. Point estimates are that about one third of production is attributable to university educated staff, and that two thirds of the resources are used on direct patient intervention time. The final lines of the table show that the four inputs and outputs are highly significant as variables in the model.

Discussion

The main results emerging from this analysis are as follows.

(i) Average efficiency is around 70%, and productivity around 65% in the BUP outpatient clinics. Based on these results there seems to be considerable room for improved activity in these clinics.

It is also interesting, although probably coincidental, that the potential for higher output is not that far from the officially stipulated goal of 50% increased productivity.[5] It should also be remembered that these measures are derived under the assumption that medical practice is efficient. If this is not the case the observed best practice and the theoretical frontier will not coincide, and there is room for further improvement in outputs.

There are, however, some qualifying remarks to be made. First of all, clinics may vary as to how much time should be spent treating outpatients. In some cases personnel are dedicated to other tasks either in the community or for

Table 4. Marginal products, marginal resource cost, implied value shares, and significance of individual inputs and outputs

		S1	S23	H-dir	H-ind
Shadow prices	Average	270	339	1.00	1.15
	Standard deviation	252	222	-	1.37
	Average variable level	4.85	7.21	1753	589
Implied value shares	Average	0.37	0.63	0.67	0.33
	Standard deviation	0.27	0.27	0.35	0.35
T-value (5% critical value 1.650)		3.323**	8.432**	6.659**	3.190**
P-value		0.001	0.000	0.000	0.001

Note: Shadow prices are normalized in units of H-dir. T-values and associated P-values are based on comparison of efficiency estimates in models with and without each variable.

inpatients at adjacent hospitals. Also there will be variations the extent to which personnel at clinics spend their time servicing primary health care. We do not capture "consultative work" as an output in our model, nor can we correct the input measures for time spent in other facilities. The implications of this are that the clinics on the frontier may not be the "real" reference units, and that the potential for output improvement of inefficient clinics could be different* from what emerges from this analysis.

Next, we note that the output measures used for the BUP sector may not capture all the aspects of case-mix differences. As noted, there may be a substantial difference in number of therapists present for patients that is not captured in our measure of number of hours spent on direct contact with patients. If this is also reflected in the outcome of the treatment, clinics that rely on using more than one therapist will get too low efficiency estimates.

Third, outpatient services are delivered by specialized personnel, e.g. physicians specialized as psychiatrists or psychologists specialized as clinical psychologists. In most cases, however, outpatient clinics are staffed with personnel undergoing training to become specialists. This implies that a substantial amount of time is spent on training, both by those undergoing it and by trained personnel acting as mentors. It is reasonable to assume that efficiency will be affected by the number of therapists engaged in some form of training. At present we have not included variables to adjust for this in our analysis, thus possibly overestimating the potential for efficiency improvement.

(ii) There are variable returns to scale in the BUP sector, specifically such that the highest productivity is achieved by small clinics and large clinics have low scale efficiency.

Initially we would expect that activity be proportional with staff. There might, however, be variations in other types of activity, in the sense that large clinics could have a higher share of consultative work related to primary care and hospitals, and thus have a lower level of productivity. In this case our estimates of low scale efficiency for the largest clinics is caused by the lack of a full set of variables, and not by real productivity differences. On the other hand there might be real reasons to expect decreasing returns in BUPs. Small organizations often have advantages in less formal reporting procedures and ways to circulate information, and in less bureaucratic systems of control. Inactivity, or less than optimal use of time, is less hidden in small units. To the extent that patient cases are discussed in full staff meetings, less time is wasted if fewer persons need to be present.

(iii) Staff composition matters, although marginal products are quite similar.

To understand how staff composition could be expected to affect efficiency, we need to look more closely at internal organization of the outpatient clinics. For the moment side-stepping the fact that many will be in training positions, there are broadly four types of therapeutic personnel in the clinics: psychiatrists, psychologists, nurses and social workers. In theory there is a division of labor between these professions. Social workers will, at the outset, have limited possibilities to perform individual therapy, psychiatrists are needed to administer medication but will be less qualified to organize the patients' living arrangements, etc. In this respect the staffing mix would be a reflection of the clinic's patient mix. What we observe in practice, however, is a production process where there is very little division of labor, and where specialized skills are utilized to transfer knowledge to other professions rather than to use it in a clinical setting.[7]

In many ways this is a way of organizing the activity that is inherently unproductive. Much time is spent on general staff meetings, with respect to sorting out patients that are admitted and discussing the treatment of individual patients. These meetings are a way of organizing the treatment process that compensates for lack of knowledge on the part of the therapist responsible for the patient, and work as a sort of internal education. On the other hand, it is probably true that people in need of psychiatric care generally are better off when they can relate to fewer persons. Thus a model where the patient would meet four or five therapists during a treatment process could

* Note that the error could go both ways.

be even less productive than the model that is dominant today.

It is also worth noting that the unwillingness to utilize specialized skills by way of a more open division of labor is founded in a fundamental uncertainty about how to diagnose and how to provide medical treatment for mental illnesses. In situations where there is uncertainty, each profession can “rightfully” maintain that it should be responsible for certain tasks. In the case of mental health services the professional disputes about who are/are not qualified to perform certain tasks have not been resolved, and the lack of specialization is as much a result of this impasse as it is the result of a well conceived treatment concept.

Conclusions

Measures of mental health illnesses are hard to find, and in this respect the analysis performed here should be treated with caution. One obvious limitation to this study is the lack of appropriate outcome measures. Such measures, in the form of global assessment of functioning scores (GAF), will soon be available and will improve the policy value of this type of analysis. A more refined data set with information about the number of personnel in training positions will also be available, and used to refine the analysis.

Still, the results in this paper seem to support the hypothesis that a lack of consensus on the issues of who should be treated, how they should be treated and by whom results in a sector where there are large variations in productive efficiency. These issues are at present a “topic” in the health policy debate in Norway. At the time of writing, the question of a revised financing system for psychiatric outpatient clinics has also been raised by central authorities. As noted previously, we are not likely to find an optimal financing system. Still, the potential for efficiency improvement that follows from the analysis performed in this paper clearly implies that a strengthening of financial incentives may be a step in the right direction.

Appendix A. Estimates of efficiency in Data Envelopment Analysis

Using the terminology of Førsund & Hjalmarsson,[17] the Farrell[12] radial estimate of technical output efficiency is defined by

$$\hat{E}_2 = \mathrm{Min}_\theta \left\{ \theta \left| (\mathbf{y}/\theta, \mathbf{x}) \in \hat{P} \right. \right\}, \qquad (A.1)$$

where $\mathbf{y}$ is a vector of K outputs and $\mathbf{x}$ is a vector of L inputs, and $\hat{P}$ is an estimate of the production possibility set or technology

$$P = \left\{ (\mathbf{y}, \mathbf{x}) \in \Re_{+}^{K+L} \left| \mathbf{y} \text{ can be produced from } \mathbf{x} \right. \right\}. \qquad (A2)$$

Figure A.1 illustrates the basic concepts. Point A is an observed input/output combination in a one-input one-output technology, and the technology set is the area below and to the right of the curved frontier. Given a constant level of input OE, the technical output efficiency of unit A is the ratio of actual output EA (=OC) to the maximum production that is feasible ED.

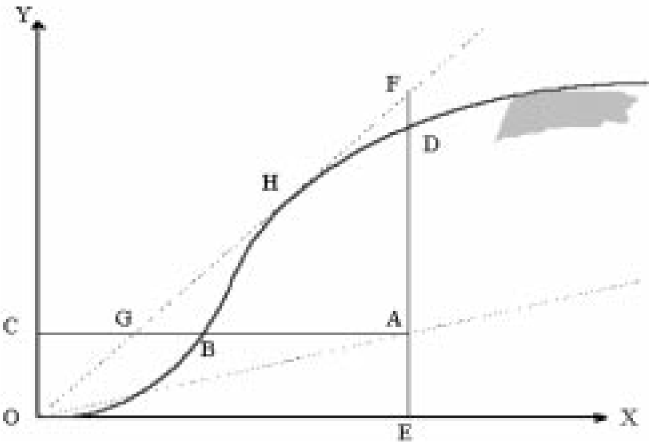


Figure A.1. Efficiency measures in input-output space. E_2=EA/ED, E_3=EA/EF, E_5=ED/EF.

The figure also illustrates the measure of technical productivity E_3 that is the ratio of the output-input ratio of observation A, the slope of the dashed line OA, and the maximal output-input ratio, the slope of the dashed line OH. Geometrically this can be seen to be equal to the ratio EA/EF. Technical productivity is sometimes termed gross scale efficiency, implying a comparison of actual production per unit of input behavior to the maximal production per unit of input had the production taken place at the technically optimal scale of point H. The estimate of this measure can be formulated as

$$\hat{E}_3 = \mathrm{Min}_\theta \left\{ \theta \left| (\gamma\mathbf{y}/\theta, \gamma\mathbf{x}) \in \hat{P} \right. \right\}, \qquad (A.3)$$

where γ is a free scalar. The inverse of the optimal value of γ is the scale indicator λ that measures the proportion of actual inputs to the inputs at the optimal scale (i.e. OF/OH in **Figure 1**). Finally we introduce the pure scale efficiency measure E_5, which is the ratio of the productivity of the technically efficient frontier point and the maximal productivity (i.e. ED/EF in **Figure 1**). The estimate is defined simply by

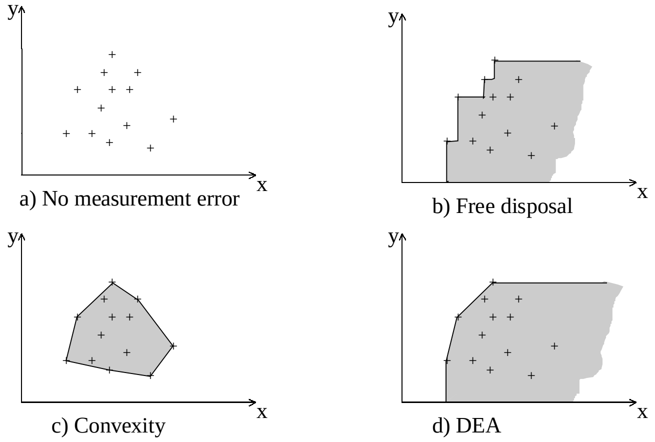
$$\hat{E}_5 = \hat{E}_3 / \hat{E}_2 \qquad (A.4)$$

One may note that if the production technology exhibits constant returns to scale (CRS), the frontier is a straight line from the origin, and the measures of technical efficiency and technical productivity coincide (E_2=E_3). This also implies that all observations are scale efficient (E_5=1).

The DEA estimate of the production possibility set is given by a set of linear constraints

$$\hat{P} = \left\{ \mathbf{Y}\lambda \geq y, \mathbf{x} \geq \mathbf{X}\lambda, \lambda \geq 0, \sum_{i \in N} \lambda_i = 1 \right\}, \qquad (A.5)$$

where $\mathbf{Y}$, $\mathbf{X}$ are the vectors or matrices of observed outputs and inputs and λ is a vector of reference weights. This corresponds to the formulation in Banker, Charnes & Cooper,[18] and is the minimum extrapolation estimator of the technology satisfying convexity, free disposability of inputs and outputs and feasibility of observed units, as illustrated in **Figure A.2.**



a) No measurement error b) Free disposal

c) Convexity d) DEA

Figure A.2. The DEA assumptions on the possibility set.

The calculations of DEA efficiency estimates are solved as a set of LP-problems by inserting (A.5) in (A.1). The shadow prices on the constraints associated with each variable in (A.5) are formally the derivatives

$$\omega_k = \frac{\partial E}{\partial y_k}, \ \omega_l = \frac{\partial E}{\partial x_l} \qquad \text{(A.6)}$$

but of more interest are the ratios ω_k / ω_l, etc, which then are the rates of substitution between the different inputs and outputs on the efficient frontier of the estimated feasibility set $\hat{P}$. If the behavior of each clinic is such that the allocation of inputs is cost minimizing, then this ratio would be equated to the factor price ratio, hence the use of the term shadow *prices*. Similarly, the ratio of an output shadow price and an input shadow price can be interpreted as the marginal product of that input with respect to that output, and the ratio of two output shadow prices is the relative resource cost of these products.

Appendix B. Testing DEA models

If no parametric assumptions are maintained about the inefficiency distributions, the Kolmogorov-Smirnov nonparametric test of the equality of two distributions is a suitable approximation. Applied to the distributions of i.i.d. efficiency estimates, and denoting the estimated cumulative distribution function of these as $S^0(E)$,$S^1(E)$, the statistic

$$D^+ = \mathrm{Max}_E \left\{ S^0(E) - S^1(E) \right\} \qquad \text{(B.1)}$$

is asymptotically distributed with a rejection probability of

$$\Pr\left(D^+ > \left(\frac{n^0 n^1}{n^0 + n^1} \right)^{-\frac{1}{2}} z \right) = e^{-2z^2}, \quad z > 0 \qquad \text{(B.2)}$$

which makes it applicable for testing one-sided hypotheses. [34]

The simple T-statistic[35] for the equality of group means for two samples of equal size *n* is:

$$T = \frac{\mathrm{Mean}_i\left(\hat{E}_i^1\right) - \mathrm{Mean}_i\left(\hat{E}_i^0\right)}{\sqrt{\dfrac{\mathrm{Var}_i(\hat{E}_i^1) + \mathrm{Var}_i(\hat{E}_i^0)}{n-1}}} \qquad \text{(B.3)}$$

which, if sample means are i.i.d. normal, is T-distributed with $2\,n$ - 2 degrees of freedom. By the central limit theorem the sample means will be approximately normal unless sample size is very small.

References

1. Verhulst FC, Berden GF, Sanders-Woudstra. Mental health in Dutch children (II): The prevalence of psychiatric disorder and relationships between measures. *Acta Psychiatr Scand* Suppl. 1985; **324**: 1-45.
2. Lavigne JV et al. Prevalence rates and correlates of psychiatric disorders among preschool children. *J Am Acad Child Adolesc Psychiatry* 1996; **35**: 204-214.
3. Halsteinli V. Nasjonale utviklingstrekk BUP. In Hagen H (ed). *Psykiatritjenesten 1998 - på rett vei?* SINTEF Unimed: Trondheim, 1999; Ch. 6: 81-89.
4. St meld 25 (1996-97): *Åpenhet og helhet. Om psykiske lidelser og tjenestetilbudene.* Ministry of Health and Social Affairs: Oslo, 1997.
5. St prp 63 (1997-98): *Om opptrappingsplanen for psykisk helse.* Ministry of Health and Social Affairs: Oslo, 1998.
6. *Nye alternativer i psykiatrien,* Norwegian Board of Health: Oslo, 1985.
7. Hatling T, J Magnussen. *Evaluering av arbeidsformer og produktivitet ved voksenpsykiatriske og barne- og ungdomspsykiatriske poliklinikker*. SINTEF Unimed: Trondheim, 1999.
8. Hagen H, Hatling T. Psykiatrien i Norden - en sammenligning. Vedlegg i St meld 25 (1996-97): *Åpenhet og helhet. Om psykiske lidelser og tjenestetilbudene.* Ministry of Health and Social Affairs: Oslo, 1997.
9. Nathan PE, Berge T, Høstmark Nielsen G. Treatment that works. *Tidsskrift for norsk psykologforening* 1999; **36**:617-619.
10. Newhouse JP. Reimbursing Health Plans and Health Providers: Efficiency in Production Versus Selection. *Journal of Economic Literature* 1996; **34**:1236-1263.
11. Debreu G. The coefficient of resource utilization. *Econometrica* 1951; **19**: 273-292.
12. Farrell MJ. The measurement of productive efficiency. *Journal of the Royal Statistical Society* 1957; **120**: 449-460.
13. Färe R, Lovell CAK. Measuring the technical efficiency of production. *J Econ Theory* 1978; **19**: 150-162.
14. Färe R, Grosskopf S, Lovell CAK. *The measurement of efficiency of production* Boston: Kluwer-Nijhof, 1985
15. Charnes A, Cooper WW, Rhodes E. Measuring th efficiency of Decision Making Units. *Euron J Oper Res* 1978; **2**: 429-444.
16. Seiford LM. Data Envelopment Analysis: The evolution of the state of the art (1978-1995). *Journal of Productivity Analysis* 1996; **7**: 99-137.
17. Førsund FR, Hjalmarsson L.On the measurement of productive efficiency. *Swedish Journal of Economics* 1974; **76**: 141-54.
18. Banker RD, Charnes A, Cooper WW. Some models for estimating technical and scale inefficiencies. *Management Science* 1984; **30**: 1078-1092.
19. Aigner, D.J. and S. Chu, On estimating the industry production function. *Am Ec Rev*, 1968; **58**: 826-839.
20. Aigner, D., C.A.K. Lovell, and P. Schmidt, Formulation and estimation of stochastic frontier production functionmodels. *Journal of Econometrics*, 1977; **6**: 21-37.
21. Banker RD. Maximum Likelihood, Consistency and Data Envelopment Analysis: A Statistical Foundation. *Management Science* 1993; **39**: 1265-1273.
22. Korostelev AP, Simar L, Tsybakov AB. Efficient estimation of monotone boundries. *Annals of Statistics* 1995; **23** (**2**): 476-489.
23. Korostelev AP, Simar L, Tsybakov AB. On estimation of monotone and

convex boundries. *Publications de l'Institut de statistique de l'Université de Paris* 1995; **39**: 3-18.

24. Coelli, T. and S. Perelman. *Efficiency measurement, multiple-output technologies and distance functions: with application t European railways*, CREPP 96/05. 1996, Université de Liège.
25. Valdmanis V. Sensitivity analysis for DEA models. An empirical example using public vs. NFP hospitals. *Journal of Public Economics* 1992; **48**: 185-205.
26. Magnussen J. Efficiency measurement and the operationalization of hospital production. *Health Serv Res* 1996; **31**: 21-37.
27. Kneip A, Park BU, Simar L. *A note on the convergence of nonparametric DEA efficiency measures.* Discussion Paper 9603, Institut de Statistique, Université Catholique de Louvain, 1996.
28. Simar L, Wilson PW. Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models. Management Science 1998; **44**: 49-61.
29. Simar L, Wilson PW. *Nonparametric tests of returns to scale.* Paper presented at the 5th European Workshop on Efficiency and Productivity Measurement, The Royal Veterinary and Agricultural University, Copenhagen, 1997.
30. Banker RD. Hypothesis Tests Using Data Envelopment Analysis. *Journal of Productivity Analysis* 1996; **7**: 139-159.
31. Kittelsen SAC. *Monte Carlo simulations of DEA efficiency measures and hypothesis tests.* Memorandum No. 9, Department of Economics, University of Oslo, 1999.
32. Grosskopf S. Statistical inference and nonparametric efficiency: A selective survey. *Journal of Productivity Analysis* 1996; **7**: 161-176.
33. Hackman ST, Passy U, Platzman LK. Explicit Representation of the Two-Dimensional Section of a Production Possibility Set. *Journal of Productivity Analysis* 1994; **5**: 161-70.
34. Johnson NL, Kotz S. *Distributions in statistics: Continous univariate distributions-2* Boston: Houghton Mifflin, 1970.
35. Bhattacharyya GK, Johnson RA. *Statistical concepts and methods* New York: John Wiley & Sons, 1977.