Measuring and Improving the Performance of Health Service Networks

Maik Hammerschmidt, Tomas Falk and Matthias Staat

Journal of Service Research 2012 15: 343 originally published online 27 April 2012
DOI: 10.1177/1094670512436804

Journal of Service Research
15(3) 343-357

DOI: 10.1177/1094670512436804
http://jsr.sagepub.com
SAGE

Measuring and Improving the Performance of Health Service Networks

Maik Hammerschmidt[1], Tomas Falk[2], and Matthias Staat[3]

Abstract

Health maintenance organizations (HMOs) have intensified their efforts to establish network-like structures with service partners who are responsible for different functions along the health value chain. To calculate the potential value and cost benefits of service production within health care networks and to improve performance in such networks, the authors propose a two-step benchmarking approach. While the first step is concerned with measuring and comparing service provider performance, the second step relates to a contact program that disseminates the lessons learned during the benchmarking process. Across two empirical studies with general practitioners and specialty physicians, the authors identify in a first step tremendous overspendings and provide suggestions on cost reductions that could be achieved without threatening output levels. With regard to the second step, the authors find that detailing efforts based on the results of performance measurement helped physicians to improve their performance. Through detailing, the hub was able to inform network partners about the benchmarking results and to reveal performance gaps in their current resource utilization patterns. In addition, the authors show that managers of HMOs should seek out physicians with smaller practices and high-referral (i.e., risk-averse) physicians as targets for detailing, who are especially responsive to these initiatives.

Keywords

service networks, benchmarking, service efficiency, performance management, health care networks, data envelopment analysis

Health maintenance organizations (HMOs) increasingly enter into relational strategies with health care providers, such as general practitioners (GPs) and specialty physicians (Stremersch and Van Dyck 2009). By establishing network-like structures, either through ownership or through formal agreements, HMOs can promote the harmonization of health care provider assets and expertise, resulting in "turnkey" solutions (Berry and Mirabito 2010; Wan and Wang 2003), as well as advance standardized service provision that might reduce health care costs (Metters and Marucheck 2007).

From a strategic perspective, HMOs thus resemble a network hub that initiates and coordinates activities in a network (Achrol and Kotler 1999). In turn, physicians represent the organizationally independent yet functionally dependent network spokes that coproduce a final service outcome (e.g., health care). Poor performance by one party thus undermines the performance of successive parties. Exploring the performance of health care networks also seems particularly worthwhile, because health care is increasingly troubled by "costing too much" and "wasting too much" (Berry and Bendapudi 2007, p. 112).

To improve performance in health care networks, two key needs emerge: (1) to identify the best performers in a given service function and (2) to improve the abilities of poor performers. Assessments of physician performance should rely on efficiency measured as the ratio of multiple service outputs to multiple service inputs (Keh, Chu, and Xu 2006; Rust and Huang 2012). Therefore, evaluating and comparing physician efficiency represents the "diagnosis" in performance management, whereas developing strategies to encourage poor performers to improve their efficiency constitutes the "therapy." By considering both steps in service performance management, diagnosis and therapy, this study advances existing literature that merely assesses performance (Donthu and Yoo 1998; Frei and Harker 1999; Hollingsworth 2008). Specifically, to recognize the best performers, we develop a benchmarking procedure that acknowledges physician heterogeneity and ensures that measured efficiency differences actually stem from varying capabilities, rather than just unique characteristics of the physicians (Brown 2006; Dyson et al. 2001). Then, to design effective performance therapy, we draw on the notion of detailing and consider the impact of communication efforts that disseminate lessons learned from best performers to the

[1] Chair of Marketing and Innovation Management, University of Goettingen, Goettingen, Germany
[2] ConCardis Endowed Chair for Consumer Behavior, EBS Business School, Oestrich-Winkel, Germany
[3] DSC Consulting, Schriesheim, Germany

Corresponding Author:
Maik Hammerschmidt, University of Goettingen, Platz der Goettinger Sieben 3, 37073 Goettingen, Germany
Email: maik.hammerschmidt@wiwi.uni-goettingen.de

members of the wider health care network. Generally, detailing refers to life sciences firms' endeavors to influence physicians' prescription and therapeutic behaviors. Considerable research notes the effectiveness of detailing (Chintagunta and Desiraju 2005; Manchanda, Rossi, and Chintagunta 2004; Narayanan, Manchanda, and Chintagunta 2005), though tests of physician responsiveness to detailing are rare (Nair, Manchanda, and Bhatia 2010; Stremersch and Van Dyck 2009). We posit that the effectiveness of detailing for improving performance strongly depends on *practice size*, reflecting the amount of resources available for implementing changes, and *referral intensity* reflecting perceived risk.

With this approach, our research not only establishes an actionable, two-step performance management approach in hub-and-spoke networks but also reveals levers that influence the effectiveness of detailing. In the next section, we present our conceptual framework, which reflects existing benchmarking research. Next, using longitudinal data from 816 GPs and 633 specialty physicians, we test the benchmarking procedure as a means to measure the efficiency of health care providers embedded in a network structure. We also explore the effect of detailing calls on efficiency, using hypotheses that we derived from prior detailing literature. We test our hypotheses based on data from a field experiment with 726 GPs who either did or did not receive a detailing call. Finally, we summarize our findings and their implications.

Research Framework

Benchmarking, a generally effective means to achieve performance improvements (Donthu, Hershberger, and Osmonbekov 2005; Vorhies and Morgan 2005), refers to "a continuous, systematic process for evaluating the products, services, and work processes of organizations that are recognized as representing best practices for the purpose of organizational improvement" (Spendolini 1992, p. 9). In particular, research shows that benchmarking can improve the performance of bank branches (Kamakura et al. 2002), physicians (Chilingerian and Sherman 1997), retail stores (Donthu and Yoo 1998), and sales forces (Horsky and Nelson 1996). However, prior studies focus on single, hierarchical organizations that conduct intraorganizational benchmarking analyses. In contrast, health care intermediaries enter into relational strategies with health care providers, creating nonhierarchical network structures. Thus, it is unclear whether existing findings transfer to network settings, which "are not tolerant to traditional instruments of authority and control" (Achrol and Kotler 1999, p. 146).

However, imitative learning and replication of service processes through benchmarking could be effective in network-like structures. Network partners cooperate to achieve competitive advantages, and this mutual goal encourages productive information flows and exchanges of ideas, which also can support benchmarking activities (Ostrom et al. 2010; Vorhies and Morgan 2005). Although prior research offers no suggestions about which part of a network should initiate, coordinate, or control benchmarking activities, we assert that in service networks, one focal organization generally performs fewer service functions and instead serves as an integrator that manages and coordinates the network—that is, the hub (Evanschitzky 2007; Singh 1991). We propose that this hub is best able to plan and monitor benchmarking activities, because it possesses the required process knowledge and can link decoupled or loosely coupled network members. Network members are highly specialized and interdependent, so without any hierarchical authority, the network hub must organize information and resource flows, as well as coordinate decisions and activities (Achrol and Kotler 1999). Thus, benchmarking activities should be initiated and monitored by the network hub, with a focus on the performance of boundary-spanning network partners, which interact with end customers and generate most of the service outcome (Singh 1991).

We empirically test these propositions by analyzing the performance of an integrated health care network (IHN). In general, IHNs combine multiple organizations in partnerships to establish comprehensive health care delivery systems (Wan and Wang 2003). Despite traditional resistance to performance management, pressure to evaluate the performance of health service providers has grown as health care costs rise. The pressure is particularly intense for IHNs (Hollfelder 2002; Meyer Goldstein and Ward 2004), which must fulfill the needs of patients by managing a network of providers (GPs and specialty physicians) who constitute a health care delivery system.

Diagnosing Service Performance: Identifying Best Performers

Background

For this study, we cooperated with a health care fund in a European country that is comparable to IHNs and has evolved from a loosely coupled service system with multiple physicians into a comprehensive system (Govind, Chatterjee, and Mittal 2008). Such organizations, through formal agreements, align health care providers to deliver integrated health care services that improve patient coverage and reduce costs (Wan and Wang 2003). The focal health care fund maintains contracts with a full spectrum of physicians who deliver all elements of health care to a clearly demarcated geographic area. Patients are assigned to a single GP; to be reimbursed for physicians' services, they must visit specialty providers who are affiliated with the same fund. These contractual agreements also limit the range of services each physician may provide, so any patients in need of extended services must be referred to specialists. Physicians' income is composed of a per capita, a fixed monthly premium for each patient, plus additional fees for any services not covered by this per capita amount, paid each quarter. The fund thus functions as a network hub, in that it manages a network of designated subcontractors, has necessary process knowledge, can control information flows, and initiates important network activities.

To optimize overall performance, the fund should benchmark the performance of its network partners. In line with existing research, we propose efficiency as a meaningful

Table 1. Descriptive Statistics for General Practitioners[a]

	District			
	1	2	3	4
Number of physicians	209	201	210	196
Fee-for-service (€)[b]	135,068	129,736	128,448	116,704
Cost of medication (€)[b]	266,280	252,744	256,104	254,904
Cost of referrals (€)[b]	147,100	140,068	146,916	133,852
Number of patients till 50 years[c]	2,176	2,012	2,012	1,700
Number of patients older than 50 years[c]	1,592	1,564	1,532	1,608
Morbidity rate (%)[d]	2.59	2.57	2.24	2.21
Mortality rate (%)	.84	.72	.78	.91
Unemployment rate (%)	4.47	7.25	8.01	8.97
	Total cost per physician (€)			
Minimum	260,967	292,741	234,264	270,670
Average	548,448	522,548	531,468	505,460
Maximum	1,876,813	753,148	1,178,169	1,166,497

[a] Descriptive statistics for physician age and gender are not be reported, for confidentiality reasons.
[b] The reported values are average costs per physician per year (2006).
[c] The reported values are average number of patients per physician.
[d] Morbidity rate = proportion of the population with diseases on a "White list" based on the international classification of diseases.

performance metric that encompasses both input and output sides (Frei and Harker 1999; Keh, Chu, and Xu 2006). Physicians are largely in control of most inputs and outputs in medical care production and are responsible for bottlenecks in service provision (Meyer Goldstein and Ward 2004). Yet, few studies consider physician efficiency. Chilingerian and Sherman (1997) examine the relative efficiency of primary care physicians in their study, which regards health care provision as a process of transforming inputs (e.g., quantity of clinical resources) into outputs (e.g., number of patients treated). However, they also mix different specialties, without accounting for potential heterogeneity in case mixes, type of services provided, or service quality. Such an approach fails to acknowledge that differences in service efficiency inherently result from the various types of physicians considered (Dyson et al. 2001). Even studies that feature more precise information about case mixes and the nature of health care services provided do not explicitly account for sample heterogeneity. Wagner, Shimshak, and Novak (2003), for example, propose a case severity index and find that by adopting best practice patterns, physicians could reduce health costs by 10% but maintain constant service outcomes. Ozcan (1998) and Pai, Ozcan, and Jiang (2000) assess efficiency for a specific diagnosis while using a case severity index, but they also include physicians from different regions and ignore environmental factors that might explain the substantial efficiency variation across regions. Thus, prior studies cannot provide robust evidence about the efficiency of health care provision.

By analyzing physicians with different specialties and failing to conduct tests of the viability of pooling these potentially heterogeneous physicians for a benchmark analysis, existing studies could produce biased results, in that they might choose the wrong role models for benchmarking. Accordingly, we explicitly incorporate system heterogeneity into our analysis to ensure the validity of our efficiency measure.

Data Description

For our analysis, we used the physician database of our fund partner, which granted us access to longitudinal input and output data for 816 GPs and 633 specialty physicians. All physicians had identical contracts with the health fund, which strictly limit the type of services they may offer. Therefore, the spectrum of services provided is homogeneous within the sample.

As inputs we used the fee-for-service (FfS) charges, the cost of medication (CoM), and the cost of referrals (CoR) to (other) specialists. The physicians in this fund do not own any significant equipment, which would permit them to charge higher fees for service relative to the time spent with each patient. Therefore, the FfS measure provides a good approximation of the time spent treating patients. These three inputs also represent the three cost categories most commonly reported in health cost statistics (Ginsburg et al. 2006). For the output, we consider the numbers of patients in two age categories (0–50 years and 50+ years). Splitting patients into age categories appropriately captures differences in the intensity and quality of the health services provided across physicians (Chilingerian and Sherman 1997; Kravitz et al. 1992).

In addition, for all physicians, we collected data on regional morbidity and mortality (proxies for case severity mix), regional unemployment rates, and physicians' age and gender. These data enable us to control for the heterogeneity of the physicians. The area covered by the fund comprises four districts, which differ with respect to the three exogenous regional characteristics mentioned above (e.g., morbidity). In Table 1, we provide the descriptive statistics for the focal GPs.

Table 2 contains the descriptive statistics for the specialty physicians and shows that the total cost per patient, as well as the proportions of the three cost factors, vary considerably across specialty groups. We use these data to detect the

Table 2. Descriptive Statistics for Specialty Physicians

Specialty	Number of Physicians	Average (per Physician)						
		Fee-for-Service (€)	Cost of Medication (€)	Cost of Referrals (€)	Number of Patients ≤ 50 Years	Number of Patients > 50 Years	Total Cost per Physician (€)	Total Cost per Patient (€)
Dermatology	73	150,870	83,817	10,440	2,514	1,680	245,127	58.45
ENT	61	183,884	21,479	22,814	1,936	1,448	228,177	67.43
Gynecology	110	129,167	18,931	54,029	2,340	728	202,127	65.88
Internal	96	173,469	92,644	68,567	572	1,690	334,680	147.96
Neurology	29	187,897	128,871	49,242	908	1,236	366,010	170.71
Ophthalmology	87	202,636	30,416	3,297	2,031	2,340	236,349	54.07
Orthopedics	26	244,970	33,874	135,959	1,608	1,864	414,803	119.47
Pneumology	40	206,376	156,892	21,685	1,152	1,648	384,953	137.48
Psychiatry	27	157,884	158,238	22,486	832	873	338,608	198.60
Surgery	38	129,132	17,507	26,154	684	827	177,459	114.36
Urology	46	201,303	94,916	40,307	773	2,303	336,526	109.40

Note. ENT = ear, nose, throat.
For ease of exposition, we focus on overall values across all districts.

inefficiencies and potential cost savings for each group, adjusted for different sample sizes (i.e., ranging from 26 to 110 observations). Because we have a sufficient number of observations for each specialty group, we do not need to mix different specialties within one model and can obtain clean efficiency results for each specialty group.

To ascertain whether our three inputs are the key factors for producing the selected output, we must determine whether there are any hidden nuances in these effects. A certain portion of output variance could be due to unobserved or omitted physician-level inputs, such as differences in the patient base associated with a certain physician (e.g., nature of disease, characteristics of the patients, compliance) or differences in expertise or willingness to adhere to clinical guidelines. Moreover, omitted time-level inputs might involve changes in physician learning and trends over time. To evaluate the relevance of these omitted inputs, we ran a linear mixed model that accounts for 2-way (physician and year) random effects (Griliches and Hausman 1986; Luo 2007); the equation for the mixed model appears in the Appendix A. For the GPs, physician random effects had a relatively small intraclass correlation coefficient of 9.2%, such that unobserved or omitted physician inputs accounted for less than one tenth of the variance in the output variable. Unobserved time-random effects were weakly significant and accounted for only 9.4% of the output variance. For specialty physicians, physician random effects (time-random effects) accounted for 18.6% (10.7%) of the output variance. The influence of omitted input variables thus was small to moderate (Baltagi 2001).

Methodology

Data envelopment analysis (DEA): Assessing physician efficiency. Introduced by Charnes, Cooper, and Rhodes (1978), DEA measures the efficiency of a decision-making unit (in our case, physicians) relative to a frontier of the most efficient units. As the true frontier is unknown, the frontier is estimated nonparametrically from a set of observed units. Thus, the true efficiency of physician k, θ_k, is not directly observable, but for any given sample of physicians $\mathcal{S} = \{(x_i, y_i) | i = 1, \ldots, n\}$, sample equivalents can be derived. In turn, $\hat{\theta}_k$ is the estimate of θ_k obtained by solving the following fractional programming format:[1]

$$\max \hat{\theta}_k = \frac{\sum_{r=1}^{s} u_{rk} y_{rk}}{\sum_{j=1}^{m} v_{jk} x_{jk}} \text{ s.t.} = \frac{\sum_{r=1}^{s} u_{rk} y_{ri}}{\sum_{j=1}^{m} v_{jk} x_{ji}} \leq 1; \quad i = 1, \ldots k \ldots, n;$$
$$u_r \geq 0; v_j \geq 0; \quad r = 1, \ldots, s; \quad j = 1, \ldots, m. \tag{1}$$

In DEA, we form a virtual input and output for each physician, using the weights u_r for outputs y_r and v_j for inputs x_j. Each physician thus has two outputs, number of patients till 50 years (NP_{0-50}) and number of patients older than 50 years (NP_{50+}). The three inputs are FfS, CoM, and CoR. Then, for any physician k, the virtual output is expressed as $u_1 \cdot NP_{0-50} + u_2 \cdot NP_{50+}$, and the virtual input is $v_1 \cdot FfS + v_2 \cdot CoM + v_3 \cdot CoR$. All estimated efficiency scores ($\hat{\theta}_k$) are equal to or less than 1 (100%), and the model runs successively with each physician in the objective function to derive individual efficiency scores.[2] Efficient physicians (i.e., best practices according to DEA) earn a score of 1 and form the efficient frontier. The remaining physicians earn scores between 0 and 1, and the portion ($1 - \hat{\theta}_k$) represents the inefficient percentage of inputs for physician k, that is, resources that could be saved, holding the output level constant (Luo and Donthu 2006).

Bootstrap procedure: Testing for sample heterogeneity. We apply a bootstrap procedure to test for heterogeneity across subsamples. This technique obtains an efficiency distribution by repeatedly drawing new data through resampling. The

original efficiency estimator gets applied to each new sample, such that the resulting bootstrap estimates (pseudo-scores) mimic the sampling distribution of the original estimator (Simar and Wilson 1998). The next step involves an asymptotic test of whether certain subsamples of a set of observations have different efficiency distributions (i.e., test of sample heterogeneity), which is of vital importance for diagnosing physicians' efficiency and potential efficiency gaps meaningfully (Dyson et al. 2001). Not accounting for heterogeneity would distort the true causes of efficiency differences; it would be unclear whether efficiency variations resulted from better input-output translations or from varying characteristics of the benchmarked physicians. We provide a detailed overview of the bootstrap procedure in the Appendix A.

Monte Carlo simulation: Calculating sample size-corrected efficiency scores. For physicians of different specialties that are incommensurate with respect to the services they provide, we likely need to run separate DEA models. Considering separate groups of specialty physicians results in samples of different sizes, and the quality of the DEA results depends on the number of observations in the sample (Kneip, Park, and Simar 1998), such that more observations provide a better approximation of the true frontier. Therefore, results obtained from data sets of different sizes are difficult to compare. To ensure comparability across efficiency scores, despite varying sample sizes, we use a Monte Carlo approach and limit the larger samples to the size of the smallest sample (Zhang and Bartels 1998).

Superefficiency analysis: Eliminating low-quality outliers. Using the numbers of patients as outputs in a frontier-based evaluation might initiate a race to the lowest service quality level (Newhouse 1994); that is, extreme efficiency can be obtained by reducing service quality, which is not explicitly captured as an output measure. Physicians who follow such a shirking strategy likely reduce inputs (e.g., treatment costs) though, so they should emerge as low-cost outliers (Hollingsworth 2008). Low-cost physicians who do not provide adequate treatment should be removed from the analysis; otherwise, they create cost targets for other state-of-the-art physicians who cannot maintain high-quality services with these extremely low costs.

To identify such low-quality outliers, we use the superefficiency procedure (Banker and Chang 2006; Simar 2003). Low-quality outliers that produce outputs with an extremely low input (cost) level receive extremely high efficiency scores and push out the frontier, leading to biased evaluations. Technically, when estimating superefficiency scores, the physician k being evaluated does not appear in the reference set, so an additional constraint ($i \neq k$) must be added to the linear program of the standard DEA model from Equation 1. The efficient frontier (reference set) for physician k then consists of observations other than k. Among physicians who are efficient in the standard DEA model, some may lie above this new frontier, such that extremely efficient physicians achieve scores significantly greater than 1. Prior literature shows that units with superefficiency scores greater than 1.5 are likely outliers and thus should be excluded (Banker and Chang 2006).

The results of the superefficiency analysis show that all GPs remain far below this cutoff level. Thus, we can ignore the risk of pursuit of the lowest service quality. Using physical performance outputs is appropriate, and we do not need sample adjustments (Hollingsworth 2008; Newhouse 1994). In contrast, for specialty physicians, we find a 12% average proportion of outliers. We remove these low-quality outliers from the data to avoid unrealistic cost targets for providers offering state-of-the-art services (Newhouse 1994; Zuckerman, Hadley, and Iezzoni 1994). The subsequent analyses and results for specialty physicians are based on this reduced, quality-adjusted sample.

Results

Efficiency of GPs. In line with existing empirical findings, we focus on five exogenous variables that constitute heterogeneity in the input-output transformation process of physicians (Chilingerian and Sherman 1997; Janakiraman et al. 2008): regional morbidity, mortality, and unemployment rate where the practice is located, as well as physician age and gender. We test for the possible effects of these factors on physician efficiency by estimating the DEA model in steps. Starting with morbidity, we calculate four-group (districts), two-group (high vs. low morbidity), and one-group (full sample) solutions and compare the results using the bootstrap procedure. We find no significant difference for either the four-group (Table 3, Panel A) or the two-group (Table 3, Panel B) solutions and thus cannot reject the null hypothesis that the true average efficiencies are identical for all subgroups. In other words, the stepwise DEA does not point to a heterogeneity problem. Similarly, we ran bootstrap tests for mortality, unemployment rate, age (with median splits), and gender. Again, all test statistics were closer to 1 than to the respective critical values, indicating no significant heterogeneity in the sample. These results provide evidence of a negligible influence of the exogenous variables on the input-output transformation and indicate that a one-group model (i.e., one DEA model for the full sample) adequately characterizes the data. The average efficiency for this solution is .82, and 726 GPs (89%) are inefficient.

Efficiency of specialty physicians. In contrast with GPs, specialty physicians likely should not be jointly examined in one model. Grouping these physicians by specialty seems conceptually necessary, considering the varying diseases treated by the different specialties (Ozcan 1998). Yet, two problems arise with this grouping: First, specialty groups involve considerable differences in sample sizes. Second, they differ remarkably with respect to treatment costs per patient.

In Table 4, the first column of Panel B lists the efficiency scores obtained by a standard DEA with the original sample sizes. The first column of Panel C shows the results of the Monte Carlo simulation with the efficiency scores adjusted to

Table 3. Assessing Heterogeneity of General Practitioners

A: Four-Group Solution (4 Districts)					B: Two-Group Solution (High vs. Low Morbidity)				
Group (District)	Average	Average (Other)[a]	τ[b]	95% Value[c]	Group	Average	Average (Other)[a]	τ[b]	95% Value[c]
1	.8232	.8157	1.009	1.0288	1	.8206	.8146	1.007	1.1028
2	.8180	.8175	1.001	.9815					
3	.8167	.8179	.9985	.9847	2	.8146	.8206	.9927	.8972
4	.8125	.8193	.9917	.9873					

[a] Average (other) = average efficiency of the respective other groups (e.g., of the groups 2 through 4 in case of group 1).
[b] τ = test statistic, or the column "average" divided by column "average (other)".
[c] Critical value, equal to the 95% value of the distribution of ratios obtained via bootstrapping, which implies a generous significance level of 10%. Average efficiency for the one-group solution (i.e., full sample): .8168. In the one-group solution, 726 GPs (89%) are inefficient.

Table 4. Detailed Results for Specialty Groups (Average Across Inefficient Physicians)

			B: Results Without Bias Correction (Standard DEA)		C: Results With Bias Correction (Monte Carlo Simulations)	
Specialty	Number of Physicians[a]	A: Total Cost per Physician (€)[b]	Efficiency	Cost-Adjusted Potential per Physician (€)	Efficiency	Cost-Adjusted Potential per Physician (€)
Orthopedics	49	436,905	.7323	116,959	.8376	70,953
Urology	20	331,948	.8225	58,920	.8789	40,199
Surgery	15	175,951	.7799	38,726	.8011	34,997
Neurology	11	371,380	.8898	40,926	.9094	33,647
Internal	58	382,796	.8284	65,687	.9192	30,930
Ophthalmology	63	240,592	.7942	49,513	.8792	29,064
Pneumology	22	411,165	.8927	44,118	.9324	27,795
ENT	35	211,113	.7271	57,613	.8763	26,115
Dermatology	46	260,067	.8503	38,932	.9045	24,836
Psychiatry	10	341,712	.9383	21,084	.9383	21,084
Gynecology	87	206,837	.8689	27,116	.9512	10,094

Note. DEA = data envelopment analysis.
[a] These values are numbers of *inefficient* physicians (score < 1) in each specialty, who exhibit cost savings potential.
[b] These values are average total treatment costs across *inefficient* physicians. All results are based on reduced samples without low-quality outliers.

the size of the smallest specialty (i.e., psychiatry). The comparison indicates that the sample size bias is considerable. With the sample size-adjusted efficiency scores, we applied our bootstrap test for heterogeneity; the average efficiency scores across all specialties differed significantly from one another, which suggests substantial heterogeneity with respect to the nature of disease (i.e., among specialties). However, we found no within-specialty heterogeneity with respect to mortality, morbidity, unemployment rate, gender or age of physicians. These findings support our specification of separate DEA models on the specialty group level but pooling all physicians of the same specialty in one DEA model.

Differences in treatment cost per physician across specialties also may distort the results of the efficiency analyses. Therefore, to identify the maximum savings potential we took sample size-related bias corrections and a trade-off between efficiency and treatment cost into account simultaneously. Panel A in Table 4 shows the total treatment cost per inefficient physician (average) for each specialty. The second column in Panel B then provides the cost-adjusted savings potential, based on the standard efficiency score (i.e., without bias correction). Finally, the second column in Panel C shows the bias-corrected and cost-adjusted savings potential per physician (total cost × bias-corrected efficiency score). Note that these values refer to inefficient physicians; for efficient physicians operating on the frontier, no projected savings exist. Our results show that the potential improvements are vastly overestimated when the cost-adjusted savings potential is calculated using standard DEA scores. For example, according to the cost-adjusted savings potential from the standard efficiency score, ear, nose, and throat (ENT) physicians take position 4. In the cost-adjusted and bias-corrected metric, ENT ranks in position 8 though. The projected savings for ENT actually are rather low (€26,115 per physician), in that this low-cost specialty has a relatively high-efficiency score when we account for sample size bias. In contrast, achieving best practices in orthopedics could save €70,953 per physician, even with constant outcomes.

The bias-corrected and cost-adjusted metric in the final column of Table 4 also provides information about trade-offs between the specialty groups, such as how many improvement initiatives for gynecologists could be substituted for by improving the efficiency of one orthopedist. Assume that the cost of contacting physicians is constant across specialties,

enhancing the efficiency of an additional orthopedist instead of a gynecologist results in increased cost savings of €60,859. Moreover, it is possible to reduce efforts and program costs, holding the level of expected gains constant. For example, the potential remains unchanged if targeting of seven gynecologists gets replaced by targeting one orthopedist.

Our diagnosis part of performance management in a health care network reveals the current performance of physicians; the resulting efficiency scores represent performance metrics and disclose efficiency gaps on the basis of a relative-to-best comparison. This diagnosis constitutes a necessary precondition for improving physician performance and overall network performance. Next, we aim to provide conceptual and empirical insights into how to provide therapy to poor performers, drawing on insights from detailing literature.

Improving Service Performance: Therapy for Poor Performers

Physician Responsiveness to Detailing

To improve physician performance in line with the identified best practice patterns, members of the network should emulate the most efficient entities, as identified during the diagnosis phase (Vorhies and Morgan 2005). Although prior literature implies that implementing benchmarking goals is easy, ''at least conceptually'' (Donthu et al. 2005, p. 1475), little practical advice indicates how to realize this task. We therefore propose the communication of the benchmarking findings through detailing calls. Detailing traditionally refers to visits and/or calls by pharmaceutical or other salespeople who attempt to influence physicians' prescription and therapeutic behaviors (Chintagunta and Desiraju 2005; Manchanda, Rossi, and Chintagunta 2004; Narayanan, Manchanda, and Chintagunta 2005). Such efforts through personal communication can convince physicians to switch to new drugs and new ways to monitor patients (Venkataraman and Stremersch 2007). Although the mean effect of detailing on physician behavior is positive, it also may be relatively small (Stremersch and Van Dyck 2009); some physicians even resist detailing and rely on inertia or habit to determine their treatment decisions, especially when they perceive poor source and message credibility (Mizik and Jacobson 2004). Less credible arguments in a detailing campaign thus get discounted, particularly if physicians believe sales representatives are neither experts nor reliable information providers (Janakiraman et al. 2008).

In response, we propose that detailing instead might be effective in network-like structures, because the information provider and detailing recipient already have entered into a strategic partnership. The network hub aims to increase patient value rather than maximize its own profits, which should enhance its source credibility (Wan and Wang 2003). That is, physicians might regard detailing activities triggered by the network hub as attempts to increase the efficiency of health care provision, not just as a communication campaign to ''sell'' drugs and therapies.

Detailing programs that are based on relative-to-best comparisons also could establish a higher degree of message credibility, because best performers represent expert peers and potentially opinion leaders among physicians who operate less efficiently in similar conditions. Recent findings empirically confirm that physician behavior is significantly influenced by expert peers (Nair, Manchanda, and Bhatia 2010), because the opinions and practices of (efficient) peers help reduce uncertainty among less efficient physicians. Physicians also should be more motivated to respond to guidelines and change their practices if the information has been derived from a precise and sound methodological procedure, such as DEA (Agrell and Bogetoft 2001; Guth and Kleiner 2005).

Finally, obtaining information through detailing requires minimal effort and can be a valuable source of information for a busy doctor that reduces search, learning, and thinking costs (Janakiraman et al. 2008). These benefits then should enhance physicians' intentions to participate in a benchmarking program, even without direct financial rewards. Because the initiating, coordinating network hub offers high source and message credibility, physician responsiveness to detailing campaigns in health care networks should be positive.

Hypothesis 1. Physicians who receive a detailing call improve their performance more than physicians who do not receive a detailing call.

Moderating Effect of Practice Size

Performance gains triggered by detailing calls likely depend on physician characteristics, including the number of patients served (i.e., practice size). We posit that low-volume physicians exhibit less inertia, such that they can be influenced more easily to improve their efficiency. Doctors with fewer patients have more time and motivation to learn, search, and think (Narayanan, Manchanda, and Chintagunta 2005); but if they work in smaller practices, they likely lack access to various information sources and thus rely more on information gleaned from network representatives. That is, for low-volume physicians, detailing calls should have a greater impact in terms of unlocking their efficiency potential.

Hypothesis 2. Among physicians who receive a detailing call, performance improvements decline with increasing practice size.

Moderating Effect of Extent of Referrals

Many physicians regard referrals as an effective means to reduce perceived treatment risk, because they shift the risk of failure (e.g., wrong diagnosis, inappropriate therapy) and responsibility to colleagues. Physicians who feel uncomfortable with their skills pertaining to a specific treatment use the referral slip to avoid potentially risky decisions (Chilingerian and Sherman 1997), and empirical research

Table 5. Comparison of Test and Control Groups

	Test Group: Detailing Call	Control Group: No Detailing Call
Efficiency score in 2006[a]	76.4	76.7
Size of efficiency gap (100—efficiency score)[a]	23.6	23.3
Change in efficiency 2006-2007[a]	2.71*	−.58
Proportion of efficiency gap closed[a] (%)	11.4	−2.5

Note. GP = general practitioners.
*Significant at $p < .01$.

[a] The reported values are group means. The GPs in the test and control group are equally distributed across the four districts. The physicians in the test and control group did not significantly differ with respect to age and gender, as well as morbidity, mortality, and unemployment rate in the respective region. We are unable to report group means for these variables for confidentiality reasons.

identifies referrals as a risk-averse behavior by physicians who engage in defensive decision making (Ruston 2004).

Generally, defensive decision making is triggered by imperfect information. To overcome their information gaps, people perform screening activities, traditionally through an active search for information (Stiglitz 2000). However, many physicians struggle with time pressures and scarce resources (Janakiraman et al. 2008), so they may screen negatively, by referring treatment decisions to others. Such negative screening allows the physicians to circumvent difficult decisions and lower treatment risk. Other physicians might use referrals to confirm their own work, which implies the redundant provision of overlapping services and reduces the efficiency of physicians' performance within the network.

In this setting, detailing calls could provide an alternative risk-reduction mechanism that does not harm physicians' performance. By providing concrete information about successful treatment patterns, the network hub actively supports physicians' information search efforts (Spence 1973). Once physicians obtain detailed feedback and suggestions through detailing, their information status improves, which lowers their perceived treatment risk (Nair, Manchanda, and Bhatia 2010). In turn, they do not need to activate alternative mechanisms such as negative screening through referrals. For high-referral physicians, detailing calls should lead to significant performance improvements through a greater willingness to refrain from unnecessary referrals.

Hypothesis 3. Among physicians who receive a detailing call, performance improvements increase with increasing extent of referrals.

Field Experiment

Background. In cooperation with the health care fund we described previously, we conducted a field experiment to test these hypotheses. The fund contacted a randomized subsample of 40% of the 726 GPs who were inefficient in 2006 (i.e., who had efficiency scores below 1) in a detailing call in the first quarter of 2007 (see Table 5). The remaining 60% represented the control group that received no detailing call. For both groups, we obtained input and output data for 2007.

Treatment. The treatment (i.e., detailing call) aimed to articulate clearly the basic clinical policies and standards for the use of resources.[3] More specifically, the detailing informed poor performers about their current resource consumption and the optimal resource utilization pattern applied by similar, comparable best practice physicians. These details included the percentage of fees, medication costs, and referral costs they "overspent" in comparison with best practices. To provide instructive insights, each detailed physician read about three typical inefficient practice styles, identified from the study diagnosis, and was assigned to one of them. The first inefficient practice style features considerable overspending on fees, whereas physicians that adopt the second style exhibit particularly high inefficiencies with respect to the use of medications, and the third style is represented by excessive resource consumption for referrals. Each inefficient practice style thus is characterized by one critical input that provides the main source of inefficiency and offers a primary lever for improvement. After having informed physicians about the extent and sources of inefficiency and potential levers for each inefficient style, the person conducting the detailing call disseminated individual targets set by the benchmark physician (i.e., preferred utilization profile for inputs and outputs), explaining that these best performing physicians used their resources optimally to maximize the number of patients served. In addition, the network hub provided suggestions and guidelines on how to reduce overspending to reach these targets. Finally, the physicians were informed that treatment costs of up to €78.4 million could be saved if all inefficient physicians would adhere to the best practices.

Results. Using 2007 data, we calculated DEA efficiency scores. To test Hypothesis 1, we compared the mean efficiency changes from 2006 to 2007 for the treatment group that received a detailing call versus the control group that received no call (see Table 5).[4] We are not interested in an analysis of the efficiency change of all GPs but rather aim to investigate how the performance of the inefficient GPs, relative to other GPs, changes from 2006 to 2007 as they move toward the best practice frontier. A change in relative efficiency also could be caused by a shifting efficiency frontier in 2007, due to factors unrelated to physicians' therapeutic or managerial skills, such as industry changes (e.g., political changes, legislation), technical progress, or environmental variables (Luo and Donthu 2006). Therefore, we apply the efficient frontier for 2006 as a reference function to calculate the efficiency scores for 2007. In other words, we examine the extent to which each individual inefficient physician caught up in 2007 compared with 2006 (Färe, Grosskopf, and Lovell 1994).

We first evaluated whether the test and control groups differed in efficiency during 2006 using a *t*-test (Wilson 2003). It revealed no significant efficiency differences, so we can rule out potential sources of bias between the test and control groups

Table 6. Results of the Moderated Regression Analyses

	β	*t*	β	*t*
Treatment (detailing call)	1.825	4.87	1.788	4.81
Practice size	−.152	3.15		
Treatment × Practice size	−.177	3.05		
Risk aversion			.00982	3.73
Treatment × Risk aversion			.00858	3.35

Note. All coefficients are significant at $p < .01$.
Dependent variable = efficiency change; mean-centered values for practice size and risk aversion enter the regressions. For the treatment variable, we used weighted effect coding. Coefficients for moderators and interaction terms are multiplied by 100.

before the detailing calls. We next tested the difference in means between the 2006 and 2007 efficiency scores using *t*-test with Bonferroni adjustment. Efficiency slightly decreased across periods for the control group (−.58, ns); whereas in the test group, we observed a significant efficiency enhancement (2.71, $p < .01$). The significant difference in the efficiency changes between the test and control groups (3.29, $p < .001$) offers a measure of the performance gain that likely is attributable to detailing calls. In particular, the test group exploited 11.4% of its efficiency enhancement potential, whereas the efficiency gap of the control group increased, by 2.5%, in support of Hypothesis 1.

However, a common problem associated with field experiments is the potential lack of internal validity. The efficiency improvements might be partially influenced by confounding factors, such as physician motivation to comply or patient mix. To minimize the distorting effects of confounding factors, we paid special attention to the design of the field experiment (Shadish, Cook, and Campbell 2002). For example, our close cooperation with the IHN should ensure a correct implementation of the intervention (detailing call). Finally, experimenter effects appear unlikely to represent a severe problem, because the success of the detailing calls did not affect the employees performing the calls, such as in terms of salary.

To test for the hypothesized moderating effects of practice size and risk aversion, we conducted moderated regression analyses. We used efficiency change as the dependent variable and treatment (with weighted effect coding due to different group sizes, where no detailing call = −.66 and detailing call = 1) and the respective moderators (practice size or risk aversion) as independent variables. We operationalized practice size by the number of patients treated in 2006 and risk aversion as the number of referrals in 2006. Both moderators are continuous variables, so moderated regression is the most appropriate method for testing the interactions: It maintains the integrity of the sample, makes complete use of the data, and retains full statistical power (Aiken and West 1991; Irwin and McClelland 2003). Following Aiken and West's (1991) procedure, we first calculated a regression that contained only the treatment variable and the moderator (i.e., no interaction term). The coefficients for the treatment variable and both moderators were significant. In the second regression, we integrated the interaction term between the treatment variable and its respective moderator. The results for the full regression models appear in Table 6.

According to Table 6, the efficiency-enhancing effect of the detailing call is highly significant, as are the regression coefficients for the moderators and the interaction terms. We thus can conclude that practice size and risk aversion predict efficiency changes and moderate the treatment-efficiency change relationship (i.e., quasi-moderators). First, we find a negative moderating effect of practice size on the relationship between the detailing calls and efficiency change. That is, with increasing practice size, the efficiency increase weakens. In other words, for low-volume physicians, detailing calls have a much higher impact in terms of unlocking efficiency potential, in support of Hypothesis 2.

Second, in support of Hypothesis 3, higher efficiency increases followed from detailing calls to highly risk-averse (i.e., high-referral) physicians, compared with those to less risk-averse (i.e., low-referral) physicians. As we show in Table 6, the moderating effect of risk aversion is positive and significant, so the efficiency-enhancing effect of detailing calls increases as risk aversion level increases.

To confirm that the identified moderating effects are not spurious or artificial, we regressed efficiency changes on the treatment variable at high and low moderator levels (Maxwell and Delaney 1993). In addition to the regressions at the mean level of the moderator variables (Table 6), we ran regressions at two other levels: one standard deviation (*SD*) above the mean and one *SD* below the mean. We plot the results of this series of regressions in Figure 1. For each moderator variable, the slopes between efficiency change and the treatment variable (i.e., difference in efficiency change between treatment conditions) differ significantly at the three levels, which indicate that the detected moderating effects are robust.

According to Figure 1, Panel A, the efficiency increase is significantly higher for low-volume physicians than for high-volume physicians. Figure 1, Panel B, also indicates that detailing calls trigger higher efficiency increases for highly risk-averse (i.e., high-referral) physicians compared with less risk-averse (i.e., low-referral) physicians.

Discussion

Implications for Service Research

We add to existing knowledge by promoting a formal approach for increasing the performance of health care networks. In

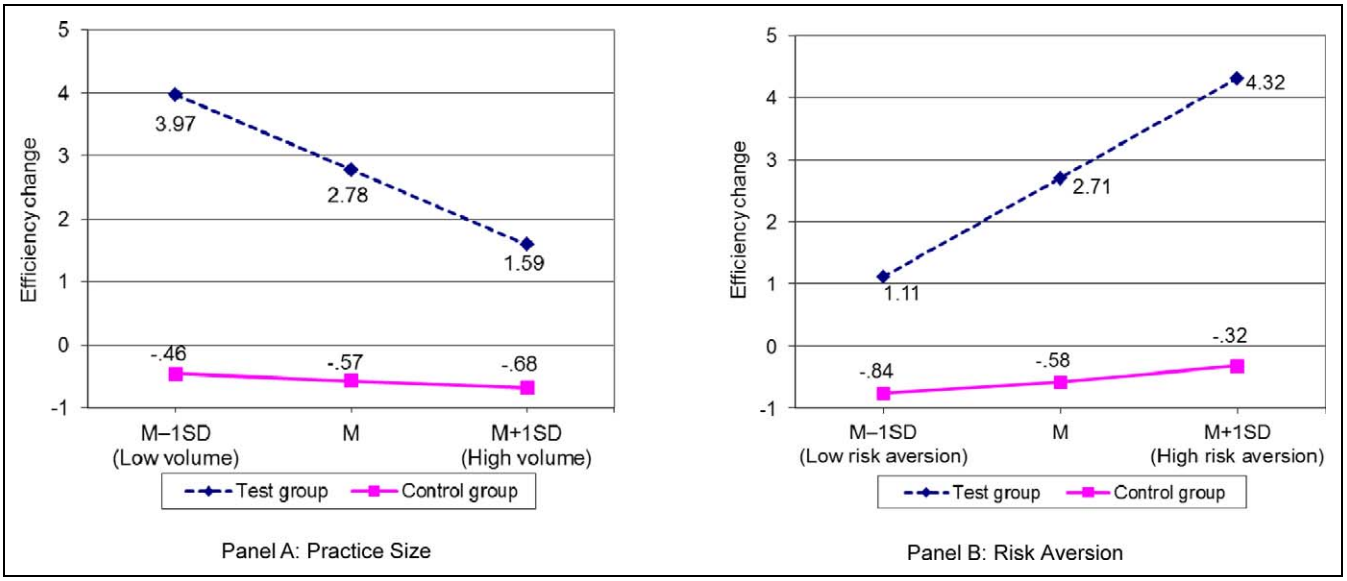

Figure 1. Moderating effects of practice size and risk aversion on efficiency change
Note. M = mean value. $M - 1\ SD$ = mean value minus standard deviation. $M + 1\ SD$ = mean value plus standard deviation.

particular, we posit that performance augmentations might best be achieved by implementing a benchmarking approach (Donthu, Hershberger, and Osmonbekov 2005). Benchmarking fits with the multiobject nature of health care networks, in that it can match the capabilities of comparable physicians as they transform inputs into outputs. Poor performers should aim to replicate the processes implemented by top-performing role models to enhance their own capabilities to transform inputs (i.e., health costs) into outputs (i.e., number of patients treated) (Vorhies and Morgan 2005). Thus, improving the performance of network partners based on benchmarking should help members build important capabilities by imitating best practices. In contrast with benchmarking efforts among competitors, imitative learning and process replication appear particularly effective in networks, assuming an integrative hub can identify and infuse best practices throughout the network.

With our diagnosis discussion, we help extend prior work on service efficiency. First, in presenting a formal methodology for assessing performance and adopting a "what gets measured gets done" perspective, our quantitative approach provides the inevitable basis for enhancing service efficiency and adds to existing conceptual discussions (Grönroos and Ojasalo 2004). Second, compared with relative-to-average approaches suggested to capture service efficiency (Brown and Dev 2000; Rust and Huang 2012), we regard a relative-to-best metric as superior, because it can detect efficiency gaps by benchmarking the efficiency of comparable units. Thus, it is possible to derive realistic goals for efficiency improvements.

The implications of the therapy part of our article are most influential for existing research on detailing (Chintagunta and Desiraju 2005; Janakiraman et al. 2008; Mizik and Jacobson 2004; Venkataraman and Stremersch 2007). In particular, detailing efforts initiated by an organization that acts as a network hub can enhance the responsiveness of network partners and thus improve efficiency. In our health care context, proactive information sharing about the best practices exerted by efficient peers yields a twofold effect on physician responsiveness: First, it helps foster a common understanding of performance improvement goals and plans, which is generally an important prerequisite for the effective implementation of performance enhancements (Vorhies and Morgan 2005). In particular, a common understanding might help increase source credibility. Second, message credibility might be improved when the disclosed best practice patterns are derived through a powerful methodological approach, such as advanced DEA. Both these aspects seem to be potential explanations for the detected efficiency improvements in our study. In addition, our results show that low-volume physicians respond more to detailing calls than do high-volume physicians, which corresponds to findings by Manchanda, Rossi, and Chintagunta (2004) that sales calls induce more prescriptions of detailed drugs, particularly among low-volume physicians. Standard detailing policies in practice, however, focus on high-volume physicians, who are significantly less open to detailing efforts. If they adapted detailing policies, HMOs likely could increase their return on (detailing) investments. Finally, spreading the word about successful service patterns constitutes a form of signaling activity conducted by the network hub. This signaling helps high-referral physicians reduce their sense of uncertainty and risk. This enhanced responsiveness to detailing among both low-volume and high-referral physicians suggests additional insights that might improve returns on detailing (Nair, Manchanda, and Bhatia 2010; Stremersch and Van Dyck 2009).

Managerial Implications

The benchmarking approach we propose can help managers compare efficiency scores across different service providers

(e.g., GPs and specialty physicians) in network structures. Accordingly, from the diagnosis part of our article, we offer a reliable input-to-output metric that HMO managers can use for their performance monitoring and benchmarking. This metric indicates overspendings, i.e. cost reductions for each input that could be achieved without threatening output levels. We therefore suggest adding this metric to HMO dashboards.

In addition, our diagnosis provides a procedure for assessing heterogeneity among network partners. Although accounting for heterogeneity among physicians is vital for obtaining valid results, running segment-specific DEA models is not a desirable strategy per se. Our results suggest for example that for GPs, a one-group model is appropriate, but for specialty physicians, efficiency measurements need separate, per-specialty DEA models. If managers erroneously employ separate DEA models for subgroups of GPs, overspendings could go undetected; in our study, these missed gaps amounted to €6.8 million for the 25% of most inefficient physicians. Moreover, our metric provides valuable information for potential trade-offs between specialty groups. This insight in turn may help HMOs to more effectively balance interests between various specialty groups.

As a third diagnosis-related contribution, we offer an integrated metric that managers should employ when they confront large differences in sample sizes and absolute service costs. This metric can reveal the true resource savings potential. For example, a pure efficiency metric indicates that our health care organization should focus on optimizing ENT, but an integrated metric (i.e., trading off efficiency with absolute costs) recommends a focus on orthopedics. Redeploying the improvement efforts from ENT to orthopedics would yield additional projected savings of €43,900 per physician. That is, focusing only on service efficiency disregarding cost levels can be myopic.

Thus, the potential improvements revealed through DEA are noticeable. Developing such analysis abilities requires that organizations focus increasingly on measurement techniques. To help employees master such techniques, the hub should offer training and seminars and perhaps even automate their use across the board and at operational levels.

In the move from diagnosis to therapy, we uncover further managerial recommendations. First, managers should use the diagnosis results proactively to develop cornerstones of actionable programs that exploit the potential efficiency gains. As our results reveal, it appears beneficial for network hubs to initiate detailing efforts that can improve GP efficiency. Through detailing, the hub was able to inform network partners about the benchmarking results and reveal their current resource utilization patterns. The high response to the detailing program in terms of willingness to reconfigure service patterns (i.e., 92% of the physicians in our study agreed to participate in a detailing call) indicates that physicians assign high source credibility to the network hub. Second, by using sound measurement techniques such as DEA, the HMO can communicate its commitment to rigorous performance monitoring, which is likely a prerequisite for acceptance of performance improvement programs. Using DEA, health care managers can precisely communicate which inputs they should adjust and to what extent if they hope to improve performance.

Third, we provide evidence of varying degrees of responsiveness to detailing calls. According to our findings, detailers should target low-volume and high-referral (i.e., risk-averse) physicians. Although these physicians offer less efficiency improvement potential, their share of potential exploited is higher (i.e., low potential—high share), so these segments should be primary targets of future communication campaigns. This low volume/high-referral pattern is typical for physicians early in their carrier and thus targeting younger physicians can mitigate the accumulation of inefficiencies across the physician life cycle. In contrast, high-volume and low-referral physicians should not be a primary target group for detailing, because they do not display high ability to exploit improvement potential (i.e., high potential—low share).

The results from both the diagnosis and therapy elements of our study can be generalized to other service industries that rely on hub-and-spoke network structures, such as retailing (outlets as providers), financial services (branches as providers), or educational services (schools as providers). Our proposed framework can measure the efficiency of networks with two levels of partners that provide distinct subservices, such as generalists (e.g., GPs who conduct regular health checks) versus specialists (e.g., orthopedic surgeons who treat problems of the musculoskeletal system) (Stremersch and Van Dyck 2009). In such two-level networks, poor service provision gets exacerbated along the value chain, similar to a negative bullwhip effect in supply chains. Comparing efficiency across providers and disseminating the processes of the best performers instead fosters service harmonization and can boost the efficiency of the entire service delivery chain.

Limitations and Future Research Directions

This study reveals interesting avenues for research, some of which result from its limitations. First, though we controlled for case mixes and service quality, the results could be improved by expanding the set of output variables. Data about case severity and service quality would be desirable; they were not available for our study. Second, tracking the sources of improvements to service provider performance could be advanced if studies incorporated more descriptive variables that offer potential segmentation criteria. For example, an application of the Malmquist index approach might identify performance changes due to technological progress—that is, the shift of the frontier—in addition to performance changes achieved through enhanced skills and routines (Luo and Donthu 2006). In combination with (more) data on exogenous variables, such an analysis would clarify the portion of performance change attributable to environmental characteristics, which physicians generally cannot control. Third, researchers should test the validity of our findings by exploring the performance of networks in other service industries, such as retailing, financial services, or education.

Conclusion

This study suggests that due to the harmonization of assets and the promotion of standards, health care services coproduced within network-like structures yield the potential to outperform services provided by single providers (i.e., physicians). However, as physicians are functionally dependent network spokes, bad performance of one party may largely undermine the performance of the succeeding party. To improve performance in health care networks, this research proposes that the initiating and coordinating network hub (e.g., a HMO) should establish a two-step benchmarking approach. The first step (the *diagnosis* step) is concerned with measuring service provider performance and identifying the best performers in a given service function. The second step (the *therapy* step) relates to a contact program that aims at improving the abilities of poor performers.

Across two empirical studies with GPs and specialty physicians, tremendous overspendings are disclosed in the first step. Moreover, suggestions on cost reductions that could be achieved for each input without threatening output levels are presented. With regard to the second step, we find that detailing efforts initiated by the network hub helped physicians to improve their performance. Through detailing, the hub is able to inform network partners about the best practices exerted by benchmark physicians and contrast them to their current resource utilization patterns. Finally, we show that managers of HMOs should seek out physicians with smaller practices and high-referral (i.e., risk-averse) physicians as targets for detailing because they are especially responsive to these initiatives. In sum, our findings not only provide empirical support for the proposed two-step approach of performance management in health care networks but also disclose policies that increase the return on (detailing) investments.

Appendix A

Linear Mixed Model:

$$\text{Number of Patients}_{it} = \mu + \beta_1 FfS_{it} + \beta_2 CoM_{it} + \beta_3 CoR_{it} + \beta_{\text{cova}}\text{Covariates}_{it} + \varepsilon_{it} + \zeta_i + \Omega_t, \quad (2)$$

where

FfS_{it} = fee-for-service;
CoM_{it} = cost of medication;
CoR_{it} = cost of referrals;
Covariates_{it}= covariates (controls) for physician i at time t, including age, gender, and district of operation;
μ = overall grand intercept;
β_j = influence of jth explanatory variable;
ε_{it} = classical disturbance term, normally and independently distributed with variance σ_ε^2;
ζ_i = random disturbance of ith physician, constant across periods and normally and independently distributed with variance σ_ζ^2; and
Ω_t = random disturbance of tth year, constant across physicians and normally and independently distributed with variance σ_Ω^2.

Bootstrap Procedure:

The estimates $\hat{\theta}_k$ and the bootstrap estimates $\hat{\theta}_k^*$ are related as follows:

$$(\hat{\theta}_k - \theta_k)|\mathcal{S} \overset{\text{approx.}}{\sim} (\hat{\theta}_k^* - \hat{\theta}_k)\Big|\mathcal{S}^*, \quad (3)$$

where $\mathcal{S}^*$ indicates a bootstrap sample of pseudodata, and $\hat{\theta}_k^*$ is a bootstrap estimate of the efficiency of observation k. Because statistical estimates of the frontier are obtained from finite samples, the corresponding efficiency measures are biased, and this bias depends on the sampling variation of the frontier. The key expression in Equation 3 enables us to estimate the bias of the DEA estimator, $\text{bias}_{\mathcal{S},k} = E_{\mathcal{S}}(\hat{\theta}_k) - \theta_k$, by its bootstrap counterpart $\text{bias}_{\mathcal{S}^*,k} = E_{\mathcal{S}^*}(\hat{\theta}_k) - \theta_k$. The latter quantity can be approximated with bootstrap values $\theta_{k,b}^*$, $b = 1, \ldots, B$, where B is the number of bootstrap replications. Then, we can compute the bootstrap bias estimate $\text{bias}_{\mathcal{S}^*,k}$ as

$$\text{bias}_{\mathcal{S}^*,k} = B^{-1}\sum_{b=1}^{B} e_{k,b}^* - \hat{\theta}_k = \bar{\theta}_k^* - \hat{\theta}_k. \quad (4)$$

Then the bias-corrected estimator $\tilde{\theta}_k$ can be computed as

$$\tilde{\theta}_k = \hat{\theta}_k - \text{bias}_{\mathcal{S}^*,k} = 2\hat{\theta}_k - \bar{\theta}_k^*. \quad (5)$$

These new efficiency scores provide an estimate and correction for the bias, using Equation 3. To test for heterogeneity across subsamples, we calculate the average efficiency score derived for one subsample, $M1$, θ_{M1}, and divide it by the average efficiency derived for the rest of the observations θ_{R1}. Our test statistic is then

$$\tau_{M1} = \bar{\theta}_{M1}/\bar{\theta}_{R1}. \quad (6)$$

If the null hypothesis is true and two subsamples do not differ in their average efficiency, $\tau_{M1} = 1$. The distribution of the test statistic can be derived by bootstrapping the efficiency scores for both samples and calculating the test statistic B times. Thus, a critical value for the test—whether any $\hat{\tau}$ differ significantly from unity—can be obtained (Simar and Wilson 2007).

The essential steps of the algorithm for deriving bias-corrected efficiency scores are as follows (Simar and Wilson 1998): First, assume that the process generating efficiency scores θ_i is $(\theta_1, \ldots, \theta_n) \sim$ i.i.d. $F(x,y)$, where $F(x,y)$ is a density function on (0,1). The process of generating x_i conditionally on the observed output values y_i and the observed proportion of inputs is completely characterized by the density function $F(x,y)$ and $X^\delta\langle x_i \mid y_i\rangle$, where $X^\delta\langle x_i \mid y_i\rangle$ is the level of the inputs the unit should reach to be on the efficient frontier

with the same level of outputs and the same proportion of inputs. That is,

$$X^{\delta}\langle x_i|y_i\rangle = \hat{\theta}_i x_i. \quad (7)$$

Now i.i.d. bootstrap samples $(x_i^*, y_i^*), i = 1, \ldots, n$ are drawn from a density $\hat{F}(x,y)$ on θ, where $\hat{F}(x,y)$ is an estimator of the joint density of (x,y) on θ. Because the input-oriented scores are bounded from above by unity, the DEA estimator produces many efficient units with $\hat{\theta}_i = 1$. Consequently, $\hat{F}(x,y)$ puts a positive mass at $\theta = 1$ and will provide poor estimates of $F(x,y)$ near the upper bound (1) of its support. To address this problem, we can smooth $\hat{F}(x,y)$ through a Kernel smoother (Silverman 1986). The optimal value for the smoothing parameter (bandwidth of the Kernel density estimator) is the value that minimizes the mean integrated square error. Having derived an optimal value for the bandwidth, we can simulate the corresponding density, which permits us to draw pseudo-scores that follow the same distribution as scores obtained with the original sample. The simulated scores produce several B pseudo data sets, used to obtain B new sets of efficiency scores. These new efficiency scores enable us to estimate and correct for bias using Equation 5.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. The axioms underlying this type of nonparametric frontier approach are well known; see Banker, Charnes, and Cooper (1984) or Parsons (1994) for example.
2. We use an input-oriented, variable returns-to-scale specification of the DEA model, which offers the additional advantage of accounting for different economies of scale. In contrast with a constant return to scale DEA model, our model does not assume that every physician operates at the most productive scale size (in terms of numbers of patients treated). A free scaling up or down of inputs and outputs thus is not allowed, and the efficiency calculation does not take into account the elimination of scale efficiencies. Even if a physician practice does not operate at the optimal scale size but instead is too small or too large, it can be identified as fully efficient and serve as a benchmark with respect to transforming inputs into outputs. This trait is desirable, because a physician is not necessarily accountable for a practice size that differs from the optimum (Rosenman and Friesner 2004).
3. The close ties between the physicians and the health care network led to 92% of the contacted GPs agreeing to discuss their resource use.
4. Recall that for GPs whose absolute cost levels are homogeneous, the standard efficiency score (without cost adjustment) provided by DEA is a sufficient basis for assessing responses to the detailing call.

References

Achrol, Ravi S. and Philip Kotler (1999), "Marketing in the Network Economy," *Journal of Marketing*, 63 (4), 146-163.

Agrell, Per J. and Peter Bogetoft (2001), "Should Health Regulators Use DEA?," in *Coordination and Incentives in Health Care*, E. G. Fidalgo, ed. Barcelona: Asociación de Economia de la Salud, 133-154.

Aiken, Leona and Stephen G. West (1991), *Multiple Regression: Testing and Interpreting Interactions*. Newbury Park, CA: SAGE.

Baltagi, Badi (2001), *Econometric Analysis of Panel Data*. New York, NY: John Wiley.

Banker, Rajiv D., Abraham Charnes, and William W. Cooper (1984), "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis," *Management Science*, 30 (9), 1078-1092.

——— and Hsihui Chang (2006), "The Super-Efficiency Procedure for Outlier Identification, not for Ranking Efficient Units," *European Journal of Operational Research*, 175 (2), 1311-1320.

Berry, Leonard L. and Neeli Bendapudi (2007), "Health Care: A Fertile Field for Service Research," *Journal of Service Research*, 10 (2), 111-122.

——— and Ann M. Mirabito (2010), "Innovative Healthcare Delivery," *Business Horizons*, 53 (2), 157-169.

Brown, James R. and Chekitan S. Dev (2000), "Improving Productivity in a Service Business: Evidence From the Hotel Industry," *Journal of Service Research*, 2 (4), 339-354.

Brown, Rayna (2006), "Mismanagement or Mismeasurement? Pitfalls and Protocols for DEA Studies in the Financial Services Sector," *European Journal of Operational Research*, 174 (2), 1100-1116.

Charnes, Abraham, William W. Cooper, and Edwardo Rhodes (1978), "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research*, 2 (6), 429-444.

Chilingerian, Jon A. and H. David Sherman (1997), "DEA and Primary Care Physician Report Cards: Deriving Preferred Practice Cones from Managed Care Service Concepts and Operation Strategies," *Annals of Operations Research*, 73 (1), 35-66.

Chintagunta, Pradeep C. and Ramarao Desiraju (2005), "Strategic Pricing and Detailing Behavior in International Markets," *Marketing Science*, 24 (1), 67-80.

Donthu, Naveen and Boonghee Yoo (1998), "Retail Productivity Assessment Using Data Envelopment Analysis," *Journal of Retailing*, 74 (1), 89-105.

———, Edmund K. Hershberger, and Talai Osmonbekov (2005), "Benchmarking Marketing Productivity using Data Envelopment Analysis," *Journal of Business Research*, 58 (11), 1474-1482.

Dyson, Robert G., Rachel Allen, Ana S. Camanho, Victor V. Podinovski, Claudia S. Sarrico, and Estelle A. Shale (2001), "Pitfalls and Protocols in DEA," *European Journal of Operational Research*, 132 (2), 245-259.

Evanschitzky, Heiner (2007), "Market Orientation of Service Networks: Direct and Indirect Effects on Sustained Competitive Advantage," *Journal of Strategic Marketing*, 15 (4), 349-368.

Färe, Rolf, Shawna Grosskopf, and C. A. Knox Lovell (1994), *Production Frontiers.* Cambridge, England: Cambridge University Press.

Frei, Frances X. and Patrick T. Harker (1999), "Measuring the Efficiency of Service Delivery Processes: An Application to Retail Banking," *Journal of Service Research*, 1 (4), 300-312.

Ginsburg, Paul B., Bradley C. Strunk, Michelle I. Banker, and John P. Cookson (2006), "Tracking Health Care Costs: Continued Stability but at High Rates in 2005," *Health Affairs: The Policy Journal of the Health Sphere*, 25 (6), 486-495.

Govind, Rahul, Rabikar Chatterjee, and Vikas Mittal (2008), "Timely Access to Health Care: Customer-Focused Resource Allocation in a Hospital Network," *International Journal of Research in Marketing*, 25 (4), 294-300.

Grilliches, Zvi and Jerry A. Hausman (1986), "Errors in Variables in Panel Data," *Journal of Econometrica*, 31 (1), 93-118.

Grönroos, Christian and Katri Ojasalo (2004), "Service Productivity: Towards a Conceptualization of the Transformation of Inputs into Economic Results in Services," *Journal of Business Research*, 57 (4), 414-423.

Guth, Kim Ann and Brian Kleiner (2005), "Quality Assurance in the Health Care Industry," *Journal of Health Care Finance*, 31 (3), 33-40.

Hollfelder, Jack (2002), "A New Era for Marketing Health Services," *Marketing Health Services*, 22 (2), 48.

Hollingsworth, Bruce (2008), "The Measurement of Efficiency and Productivity of Health Care Delivery," *Health Economics*, 17 (10), 1107-1128.

Horsky, Dan and Paul Nelson (1996), "Evaluation of Salesforce Size and Productivity through Efficient Frontier Benchmarking," *Marketing Science*, 15 (4), 301-320.

Irwin, Julie R. and Gary H. McClelland (2003), "Negative Consequences of Dichotomizing Continuous Predictor Variables," *Journal of Marketing Research*, 40 (3), 366-371.

Janakiraman, Ramkumar, Shantanu Dutta, Catarina Sismeiro, and Philip Stern (2008), "Physicians' Persistence and Its Implications for Their Response to Promotion of Prescription Drugs," *Management Science*, 54 (6), 1080-1093.

Kamakura, Wagner A., Vikas Mittal, Fernando De Rosa, and José Afonso Mazzon (2002), "Assessing the Service-Profit Chain," *Marketing Science*, 21 (3), 294-317.

Keh, Hean Tat, Singfat Chu, and Jiye Xu (2006), "Efficiency, Effectiveness and Productivity of Marketing in Services," *European Journal of Operational Research*, 170 (1), 265-276.

Kneip, Alois, Byeong U. Park, and Léopold Simar (1998), "A Note on the Convergence of Nonparametric DEA Estimators for Production Efficiency Scores," *Econometric Theory*, 14 (6), 783-793.

Kravitz, Richard L., Sheldon Greenfield, William Rogers, Willard G. Manning, Jr., Michael Zubkoff, Eugene C. Nelson, Alvin R. Tarlov, and John E. Ware, Jr. (1992), "Differences in the Mix of Patients Among Medical Specialties and Systems of Care: Results From the Medical Outcomes Study," *Journal of the American Medical Association*, 267 (12), 1617-1623.

Luo, Xueming (2007), "Consumer Negative Voice and Firm-Idiosyncratic Stock Returns," *Journal of Marketing*, 71 (July), 75-88.

——— and Naveen Donthu (2006), "Marketing's Credibility: A Longitudinal Investigation of Marketing Communication Productivity and Shareholder Value," *Journal of Marketing*, 70 (4), 70-91.

Manchanda, Puneet, Peter E. Rossi, and Pradeep K. Chintagunta (2004), "Response Modeling with Nonrandom Marketing-Mix Variables," *Journal of Marketing Research*, 41 (4), 467-478.

Maxwell, Scott E. and Harold D. Delaney (1993), " Bivariate Median Splits and Spurious Statistical Significance," *Psychological Bulletin*, 113 (1), 181-190.

Metters, Richard and Ann Marucheck (2007), "Service Management—Academic Issues and Scholarly Reflections from Operations Management Researchers," *Decision Sciences*, 38 (2), 195-214.

Meyer Goldstein, Susan and Peter T. Ward (2004), "Performance Effects of Physicians' Involvement in Hospital Strategic Decisions," *Journal of Service Research*, 6 (4), 361-372.

Mizik, Natalie and Robert Jacobson (2004), "Are Physicians 'Easy Marks'? Quantifying the Effects of Detailing and Sampling on New Prescriptions," *Management Science*, 50 (12), 1704-1715.

Nair, Harikesh S., Puneet Manchanda, and Tulikaa Bhatia (2010), "Asymmetric Social Interactions in Physician Prescription Behavior: The Role of Opinion Leaders," *Journal of Marketing Research,* 47 (5), 883-895.

Narayanan, Sridhar, Puneet Manchanda, and Pradeep K. Chintagunta (2005), "Temporal Differences in the Role of Marketing Communication in New Product Categories," *Journal of Marketing Research*, 42 (3), 278-290.

Newhouse, Joseph P. (1994), "Frontier Estimation: How Useful a tool for Health Economics," *Journal of Health Economics*, 13 (3), 317-322.

Ostrom, Amy L., Mary Jo Bitner, Stephen W. Brown, Kevin A. Burkhard, Michael Goul, Vicki Smith-Daniels, Haluk Demirkan, and Elliot Rabinovich (2010), "Moving Forward and Making a Difference: Research Priorities for the Science of Service," *Journal of Service Research*, 13 (1), 4-36.

Ozcan, Yasar A. (1998), "Physician Benchmarking: Measuring Variation in Practice Behavior in Treatment of Otitis Media, " *Health Care Management Science*, 1 (1), 5-17.

Pai, Chih-Wen, Yasar A. Ozcan, and H. Joanna Jiang (2000), "Regional Variation in Physician Practice Pattern: An Examination of Technical and Cost Efficiency for Treating Sinusitis," *Journal of Medical Systems*, 24 (2), 103-117.

Parsons, Leonard J. (1994), "Productivity Versus Relative Efficiency in Marketing: Past and Future?," in *Research Traditions in Marketing*, Gilles Laurent, Gary L. Lilien, and Bernard Pras, eds. Norwell, MA: Kluwer Academic, 169-196.

Rosenman, Robert and Dan Friesner (2004), "Scope and Scale Inefficiencies in Physician Practices," *Health Economics*, 13 (11), 1091-1116.

Rust, Roland T. and Ming-Hui Huang (2012), "Optimizing Service Productivity," *Journal of Marketing*, 76 (2), 47-66.

Ruston, Annmarie (2004), "Risk, Anxiety and Defensive Action: General Practitioner's Referral Decisions for Women Presenting with Breast Problems," *Health, Risk & Society*, 6 (1), 25-38.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, 2nd ed. Boston, MA: Houghton Mifflin Harcourt.

Silverman, Bernard W. (1986), *Density Estimation for Statistics and Data Analysis*. London, England: Chapman and Hall.

Simar, Léopold (2003), "Detecting Outliers in Frontier Models: A Simple Approach," *Journal of Productivity Analysis*, 20 (3), 391-424.

Simar, Léopold and Paul W. Wilson (1998), "Sensitivity Analysis of Efficiency Scores: How to Bootstrap in Nonparametric Frontier Models," *Management Science*, 44 (1), 49-61.

——— and ——— (2007), "Statistical Inference in Non-Parametric Frontier Models: Recent Developments and Perspectives," in *The Measurement of Productive Efficiency*, H. O. Fried, C. A. K. Lovell, and S. S. Schmidt, eds. Oxford, England: Oxford University Press.

Singh, Jagdip (1991), "Understanding the Structure of Consumers' Satisfaction Evaluations of Service Delivery," *Journal of the Academy of Marketing Science*, 19 (3), 223.

Spence, Michael (1973), "Job Market Signaling," *Quarterly Journal of Economics*, 87 (3), 355-374.

Spendolini, Michael J. (1992), *The Benchmarking Book.* New York, NY: American Management Association.

Stiglitz, Joseph E. (2000), "The Contributions of the Economics of Information to Twentieth Century Economics," *Quarterly Journal of Economics*, 115 (4), 1441-1478.

Stremersch, Stefan and Walter Van Dyck (2009), "Marketing of the Life Sciences: A New Framework and Research Agenda for a Nascent Field," *Journal of Marketing*, 73 (4), 4-30.

Venkataraman, Sriram and Stefan Stremersch (2007), "The Debate on Influencing Doctors' Decisions: Are Drug Characteristics the Missing Link?," *Management Science*, 53 (11), 1688-1701.

Vorhies, Douglas W. and Neil A. Morgan (2005), "Benchmarking Marketing Capabilities for Sustainable Competitive Advantage," *Journal of Marketing*, 69 (1), 80-94.

Wagner, Janet, Daniel Shimshak, and Michael A. Novak (2003), "Advances in Physician Profiling: The Use of DEA," *Socio-Economic Planning Sciences*, 37 (2), 141-163.

Wan, Thomas T. H. and Bill B. L. Wang (2003), "Integrated Healthcare Networkś Performance: A Growth Curve Modeling Approach," *Health Care Management Science*, 6 (2), 117-124.

Wilson, Paul W. (2003), "Testing Independence in Models of Productive Efficiency," *Journal of Productivity Analysis*, 20 (3), 361-390.

Zhang, Yun and Robert Bartels (1998), "The Effect of Sample Size on the Mean Efficiency in DEA with an Application to Electricity Distribution in Australia, Sweden and New Zealand," *Journal of Productivity Analysis*, 9 (3), 187-204.

Zuckerman, Stephen, Jack Hadley, and Lisa Iezzoni (1994), "Measuring Hospital Efficiency with Frontier Cost Functions," *Journal of Health Economics*, 13 (3), 255-280.

Bios

Maik Hammerschmidt, PhD (University of Mannheim), is a Professor of Marketing and Chair in Marketing and Innovation Management at the University of Goettingen, Germany. His current research focuses on service marketing, brand management, innovation management, and marketing performance measurement. He has coauthored and coedited four books in these fields and has published in journals such as *Journal of Marketing, Journal of the Academy of Marketing Science,* and *Journal of Service Research.* He has won seven research awards, including an Overall Best Paper Award of the American Marketing Association.

Tomas Falk, PhD (University of Mannheim), is the ConCardis Professor of Consumer Behavior and Marketing at the EBS Business School, Germany. His research interest and expertise focuses on service channel management, service quality management, self-service technologies, and service employee behavior. His research has been published in journals such *Journal of Marketing, Journal of the Academy of Marketing Science,* and *Journal of Service Research*. He has won several research awards, including an Overall Best Paper Award of the American Marketing Association.

Matthias Staat, PhD (University of Mannheim), is a Senior Project Manager at DSC Consulting, Germany. His research interest and expertise focuses on efficiency analysis, health economics, pricing, and value-based management. He has coedited a book on marketing efficiency and his research has been published in journals such *European Journal of Operational Research, Annals of Operational Research* and *Applied Economics.*