Benchmarking the Health Sector in Germany –
An Application of Data Envelopment Analysis

Matthias Staat, School of Economics, University of Mannheim, Germany
Maik Hammerschmidt, Business School, University of Mannheim, Germany

Research Paper, Institute for Market-Oriented Management, University of Mannheim

October 2000

JEL-Classification: I19, C69, D24

Keywords: Health care reform, benchmarking, relative performance evaluation, Data Envelopment Analysis

Abstract:

At present, a first round of hospital benchmarking as required by German law on health care reform takes place. After extensive discussions between hospitals and insurance companies, which are jointly responsible to deliver benchmarking results, a method with some peculiar characteristics was chosen. In this paper it is argued that the deficiencies of said method could be overcome by using Data Envelopment Analysis (DEA). The reasons that make DEA an advisable tool for policy decisions within the context of relative performance evaluation in the health care sector are discussed.

In order to illustrate the potential of nonparametric frontier estimation for hospital benchmarking in Germany, a comparison of hospitals, which provide the same basic clinical care, is carried out. Controlling for differences in the case mix and for possible heterogeneity of the services which hospitals provide, substantial productivity differences can be detected. Beyond simply identifying inefficient providers DEA leads to additional insight about the reasons of inefficiency and to useful management implications.

1. Introduction

Hospital costs are the largest portion of health related expenditures in Germany. In 1994, the year for which the data necessary for our analysis are available, outlays for health related expenditures amounted to 12.5% of GDP, a third of which were hospital cost.[1] Therefore, cost containment in the hospital sector is a key issue in stabilizing health cost at a sustainable level.

This has been recognized by German lawmakers and is reflected in recent reforms of and reform proposals for the German health system.[2] A key element of all reforms and proposals is to introduce some elements of a market mechanism into the financing system relevant for all German hospitals.[3] The planning mechanisms that dominate the health system to date are viewed to be one of the causes of steadily rising cost. Hospitals used to be paid a *per diem* plus fees for certain tasks they performed. This rather low powered regulation scheme[4] led to prolonged stays of patients in hospitals and redundant provision of tasks.[5] According to a new payment scheme which covers the most common diagnoses in the International Classification of Diseases (ICD) hospitals are now paid a treatment flat rate to cover variable cost, regardless of how long the stay and what efforts are needed to treat any particular referral.[6]

Clearly, a budgeting scheme based only on reported hospital cost ("input budgeting") has nothing to do with performance oriented resource allocation. An apparent solution would be to base payments on provided treatments/tasks ("output budgeting"). Such a budget allocation mechanism would separate the payments from cost and ensure only efficient providers to be able to generate positive profits. But in this case another problem arises because a budget paid completely independent from cost (a so called „global budget“) would assume, that at least some hospitals have a cost structure that allows gaining a positive profit. Even if this were to be the case, in a phase of a generally increasing demand for higher quality medical care or more advanced treatment methods an improvement of health care provision in the future would be blocked from the start.[7] Only by using performance measures that simultaneously integrate input **and** output variables (i.e., an efficiency measure!) it is possible to overcome

---

1 See Geil et al. (1997).

2 The reforms and the key arguments behind various reform strategies are well documented in Arnold and Paffrath (1995); Arnold and Paffrath (1996); Arnold and Paffrath (1997).

3 See Thomae (1995).

4 See Laffont and Tirole (1993).

5 The duplication x-rays already taken by the referring physicians in hospitals is one popular example.

6 For an overview on international systems of hospital financing, including Germany, see Stepan (1997).

7 See Knieps and Reichelt (1999).

these drawbacks. We will come back to this point in the sequel by introducing a method that calculates performance as an overall efficiency value.

It is well known that the question of how to set the appropriate treatment flat rate cannot be based on the status quo, i.e., by dividing total cost accrued by number and type of cases treated. The appropriate norm figure would have to be calculated on the basis of an efficient, „best practiced“ provision of services by the hospital. In Germany, the issue of how to set an appropriate fee is dealt with at present in various negotiation panels between individual hospitals and insurers. A negotiation round starts with a hospital reporting on its cost structure from which it derives the fees it demands. The insurance company accepts or declines the *per diem* and treatment flat rates calculated by the hospital from its outlays and in the latter case makes an offer based on some other, presumably more efficient, hospital’s cost structure, i.e., the benchmark or reference hospital. In case the hospital declines the insurer’s offer an arbitration panel will settle the case. If the panel fails to negotiate a compromise the hospital or the insurance company may take the matter to court. As of now, insurance companies have not won one single case of this type.

In addition to the fact that the system is hardly workable, i.e., that benchmarks provided by the insurers are often not accepted by the hospitals on grounds that reference hospitals cannot be compared to their own institutions, there is no guarantee when starting out at the status quo established under the planned system appropriate fees will be found through negotiations based on benchmarking and that in the long run maximum savings will be realized.

Efforts to enhance service, quality and efficiency have over the past decade gained momentum on all levels of government - and performance measurement is being promoted not only in the health care sector but also in the entire public administration sector. Assessing the efficiency impact of government programs (e.g. health care reform) and service providers (e.g. hospitals) is thus being studied with increasing intensity. Performance measurement is the key element within a new system of controlling of public organizations, which is well known under the term New Public Management.[8]

Most government agencies, as well as those private for-profit and nonprofit organizations delivering government services under grants and contracts, will become more and more involved in performance evaluation. Nevertheless, adopting productivity evaluation is still the exception rather than the norm in German government and, hence, there is a great potential to

[8] See Klingebiel (1997).

improve performance, accountability, and responsiveness by implementing systematic performance evaluation and by integrating performance information into regular policy and public management processes.

There are mainly two reasons for the impetus to implement performance measurement systems. First, performance dependent measures of budget allocation, granting, and contracting can be established, which will allow for a competition for financial resources. A second reason is an alignment concerning quality and efficiency of the provision of similar or homogeneous services by improving the accomplishment of the several providers. Analyzing the operations of best performing „reference providers“ offers insights that assist policymakers and policy evaluators to identify strategies on how to effect increases in performance.

In the health care sector public performance measurement was established in German federal law with § 3 and § 5 of the Hospital Cost Reimbursement Act (*Bundespflegesatzverordnung - BPflV*) of 1995.[9] But the political impact and economic effects of performance measurement will depend heavily on the choice of the evaluation method. With the wrong approach even a good end will lead to bad results. Or put in another way: Policymakers and lawmakers have to pay attention that the good ends do not justify bad means.

To be advisable for use in policy decisions a method for measuring and effecting performance should meet particular implementation criteria,[10] which are stated as follows:

(1) Flexibility in data selection

To be comparable across disparate programs and organizational units the performance measure should be independent from the scale level (continuous, ordinal, or categorical) and different units of analysis (monetary, time, scores, rates, quantity units).

(2) Robustness for data complexity

A meaningful evaluation tool has to take into consideration the multidimensional character of the performance construct. Performance cannot sufficiently be described by a set of single

[9] In the U.S. a performance measurement for the entire public sector is already legally established with the Chief Financial Officers Act (CFOA) of 1990, see the Government Performance and Results Act (GPRA) of 1993. For an overview on the state of the U.S. reforms concerning public performance measurement see Nyhan and Martin (1999) and also Stirring Committee for the review of commonwealth/state service provision (1997) for Australia. For an overview on the opportunities and approaches of public performance measurement see Nyhan and Marlowe (1995).

input-single output-ratios.[11] Otherwise, it may well happen that a provider performs as best practice on the one ratio but is inefficient in term of another ratio. If a government agency were forced to reduce the number of providers, then which one should be dropped?

(3) Competitive view

Performance measurement should be applied only as relative performance measurement. The efficiency value of a particular provider has to be calculated in comparison to the best practice player (benchmark), not the average performance. Each provider is judged against all other "competitors". Maximization instead of a satisficing assumption is needed.[12]

(4) General acceptance and fairness of the methodology

The measure of performance should be replicable and comprehensible. Therefore, the weights for the inputs and outputs should be derived endogenously, i.e., in an objective way and should not be influenced by the preferences and personal predispositions of (groups of) policymakers. In contrast, the need for a priori specification of weights does presumably not lead to acceptable and credible benchmarking results, because of the dependence on negotiation processes, which are influenced by power and non-pertinent considerations. Assigning deterministic values to each input and output variable upon which all decision makers finally agree could be an endless and resource consuming process.

The best practice provider against a particular unit is benchmarked has to be comparable. Only units with a similar input-output-structure should be compared to ensure that the performance evaluation is realistic and focuses not on extreme performance differences but on achievable best practice.

For the evaluation results to be easy to communicate and close to practice the benchmarks have to be real existing providers not hypothetical or prescriptive ideals.[13]

---

[10] For general implementation criteria for scientific models see Leeflang (2000), ch. 7. Data estimation methods (e.g., DEA as a nonparametric method) can be understood as measurement models that partly require the same "implementation criteria" as economic market models.

[11] See Young (1992).

[12] See Nyhan and Martin (1999).

[13] See Metzger (1993).

(5) Explanatory content

For a performance measurement method to be theoretically substantial it does not suffice to demonstrate that one provider is more successful than another. Beyond that it is necessary to explain these differences, i.e., to show the causes of the relative inefficiency.[14]

(6) Target orientation (comparisons of targets and achievements)

A relative performance evaluation should not only assess the current efficiency measure but also the future performance, which could be achieved. This means, information about the targeted efficiency position are needed.

(7) Management orientation

A relative performance evaluation should not only provide a target performance position but also an indication as to how a provider should attempt to vary its inputs and outputs so as to achieve a performance comparable to the best observed. Management strategies to improve performance may be developed. Output increases and / or input reductions necessary to achieve efficiency should be determined.

However, the ad hoc and theoretically unsound nature of the classic benchmarking technique is, as will be argued in the remainder of this paper, somewhat unsatisfactory concerning the above stated „implementation criteria“. And even if well suited to give to a service provider an indication of his efficiency relative to that of others w. r. t. one particular input and / or output it can hardly serve as a basis for negotiations with respect to hospital funding.

It is argued that a nonparametric method to estimate production frontiers known as Data Envelopment Analysis, which can be interpreted as a generalized and systematic form of benchmarking, could be used to evaluate the relative performance of public service providers and government agencies. By applying it to performance evaluation of German hospitals we show that DEA in contrast to simple benchmarking techniques meets the above mentioned conditions and thus is advisable as a basis for budgeting, granting, contracting, and other resource allocation decisions especially in the field of health care economics.

DEA is particularly suited for the analysis of technical efficiency of public sector service provision as no price information for inputs and outputs is needed to determine the degree of technical (in)efficiency in a multiple-input multiple-output setting. The DEA results can form the basis of a dynamic regulation scheme, which results in maximum efficiency gains in the long run. The method has become an innovative tool to evaluate the performance of hospitals and other public services.[15]

DEA is especially suited to benchmark hospitals in Germany according to legal requirements. The agreement between hospitals and health insurers on the relative performance evaluation of hospitals does call for the benchmarking of the hospital as a whole.[16] The methodology agreed upon (for a brief description, see 2.1.3) consists of aggregating individual data and to benchmark on this aggregate level. Entire hospitals can be evaluated by DEA in a much more transparent way, as will be explained below.

We will use data on German hospitals that contain the necessary information on the structure of the hospitals to ensure that only nearly identical hospitals with respect to their fields of specialization, case mix and services provided to patients are compared. In addition, as will be demonstrated below, DEA does benchmark only observations with similar input-output structure. Thus, the often raised concern that benchmarking results in meaningless comparisons does not apply to our benchmarking method.

The paper is organized in 6 sections. The next section describes some of the benchmarking approaches relevant for the methodological discussion about the official benchmarking procedure in Germany. The following section will illustrate the problems of benchmarking using a simple example. Next, DEA is introduced as a generalized benchmarking technique. This is followed by the presentation of DEA results for German hospitals. A final section briefly discusses the potential benefits of using DEA over simple benchmarking with respect to the long run optimality of the regulation of hospital cost compensation.

2. Relative performance evaluation of hospitals in Germany

After more than a decade of discussion on the issue of relative performance evaluation of hospitals in Germany (Krankenhausbetriebsvergleich) it became evident at the beginning of the 90's that German lawmakers were prepared to pass legislation forcing hospitals and insurance companies to cooperate on the issue. Hospital managers and owners had strong concerns about the fact that insurers had been able to compile ample information on the cost structure

---

[14] See Bauer et al. (2000).
[15] For the analysis of hospitals, see Hollingsworth et al. (1999).

of all hospitals. Given this, it is obvious why hospitals complain about the asymmetric distribution of information between them and the insurance companies. Efforts were undertaken to reduce the asymmetry by generating more information on the relative standing of individual hospitals.

These efforts include the so-called hospital compass devised by *führen & wirtschaften im Krankenhaus* (f&w), a periodical on hospital management.[17] Consultants also devised hospital benchmarks (*Henke / Paffrath / Wettke*) with data they had compiled.[18] The *Wissenschaftliches Institut der Ortskrankenkassen* (WIdO), the research institute run by the largest German health insurer, *Allgemeine Ortskrankenkasse* (AOK), set up its own procedure to benchmark hospitals. These three benchmarking approaches will be briefly described and discussed.

In addition to these, a number of other relative performance evaluation schemes are in operation. Best known are the Echolot by an independent consultant[19], as well as a survey approach used in a study published in *Focus* magazine.[20]

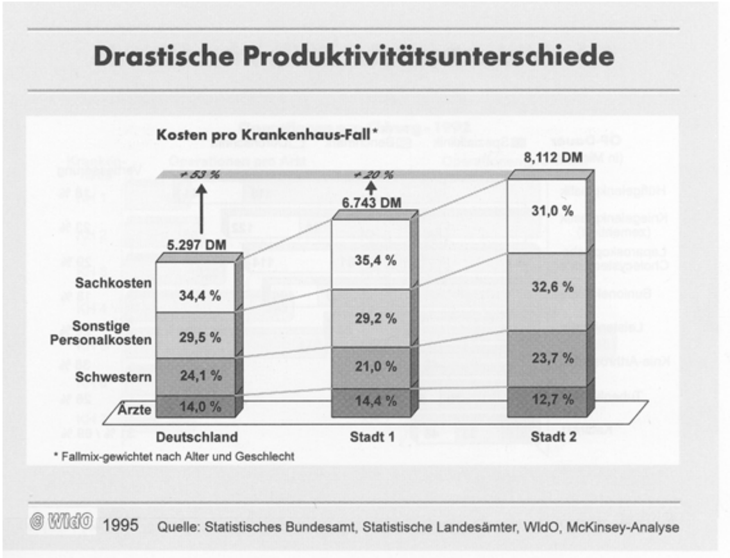
2.1. Approach of Henke/Paffrath/Wettke (1995)

This study discusses many potential benchmarking strategies for hospital performance. One of the examples based on German data in *Henke et al.* (1995) is a comparison of case cost for several ICDs across different cities in Germany.

---

16 See Deutsche Krankenhausgesellschaft (1999).
17 See Meurer (1995).
18 See Henke et al. (1995).
19 See Kaufmann and Wolf (1998).
20 These benchmarks are contained in Hildebrandt (1996).

Figure 1:
Differences in case cost across German cities[21]



The findings (see Figure 1) include a 53% difference for average case cost between one German city and the nation wide average. The authors claim that this "dramatic difference in productivity" is due to allocative and technical inefficiency. Furthermore, the authors maintain that differences between cities are not due to differences in case mix, which they hold can be accounted for by a weighting scheme (Wichtungsfaktoren). The data are weighted by patients' age and sex. The weighting scheme itself is not described nor is there an indication as to how the results derived by simple benchmarking were affected by the weighting scheme applied to them.[22]

2.2. f&w hospital compass

The second approach described here is the result of benchmarking program initiated by *f&w*. Participating hospitals are offered a relative performance evaluation of their data vis-à-vis other participating hospitals. In addition to benchmarks on average personnel cost and length of a stay, special indicators, e.g. for laboratory services, are surveyed from time to time. Between 200 and 300 hospitals take part in the hospital compass. The hospitals are divided into nine groups according to the scheme presented in Table 1.

---

21 Figure 1 reproduces graph 14-10 from Henke et al. (1995), p. 203.
22 See Henke et al. (1995), p. 206.

Table 1:
f&w hospital compass, grouping scheme[23]

Group	1	2	3	4	5	6	7	8
No. of beds	-200	200+	201- 400	201- 400	401-600	401-600	601 - 800	800+
No. of fields	- 3	3+	- 6	6+	- 8	8+		

Each hospital is provided with information on the mean and the distribution of the relevant parameters within its group.

2.3. Approach of Wissenschaftliches Institut der Ortskrankenkassen (WIdO)

The final benchmarking methodology outlined here is the official hospital program of the "Krankenhausbetriebsvergleich" carried out by the *WIdO*. The *WIdO* collects data, which hospitals are required to provide on their number and type of cases treated and the respective cost.[24] These data are more exhaustive than any other German data. They comprise more than 2000 hospitals.

The data are used to form groups of hospitals with similar case mix. The (dis)similarity of case mixes is measured by calculating the differences between the shares, which the leading ICDs have among the referrals to the hospitals. The homogeneity of the groups formed by means of the ICD-criterion is then estimated using cluster analysis (see Table 2). Homogeneity refers to the shares of the ICDs but not to the departmental or hospital structure. It is claimed that large and small departments may be compared in this way. The individual departments are benchmarked against the mean of the *per diem* and case cost of their respective groups. Also, these data are aggregated to derive a benchmark for the entire hospital.

2.4. Benchmark selection and comparability

Some aspects of the approaches described above merit further discussion. All three approaches use different criteria to judge whether hospitals can be compared in a meaningful way and thus can be benchmarked via the same reference units. *Henke et al*. (1995) consider

---

23 The entries in the Table read: -x „up to x" and x+ „more than x".

24 See Deutsche Krankenhausgesellschaft (1999).

all hospitals, which treat the same ICD as potential benchmarks for each other. According to these authors, existing differences can be handled by "weighting". In *f&w*, hospitals of similar size and structure are benchmarked without further weighting. *WIdO* does benchmark hospitals, which treat a similar mix of cases.

Instead of searching for exogenous indicators of similarity, which only can be incomplete it would be more appropriate to compare units with a similar input-output-structure, which is a direct measure of comparability. All relevant differences in hospital structure will be reflected in a specific input-output-structure. With DEA, the comparability is determined endogenously when calculating the relative efficiency score and no need for exogenous specification of similarity exists. The clustering of several groups of homogenous hospitals is achieved by the DEA method itself in a direct and transparent way as opposed to e.g., cluster analysis, a statistical technique designed to generate segments of maximum intra segment homogeneity and inter segment heterogeneity. Cluster analysis will not yield specific hospitals as benchmarks and allows only comparisons of single observations with segment means.[25]

In all three procedures benchmarks are selected which are not observed values but rather group means calculated from groups with possibly high intra-group heterogeneity. It is not clear if these group means are relevant for all hospitals in the group as a benchmark. Even if the group is homogeneous, taking means as benchmarks will render benchmarking partly meaningless. A mean is not produced by any single hospital, therefore an inefficient hospital will not be provided with a role model to guide its own choices. This, however, is one of the main advantages of benchmarking.

Second, and more importantly, a mean does not describe Best Practice. Therefore, even if all hospitals with below average performance (performance less than the mean) will improve their performance to match mean performance, potentially large efficiency gains – the difference between mean and best practice - will not be realized because convergence of common practice to Best Practice, although theoretically possible, will be achieved at a very slow rate.

Also, the use of weighting procedures will make benchmarking result hard to interpret if not unintelligible. A hospital ranked somewhere in the mid field will have to "re-engineer" its benchmarking results to find out why its performance is rated average, i.e. how the weights influenced the ranking. This destroys some of the advantages of the clear-cut information benchmarking ordinarily provides.

---

[25] The WIdO uses this method as a means of hospital benchmarking, see Gerste (1996).

Finally, standard benchmarking is an informative but nevertheless ad hoc procedure, which has no theoretical foundations. It is obvious that if one wants to proceed from detecting differences in the performance of hospitals to regulating hospitals such that efficient performance is rewarded and inefficiency is discouraged, it will be of advantage to apply a method which is strongly rooted in economic theory.

In the sequel, we will describe a methodology which benchmarks comparable units using observed (not mean) performance. Special emphasis will be put on the comparison of standard benchmarking techniques and DEA in order to make the advantages and disadvantages of the two approaches transparent.

3. Benchmarking vs. DEA

3.1. Some problems with standard benchmarking

In the sequel, an extensive treatment of DEA will be given such that the implications of the results presented should be easily understood. Further material on DEA can be found in the relevant literature.[26] The focus of this subsection are the advantages of DEA vis-à-vis benchmarking which make DEA advisable for policymakers and administrative decision makers as an adequate tool for relative performance evaluation. The following simple example will demonstrate that while DEA retains the two most desirable characteristics of standard benchmarking - best practice evaluation and comparison with observed as opposed to hypothetical practice - it does not have the above-mentioned undesirable features while it fulfills the implementation criteria.

The panel of graph 1 below illustrates the difficulties when benchmarking on more than one criterion. This is especially relevant for the German situation where a global hospital benchmark consisting of several (weighted) sub-benchmarks is required. The example will serve as a demonstration of what is possible with DEA but not with standard benchmarking. The two benchmarking criteria selected are case cost and length of stay. Of course, with a *per diem* the two parameters will both move in the same direction. However, there may as well exist a possibility substitution between time and money. For instance, minimal invasive surgery may be more expensive but will make shorter stays in the hospital possible.

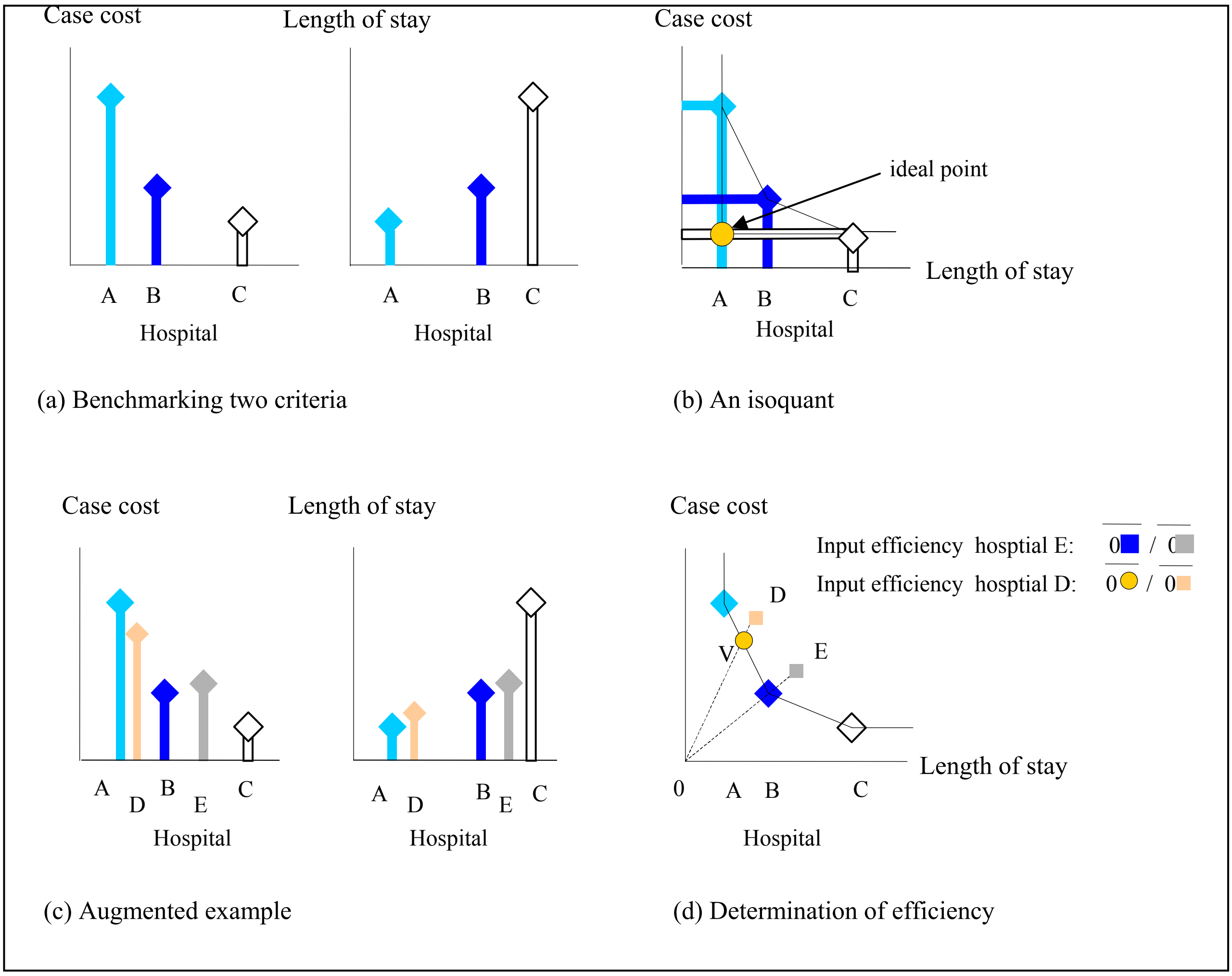
---

26 See for example Charnes et al. (1997).

Inspecting the part (a) in Figure 2 it would be hard to justify any ranking of the hospitals A, B and C. No single hospital dominates any other on both criteria. With standard benchmarking, the ranking would therefore depend completely on the relative weight assigned to the two parameters. The more weight is put on case cost, the better hospital C will be ranked; the more weight is put on length of stay, the better will hospital A's performance be ranked. Thus with simple benchmarking approaches (simple ratio analysis) the results can be manipulated by a specific and more or less arbitrary set of weights. A benchmark, which is chosen in this particular way, can hardly be a transparent and thus a defensible performance standard.

Part (b) illustrates a possibility to simultaneously benchmark the two criteria. Shifting the graph on length of stay from (a) by 90° and moving it into the case cost graph will produce the isoquant on the two inputs used to treat a case: time and money. The three hospitals A, B and C, none of which dominates any of the others are all located on the isoquant (efficiency frontier) and are therefore viewed as efficient producers. The lower panel (c) shows the same three hospitals plus hospitals D and E. The latter two lie above the isoquant in (d) and are therefore rated as inefficient producers. Also, it becomes evident that so-called "ideal points" of performance often proposed in benchmarking studies may not be feasible given current technology.

Figure 2:
Limits to standard benchmarking



(a) Benchmarking two criteria

(b) An isoquant

(c) Augmented example

(d) Determination of efficiency

The degree of efficiency can easily be calculated, see figure (d) constructed in analogous fashion from figure (c) as figure (b) from figure (a). The simplest case is that of hospital E, which is completely dominated in its performance by hospital B. The latter has lower case cost and shorter length of stay. In the case of hospital E, being dominated by hospital B, efficiency is calculated as follows: the closer any input combination for a given output is to the origin, the more efficient is the production of this output. Consequently, one may compare the relative efficiency of the two producers or hospitals by comparing the distance of their input combinations to the origin. In the case of hospitals B and E, this relative efficiency is calculated as the ratio of the two distances, i.e. $\overline{OB}/\overline{OE}$.

Note that this is the maximum efficiency, which can be ascribed to hospital E given its combination of inputs. Point B is the input combination observed, which is closest to E. Therefore, the efficiency of E would be lower if benchmarked relative to any other reference point, e.g., the “ideal” reference point mentioned above.

For hospital D efficiency is measured by the ratio of the distance from the origin to point D and the distance between the origin and the point on the isoquant next to D, i.e., the point on the isoquant crossed by the line between the origin and D. For inefficient producers, the efficiency measure will therefore always be strictly smaller than 1, for efficient producers it will always be equal to 1. The efficiency of D will be equal to $\overline{0V}/\overline{0D}$. Farrell (1957) has proposed this ratio of distances as a measure of efficiency. It corresponds to Shepard's distance developed in Shepard (1970). Thus, this way of "benchmarking" productive efficiency is firmly based on microeconomics production theory. We will develop DEA in a formal way in the next section.

3.2. DEA-related techniques

DEA is similar to econometric techniques that estimate production frontier functions, i.e., to stochastic frontier analysis (SFA).[27] Estimating the production function underlying the observed combinations of inputs and outputs DEA aims at the same goal as SFA. It does so by evaluating the efficiency of each producer or, in DEA parlance, decision-making unit (DMU) through comparison of its set of inputs and outputs to that of any other producer. Instead of deriving a hypothetical production frontier a comparison with observed best practice determines whether a producer is rated efficient or not. This makes it similar to benchmarking and in fact, DEA can be interpreted as a generalization of simple benchmarking techniques.

3.3. Illustration of the basic principles

In the single-input single-output case, DEA is very similar to benchmarking. However, DEA is capable of handling multiple-input multiple-output situations systematically. It is also capable of generating useful information on how to improve on the efficiency position of an inefficient hospital. For instance, with standard benchmarking one would be unable to compare hospitals that take different approaches to handling certain cases. If there is a choice between conservative treatment, surgery and minimal invasive surgery (as might be the case for the rupture of a ligament) different hospitals might prefer different forms of treatment. The conservative alternative may be the cheapest but may also take the longest whereas minimal invasive surgery will be the most expensive but will also make very short stays possible. Standard surgery will be in between the two extremes.

Instead of applying the same vector of weights to the parameters for all hospitals, as would be the case with standard benchmarking, DEA assigns an individual, mathematically optimal vector of weights to each hospital in a way that maximum weights are placed on those variables where a provider compares favorably and minimum weights are placed on those variables where it compares unfavorably. Below, input weights are denoted v_i and output weights u_r with index i for inputs and index r for outputs whereas index j runs over the hospitals. The sum of these weighted output-to-input-ratios, h, is then maximized under the restriction that no other hospital achieves a score greater than 1 (or any other number chosen for normalization) with the weights that maximize the index of the hospital that is being evaluated. This results in the following expression:[28] The figure for the hospital -or more generally DMU- being evaluated are indexed with 0.

$$\max_{u,v} \quad h_0 = \frac{\sum_r u_r y_{r0}}{\sum_i v_i x_{i0}} \quad \text{s.t.} \quad \frac{\sum_r u_r y_{rj}}{\sum_i v_i x_{ij}} \leq 1. \tag{1}$$

The formulation of the problem makes clear that the weights underlying the comparison between DMU_0 and the other DMUs maximize the efficiency of DMU_0. Comparatively high y's therefore will carry high u's, high x's on the other hand will have low v's.

Proceeding in this way, all three hospitals in the example would be rated efficient. The hospital using minimal invasive surgery (hospital A) would be assigned high weights v_l for the parameter length of stay and a low weight v_c would be assigned to its high input (cost). The hospital preferring the conservative approach (hospital C) would be assigned a low v_l to length of stay and a high v_c to its low cost. Each hospital is considered efficient in its own way. All hospitals would then be part of the efficient production frontier. In standard benchmarking, as stated above, only one hospital would be rated efficient or different hospitals would be rated efficient by different criteria.

One would be unable to observe which hospitals employ surgeons are capable of performing minimal invasive surgery. Therefore, it makes sense to allow for several alternative input-output-combinations to be rated efficient. Using DEA, it does not really matter for the efficiency position of the hospital whether or not its surgeons are capable of one particular technique as long as they employ their method of treatment in an efficient way and as long as this method

---

[27] For a comprehensive treatment of both methodological approaches, see Coelli et al. (1998).

[28] See Charnes et al. (1997).

is not completely dominated (slower and more expensive) by other available techniques. The use of such dominated techniques would correctly be interpreted as a sign of inefficiency.

3.4. Formal description

Continuing the formal discussion of DEA, the optimization problem (1) can be transformed into a common linear programming problem by a change of variables. The problem on the right of formula (2) below (in vector notation) corresponds to equations (1) above[29] and the primal program on the left is the input oriented formulation of the same problem. Here, efficiency is measured as the maximum input reduction possible for an inefficient producer if he produced by means of a technology, which would put him on the efficient frontier.

$$
\begin{array}{rl}
\min\limits_{\theta,\lambda,s^+,s^-} & z_0 = \theta - \varepsilon s^+ - \varepsilon s^- \\
\text{s. t. } & Y\lambda - s^+ = Y_0 \\
& \theta X_0 - X\lambda - s^- = 0 \\
& \vec{1}\lambda = 1 \\
& \lambda, s^+, s^- \geq 0
\end{array}
\qquad
\begin{array}{rl}
\max\limits_{\mu,\nu} & w_0 = \mu\, Y_0 + u_0 \\
\text{s. t. } & \nu X_0 = 1 \\
& \mu Y - \nu X + \vec{1}u_0 \leq 0 \\
& -\mu \leq -\varepsilon \\
& -\nu \leq -\varepsilon
\end{array}
\qquad (2)
$$

The parameters μ and ν are the analogues of the weights *u* and *v* in the ratio formulation and ε is a non-archimedian.[30] The interpretation of μ and ν is that of prices on which the optimization of DMU_0 with respect to its inputs and outputs would have been based. In the input-oriented formulation on the left, the slacks for output and input are labeled s^+ and s^-, respectively, and an efficient peer unit enters the reference technology for DMU_0 with a factor λ.

In case a DMU is efficient, its reference technology is its own input-output-combination and we have λ = 1. For an inefficient DMU, some λ-weighted average of efficient DMUs will be able to produce at least as much output ($Y\lambda - s^+ = Y_0$) using only a fraction θ of its inputs ($\theta X_0 - X\lambda - s^- = 0$). The score θ indicates by which common factor all inputs could be reduced if an efficient production technology was used. The formulation above on the left is the one on which the results to be presented are based.

---

29 Formula (1) is a constant returns to scale formulation that was chosen to illustrate the similarity to benchmarking whereas (2) represents a variable returns to scale formulation that will be used in the analysis.

30 In principle, the solution depends on the value of ε, usually chosen between 10^{-4} and 10^{-6} but the issue is of little practical relevance (see Olesen and Petersen, 1994). A two-stage algorithm determining first the efficiency value and during the second stage the slacks without the necessity of introducing the non-archimedian exists.

Furthermore, the results of a DEA consist of the slacks as well as the corresponding virtual multipliers or shadow prices from the solution of the dual problem. Slacks indicate that inputs would have to be reduced by more than (1 - θ) to match the input use of the reference hospital and that the outputs of the reference technology exceed those of DMU_0, respectively. Thus, DEA also gives directions on how to improve the efficiency of any DMU.

3.5. Extensions

DEA can also allow for parameters, which are important in determining efficiency but cannot be controlled (at least not in the short run) by the DMU. This makes the comparison of DMUs, which face different environmental settings, possible. Standard benchmarking techniques would necessitate "weighting" and thus give results, which are not readily interpretable.

In the case of German hospitals, a non-controllable parameter would be the number of beds. An authority of the federal state the hospital is located in determines the size of the hospital. The extension of DEA allowing for the use of non-discretionary indicator variables is developed in Banker and Morey (1986).[31] These non-discretionary parameters enter the optimization but do not directly determine the efficiency score, θ. This alters the formulation of the left hand side expression in equation (2) in the following way:

$$\begin{aligned} \min_{\theta,\lambda,s^+,s_D^-,s_I^-} \quad & z_0 = \theta - \varepsilon s^+ - \varepsilon s^- \\ \text{s. t. } Y\lambda - s^+ &= Y_0 \\ \theta X_{0D} - X_D\lambda - s_D^- &= 0 \\ X_{0I} - X_I\lambda - s_I^- &= 0 \\ \lambda, s^+, s_D^-, s_I^- &\geq 0 \end{aligned} \tag{3}$$

Efficiency is now defined by the maximum input reduction possible with respect to the parameters under the control of the DMU, X_D. The non-discretionary parameters X_I do enter the optimization indirectly such that no reference technology is allowed to have higher inputs

[31] See Ruggiero (1996) and Staat (1999) for recent discussions on the topic of non-discretionary variables.

than X_{I0}. They do not play a role in calculating θ other than restricting the reference technology.

4. Data

The data on German hospitals that will be analyzed are taken from the 1995 and 1996 issues of the Krankenhaus-Report. All data on hospitals refer to the years 1994/95, i.e., before the current system of financing hospitals was introduced. Individual hospitals are not anonymous but can be identified, i.e., name, location etc. are known. More than 2000 hospitals are contained in the 1994 raw data set and some 1800 in the 1995 data, 1700 hospitals were contained in both data sets.

The information[32] available for the 1700 hospitals includes type of ownership (public, private or a non-profit), size (number of beds), structure of the hospital (number and type of departments) and the *per diem*. Unfortunately, the *per diem* is available only for the entire hospital and not for the individual departments. The *per diem* indicates the cost of the provision of a bed for a day by a specific hospital. This cost will vary with the structure and the facilities of a hospital. Therefore, to make our analysis meaningful, it is important to ensure that only hospitals with a very similar structure and very similar facilities are compared.

Hospitals with (nearly) identical departmental structures are grouped into one of 77 different structure classes. They are also grouped in four classes according to their function in the German health system: hospitals of only local importance providing basic care without any large scale technical facilities (type I), basic care hospitals with some technical facilities that are of regional importance (type II), hospitals with various departments which are of central importance for the region (type III) and hospitals with a maximum of facilities (type IV).[33] Type I, local, and type II, regional, hospitals are relevant for this analysis.

The average length of stay is known for each of the following five departments: internal medicine, surgery, gynecology, orthopedics and ENT. In addition, the hospital data are subdivided in clusters according to the case mix of their referrals.

The available information is used in the following way: To ensure that only hospitals that are absolutely comparable are contained in the data we restrict the analysis to 3 out of the 77

---

[32] For a full description see Arnold and Paffrath (1995), p. 273 ff. and Arnold and Paffrath (1996), p. 279 ff.

[33] Definitions taken from Steinmetz (1991), p. 41 f. and translated by the authors.

structural groups (numbers 11 to 13)[34]. These are hospitals that provide basic care and have two main departments: one for internal medicine and one for surgery. The only difference between the three groups is that the fraction of patients who are treated by external specialists differs.[35] The hospitals are either of type I (local) or of type II (regional); therefore, the data are split into two sets. This leaves 160 hospitals, 108 of type I and 52 of type II.

The indicators used to compare these hospitals are the *per diem*[36] and number of beds as inputs. As the number of beds cannot be changed in the very short run it is treated as a non-discretionary variable.

Output is modeled by number of cases processed per year and average speed of treatment in the two fields of specialization.[37] The speed of treatment measure is calculated as one over average length of stay in each of the two departments. If capacity use were at 100% the two indicators would be redundant given the size of the hospital. But average capacity use in hospitals is well below 100%.[38]

To control for differences in case mix we classify hospitals according to the fraction of different leading (three digit) ICDs in each department. Surgery divisions are subdivided in 6 case mix clusters (see Table 2).[39] The only criterion by which one of the clusters differs substantially from the others is intensive care days per 100 cases.[40] The cluster with the most heterogeneous case mix (code 99) has about twice as many intensive care days (59.1) as classes 1 to 5 (between 22.6 and 33.1 days). This is reflected in the comments made by Gerste: "Although the hospitals treat different case mixes there is no difference in the services they provide. ... Only the heterogeneous classes with their high values for intensive care stand out".[41]

For the internal medicine department again only the heterogeneous group that has an average patient stay of 16.1 days stands out from the other three groups where length of stay ranges

---

34 See Arnold and Paffrath (1995), p. 274.

35 The fraction is below 10% for group 11, between 10% and 20% for group 12 and over 20% for group 13.

36 Average treatment cost is also available. To check the robustness of the results, the analysis that follows was carried out with this cost measure. There were only slight differences between the two data definitions as far as the results on efficiency are concerned.

37 Cases per year are only listed in intervals of 1000 cases, which may create some imprecision. We therefore used an alternative measure calculated as number of beds times days per year times average capacity use (see Table A 2 in the appendix) and reran the analysis to check the robustness of the results. No significant between the two models were detected and the subsequent analysis was carried out using the number of cases listed.

38 See Table A 2 in the appendix.

39 See also Gerste (1996), chap. 11.2 as well as Arnold and Paffrath (1996), p. 281 for a description of the distribution of ICD over case mix clusters.

40 See Table 11-3 in Gerste (1996), p. 123.

between 12.9 and 13.8 days.[42] For each department a dummy for adverse case mix is set to one if the hospital treats a heterogeneous case mix.

Table 2:
Characteristics of case mix clusters

Department / ICD-Cluster	Code	Length of stayin days	Intensive care days in 100 cases
Surgery			
Inner knee-joint	1	10.6	22.6
No leading ICD.	2	11.2	30.0
Hernia; cholelithiasis	3	11.3	28.1
Arteriosclerosis; varicose veins	4	11.4	33.1
Commotion; appendicitis	5	11.2	28.6
Rest	99	10.8	59.1
Internal medicine			
Low fraction of common ICDs	1	13.8	41.7
No leading ICD	2	13.5	42.4
Chronic heart disease (ICD 414)	3	12.9	39.0
Rest	99	16.1	39.9

As mentioned above, the hospitals may admit patients who are treated by external specialists, for instance an ophthalmologist or an urologist. This does not affect the measure of length of patient stay of, say patients of the surgery ward of the hospital, directly but may complicate hospital management in general. To take this into account, an indicator consisting of the number of all fields of specialization including those represented by external specialists is included.

The set of two input and five output indicators is used to assess the efficiency of the hospitals. Table 3 contains some summary statistics of the data. The fact the hospitals of regional importance are slightly larger than the ones of local relevance is reflected in their higher figure for maximum number of beds and cases treated. Hospitals of regional importance have slightly more fields of specialization than only locally relevant hospitals. All other indicators have very much the same range with the exception of average length of stay in the surgery department, which takes a maximum of three weeks in the local and of only two weeks in the

[41] Gerste (1996), p. 123, translated by the authors.

regional hospitals. The difference in average lengths of stays is not statistically significant, however. Treatment duration is longer for the internal medicine departments and shorter for the surgery departments of the type I hospitals when compared to the hospitals of type II. A significant difference for the means of the dummy variables for adverse case mix indicates a more heterogeneous case mix which surgery departments in regionally important hospitals handle. As mentioned above, the two types of hospitals are analyzed separately in the sequel.

---

[42] Data kindly provided by *B. Gerste*, *WIdO*.

Table 3:
Summary statistics

	variable	mean	s. d.	min	max
Local hospitals	per diem	393.61	51.67	260.56	574.76
	avg. cost per case	4736	844	3500	8000
	number of beds	162	44	70	265
	cases treated p. a.	4907	1531	2000	8000
	avg. length of stay	11.39	1.86	7.5	18.4
	avg. length internal med. dept.	12.39	1.95	7.3	19.1
	avg. length surgery dept.	10.83	2.22	6.6	21.1
	adverse case mix internal med.	.019	.14	0	1
	adverse case mix surgery	.046	.211	0	1
	total no. of fields of specialization	3.71	1.15	2	8
Regional hospitals	per diem	389.77	50.58	310.39	581.75
	avg. cost per case	4606	1063	3000	10000
	Number of beds	196	73	50	441
	cases treated p. a.	6192	2368	2000	14000
	avg. length of stay	11.26	1.80	7.9	20.5
	avg. length internal med. dept.	11.89	2.01	7.8	16.5
	avg. length surgery dept.	11.13	1.66	7.7	15.6
	Adverse case mix internal med.	.0577	.24	0	1
	Adverse case mix surgery	.212	.417	0	1
	total no. of fields of specialization	4.19	1.192	2	7

Only very few hospitals face an adverse case mix of patients. This reflects the fact that the hospitals in the data belong to the segment of the hospital system that provides the most basic form of clinical care. Cases that are viewed as very complex or unclear are mostly referred to more advanced (type I or type II hospitals with a broader range of departments or type III or IV) hospitals. Only among the surgery departments of the type II hospitals (see appendix, Table A 1) are there nearly 20% observations with an adverse case mix. For the hospitals handling an adverse case mix, *per diem* is higher (though not statistically significant) than for the other hospitals whereas average length of stay is the same for both.

5. Results

5.1. Efficiency distribution in the hospital samples

Table 4 below shows a summary of the efficiency scores of the hospitals analyzed. There are 28 efficient hospitals among the 108 type I hospitals; 18 type I hospitals have an efficiency below 75%. Average or structural efficiency is 86% for type I and 89 % for type II (regional) hospitals. A higher fraction of the latter (about a third) is efficient. Eight type II hospitals have an efficiency rating below 75%. The results are very much in line with results found by previous DEA studies on hospitals.[43]

Because of the fact that the cost information available is not too detailed, there would be some cause for concern if the hospitals analyzed were not homogeneous enough to be comparable. Then, some hospitals may be rated systematically as (in)efficient simply because they are not truly comparable with the rest of the data in the sample. Table A 3 in the appendix gives no evidence that the data consist of a heterogeneous mixture of hospitals.

Table 4:
Efficiency scores

	observed	$\theta = 1$	mean	s.d.	$\theta < 75\%$	min	max
type I hospital	108	28	.86	.1256	18	.48	1
type II hospital	52	17	.89	.1175	8	.58	1

The efficient and the inefficient hospitals appear to be strikingly similar. Average *per diem* differs by less than 20 DM, number of beds is smaller for efficient type I hospitals and larger for efficient type II hospitals than for the respective inefficient counterparts. Average length of stay in the internal medicine department is even longer for efficient type II hospitals than for inefficient ones.

Thus, there is no single parameter that makes efficient hospitals stand out from the rest; rather it is a combination of inputs and outputs similar to inefficient hospitals that nevertheless cannot be matched by other hospitals, which makes a hospital efficient. There is no way to assure

[43] See Hollingsworth et al. (1998). Byrnes and Valdmanis (1997) evaluat 123 hospitals and find 6 hospitals to be efficient with an average efficiency of 61%. The lowest pure technical efficiency in their sample was 48%.

that hospitals, which are efficient in this way, will be identified by any (weighted) standard benchmarking technique. It is interesting to note that a benchmarking technique placing a high weight on length of stay in the medicine department would result in a bad rating to efficient type II hospitals.

The three structure groups also do not seem to be play a role in explaining the differences in productivity among the hospitals analyzed as can be seen by inspection of Table 5.

Table 5:
Efficiency distribution over structure groups

	inefficient	efficient	Total
structure group	number (rel. freq.)	number (rel. freq.)	number
11	32 (76.19%)	10 (23.81%)	42
12	52 (73.24%)	19 (26.76%)	71
13	30 (63.83%)	17 (36.17%)	47
Total	114 (71.25%)	46 (28.75%)	160

Table 6 reveals the influence of the type ownership on efficiency. One can conclude that the hospitals run by non-profit organizations are managed less efficiently than other hospitals.

Table 6:
Efficiency by type of ownership

type of ownership	inefficient		efficient		total
	number	rel. freq	number	rel. freq	number
non-profit	58	81.69%	13	18.31%	71
Private	3	60.00%	2	40.00%	5
Public	53	63.10%	31	36.90%	84
Total	114		46		160

5.2. Results for individual hospitals

The following paragraphs provide an example of how the degree of inefficiency of an individual hospital is determined by DEA.

5.2.1. Efficient hospitals

The first three efficient hospitals (code K77, K110, K148) contained in Table 7 are the ones dominating the hospital K88. Hospital K148 is a smaller type I hospital of local importance with a very low *per diem*. Length of stay is well below average in both departments, it has one additional field of specialization and the number of cases it treats is in line with the cases calculated as described above. Its low cost and short patient stays make it efficient. Hospitals K110 differs from hospital K148 in that it only has average length of stay for its internal medicine patients and above average for its surgery patients but the minimum *per diem* observed in these data which is what makes it efficient. It is smaller than hospital K148. Hospital K77, on the other hand, achieves efficiency by generating close to maximum output with respect to cases and fields of specialization. The patient stays in this hospital, however, are longer than average.

Table 7:
Efficient peers and inefficient hospitals

	type I hospitals				type II hospitals	
	efficient			inefficient	efficient	inefficient
Code	K77	K110	K148	K88	K125	K4
per diem	334.34	260.6	294.6	349.9	342.2	393.9
treatment flat rate	4500	3500	3500	4000	4000	4500
number of beds	262	80	105	158	200	210
number of cases	8000	2000	4000	5000	6000	6000
calculated number cases	6140	1730	3400	4503	7400	5894
avg. stay internal	13.9	12.4	9.3	11.5	10.6	11.9
avg. stay surgery	12.5	15.5	7.8	10.5	8.2	9.7
adverse case mix med. Dept.	-	-	-	-	-	
adverse case mix surgery dept.	-	-	-	-	adv.	
number of specialisation fields	6	3	3	4	4	4
θ	1	1	1	.86	1	.87
ownership	non-profit	public	public	non-profit	public	non-profit
structure group	13	11	12	12	13	11

Hospital K125 is an efficient hospital of the regional type, which differs from other efficient hospitals in that it faces an adverse case mix among its surgery patients. It combines relatively low cost with below average lengths of stay.

Table 8:
Virtual multipliers for inefficient hospitals[44]

				treatment speed		adverse case mix		
code	*per diem*	beds	cases	internal	surgery	internal	surgery	nr. fields
K88	-2.86		.03		.46	.25		
K4	-2.54						.16	.08

Three of the efficient hospitals listed are owned by public institutions, one by a non-profit institution. With one exception, they are like most of the hospitals analyzed here located in

44 *Per diem* was divided by 100 when calculating the efficiency scores. The virtual multipliers have to be interpreted accordingly. A 10 DM reduction in *per diem* would mean a 25% increase in efficiency. The number of beds was also divided by 100, cases were divided by 1000.

communities with a population of 20.000 or less. The inefficient hospitals, on the other hand, are all owned by non-profit organizations and are located in communities of similar size.

5.2.2. Inefficient hospitals

It would be interesting to know how the inefficient hospitals can move away from their inefficient position. The efficiency scores and the weights, which indicate the relevance of an efficient peer for a certain reference technology, give an indication of the degree of inefficiency of a hospital and therefore indicate potential cost savings. The virtual multipliers which are also part of the DEA results, on the other hand, show whether changes of parameters in the appropriate direction will lead to an immediate improvement in efficiency –when multipliers are non-zero- or whether small changes in parameters values will lead to no immediate effect –when multipliers a equal to zero. Thus, the question what influences the efficiency position of an inefficient (in fact, any) DMU is best answered by looking at the virtual multipliers.

Hospital K88 was included in Table 7 and Table 8 as a typical inefficient type I hospital. It could reduce its inputs by 14% (which is about average) were it to produce with the technology of the virtual reference hospital. Its reference technology consists of hospitals K77, K110 and K148 where the set of weights for this virtual DMU is $\lambda_{88} = \{.37, .24, .39\}$.

The set of virtual multipliers is listed in Table 8. The non-zero multipliers indicate that K88 could change the parameter marginally to change the efficiency position. For instance, its surgery department is relatively efficient according to the length of stay criterion. The current virtual DMU could not dominate K88 with the same set of weights were this parameter to be improved slightly. Hence, θ_{88} would change if its average stay in its surgery department could be shortened. Treating more cases would also improve the efficiency of K88.

An interesting special case is type II hospital K4. It is, as can be seen comparing the data in Table 7, completely dominated by K125.

5.3. Calculation of the savings potential

Having made transparent how the results were derived and how individual hospitals can interpret the findings with respect to their efficiency position the efficiency scores are now used to calculate possible savings for the class of hospitals analyzed. One has to remember that the efficiency score is the input contraction (cost reduction), possible if an inefficient hospital applies an efficient production technology. Therefore, efficiency score times cost gives efficient cost. Since the *per diem* was used as a cost indicator in the analysis total hospital cost are now calculated as *per diem* times average length of stay times cases. Individual hospital cost is multiplied with the efficiency score to give efficient cost. Summing over all hospitals, efficient cost are only 83% of current total cost for the type I hospitals. For the type II hospitals, the ratio of efficient to current cost is 92%.

6. Conclusions

The main finding of the study is that according to 1994/95 data significant productivity differences between very similar hospitals exist. The most inefficient hospitals could half their cost by applying an efficient technology.

The average productivity was found to be 86% for the smaller (type I) hospitals and cost savings calculated on this basis amounted to 17% of total hospital cost. This is well in line with other DEA studies on relative hospital efficiency.

Ownership seems to play a role for the efficiency of a hospital, as the hospitals run by non-profit organizations seem to be less efficient than other hospitals.

Once the results have been derived there is one particular advantage in comparison with figures established through simple benchmarking. The DEA results can be used to evaluate the present reimbursement system for hospitals against an optimal regulation system. An optimal regulation under asymmetric information can by based on DEA results using the concept of yardstick competition. A regulation scheme based on the concept of yardstick competition in conjunction with DEA has been described and successfully implemented before, for the regulation of service providers.[45]

Hence the main advantages of DEA over the current system of relative performance evaluation of hospitals in Germany can be summed up as follows: DEA provides systematic instead of ad hoc evaluation, is able to evaluate entire hospitals based on disaggregated data, is thus transparent an is firmly based on production theory.

[45] See Bogetoft (1997).

References

Arnold, M. and D. Paffrath (Eds.) (1995) Krankenhaus-Report ´95 Schwerpunkt: Neue Versorgungsstrukturen. Stuttgart, Jena.

Arnold, M. and D. Paffrath (Eds.) (1996) Krankenhaus-Report ´96 Schwerpunkt: Managed Care. Stuttgart, Jena.

Arnold, M. and D. Paffrath (Hrsg.) (1997) Krankenhaus-Report ´97 Schwerpunkt: Sektorübergreifende Versorgung. Stuttgart, Jena.

Banker, R. D. and R. C. Morey (1986) Efficiency analysis for exogenously fixed inputs and outputs. *Operations Research 34(4)*: 513-521.

Bauer, H. H., M. Staat and M. Hammerschmidt (2000) Ein integriertes Konzept für Markt-segmentierung, Produktpositionierung und Benchmarking. Working Paper W 45, Institute for Market-Oriented Management, University of Mannheim.

Bogetoft, P. (1997) DEA-based yardstick competition: The optimality of best practice regulation. *Annals of Operations Research 73*: 277 – 298.

Byrnes, P. and V. Valdmanis (1997) Analyzing technical and allocative efficiency of hospitals. In A. Charnes et al. (Eds.) Data Envelopment Analysis: Theory, Methodology, and Application. 3rd ed. Boston: 129-144.

Charnes, A., W. W. Cooper, A. Y. Lewin and L. M. Seiford (Eds.) (1997) Data Envelopment Analysis: Theory, Methodology, and Application. 3rd ed. Boston.

Coelli, T., D. S. Prasada Rao and G. E. Battese (1998) An introduction to efficiency and productivity analysis. Boston.

Deutsche Krankenhausgesellschaft (1999) Vereinbarung nach §5 BPflV zum Krankenhausvergleich. http://www.dkgev.de/1_khv.htm

Färe, R., S. Grosskopf, B. Lindgren and P. Roos (1997) Productivity developments in Swedish hospitals: a Malmquist output index approach. In A. Charnes et al. (Eds.) Data Envelopment Analysis: Theory, Methodology, and Application. 3rd ed. Boston: 253-272.

Farrell, M. J. (1957) The Measurement of Productive Efficiency. *Journal of the Royal Statistical Society, Series A, 120(3)*: 253 – 290.

Geil, P., A. Million, R. Rotte and K. F. Zimmermann (1997) Economic Incentives and Hospitalization in Germany. *Journal of Applied Econometrics 12*: 295 – 311.

Gerste, B. (1996) Bildung von Krankenhausgruppen auf Fallmix-Basis. In M. Arnold and D. Paffrath (Eds.), Krankenhaus-Report ´96. Stuttgart, Jena: 115–126.

Greißinger, P. and K. J. Schmidt (1998) f&w-Krankenhaus-Kompaß: aktuelle Trends und Konsequenzen. *führen & wirtschaften im Krankenhaus 15*: 306–309.

Henke, N., D. Paffrath and J. Wettke (1995) Benchmarking im Krankenhausmarkt. In M. Arnold and D. Paffrath (Eds.) Krankenhaus-Report ´95. Stuttgart, Jena: 191 – 210.

Hildebrandt, H., H. Bexfield and G. Besser (Eds.) (1996) Krankenhäuser in Deutschland: ein Vergleich. Sankt Augustin.

Hollingsworth, B., P. J. Dawson and N. Maniadakis (1999) Efficiency measurement of health care. *Health Care Management Science 2*: 161-172.

Kaufmann, W. and E. Wolf (1998) Die vergleichende Datenbank. *führen & wirtschaften im Krankenhaus 15*: 208-209.

Klingebiel, N. (1997) Leistungscontrolling im New Public Management. *Betriebswirtschaftliche Forschung und Praxis 49*: 629-652.

Knieps, F. and H. Reichelt (1999) Globalbudget und Arzt-Netzwerk. *Soziale Sicherheit 48*: 96-105.

Laffont, J.-J. and J. Tirole (1993) A theory of incentives in procurement and regulation. Cambridge, Mass..

Leeflang, P. S. (2000) Building models for marketing decisions. Boston.

Metzger, L. M. (1993) Measuring quality cost effects on productivity using data envelope analysis. *Journal of Applied Business Research 9 (3):* 69-81

Meurer, U. (1995) Nicht zur Veröffentlichung bestimmt .... *führen & wirtschaften im Krankenhaus 12*: 226-227.

Nyhan, R. C. and H. A. Marlowe (1995) Performance Measurement in the Public Sector: Challenges and Opportunities. *Public Productivity & Management Review* 18 (4): 333 - 348.

Nyhan, R.C. and L. L. Martin (1999) Comparative performance measurement. *Public Productivity & Management Review 22 (3)*: 348-364.

Olesen, O. B. and N. C. Petersen (1994) A presentation of GAMS for DEA. *Omega* 22: 323 – 339

Reister, M. (1995) Statistische Krankenhausdaten. In M. Arnold and D. Paffrath (Eds.) Krankenhaus-Report ´95. Stuttgart, Jena: 233 – 255.

Ruggiero, J. (1996) On the measurement of technical efficiency in the public sector. *European Journal of Operations Research 90*: 553 – 565.

Shepard, R. W. (1970) Theory of Cost and Production Functions. Princeton, NJ.

Staat, M. (1999) Treating non-discretionary variables one way or the other: implications for efficiency scores and their interpretation. In G. Westermann (Ed.), Data Envelopment Analysis in the Service Sector. Wiesbaden: 23-50.

Steinmetz, F. (Ed.) (1991) Praxis moderner Krankenhausführung. Hamburg.

Stepan, A. (Ed.) (1997) Finanzierungssysteme im Gesundheitswesen. Wien.

Stirring Committee for the review of commonwealth/state service provision (Ed.) (1997) Data Envelopment Analysis: A technique for measuring the efficiency of government service delivery. Canberra.

Thomae, D. (1995) Marktstrukturen im Krankenhausbereich. In M. Arnold and D. Paffrath (Eds.) Krankenhaus-Report '95. Stuttgart, Jena: 49 - 58

Westermann, G. (Ed.) (1999) Data Envelopment Analysis in the Service Sector. Wiesbaden.

Young, S. T. (1992) Multiple productivity measurement approaches for management. *Health Care Management Review 17 (2)*: 51-58.

Appendix

Table A 1: Hospitals with normal and adverse case mix

surgery department		*per diem* mean (s. d.)	*avg. stay* mean (s. d.)
type I hospitals	normal (103)	392.63 (52.62)	10.84 (2.24)
	adverse (5)	413.97 (16.36)	10.58 (1.90)
type II hospitals	normal (41)	384.69 (44.50)	11.09 (1.47)
	adverse (11)	408.68 (68.01)	11.27 (2.35)
internal medicine department			
type I hospitals	normal (106)	393.77 (52.13)	12.45 (1.90)
	adverse (2)	385.12 (15.05)	9.15 (2.05)
type II hospitals	normal (49)	386.7 (43.62)	12.07 (1.92)
	adverse (3)	439.87 (124.82)	9.03 (1.47)

Table A 2: Capacity use by federal state in 1993[46]

Federal state		Federal state	
Baden-Wuerttemberg	.826	Hessen	.825
Bayern		Mecklenburg-Vorpommern	.789
Berlin	.867	Niedersachsen	.814
Brandenburg	.777	Nordrhein-Westfalen	.819
Bremen	.811	Rheinland-Pfalz	.827
Hamburg		Schleswig-Holstein	.842

46 See Table 17-11 in Reister (1995).

Table A 3: Inefficient and efficient hospitals: Summary statistics

variable	Type I hospitals inefficient	Type I hospitals efficient	Type II hospitals inefficient	Type II hospitals efficient
per diem	396.95	381.27	397.36	380.19
treatment flat rate	4882.35	4195.65	4948.27	4173.91
beds	165	154	192	201
avg. stay internal medicine department	12.60	11.60	11.27	11.41
avg. stay surgery department	10.94	10.39	11.45	10.71
adv. case mix internal medicine department	0	.09	0	.13
adv. case mix surgery department	.01	.17	.24	.17
number of fields of specialisation	3.61	4.09	3.97	4.48
observations	85	23	29	23