Performance Profiling in Primary Care: Does the Choice of Statistical Model Matter?

Frank Eijkenaar and René C. J. A. van Vliet

Med Decis Making 2014 34: 192 originally published online 6 August 2013
DOI: 10.1177/0272989X13498825

Performance Profiling in Primary Care: Does the Choice of Statistical Model Matter?

Frank Eijkenaar, MSc, René C. J. A. van Vliet, PhD

Background. *Profiling is increasingly being used to generate input for improvement efforts in health care. For these efforts to be successful, profiles must reflect true provider performance, requiring an appropriate statistical model. Sophisticated models are available to account for the specific features of performance data, but they may be difficult to use and explain to providers.* ***Objective.*** *To assess the influence of the statistical model on the performance profiles of primary care providers.* ***Data Source.*** *Administrative data (2006–2008) on 2.8 million members of a Dutch health insurer who were registered with 1 of 4396 general practitioners.* ***Methods.*** *Profiles are constructed for 6 quality measures and 5 resource use measures, controlling for differences in case mix. Models include ordinary least squares, generalized linear models, and multilevel models. Separately for each model, providers are ranked on z scores and classified as outlier if belonging to the 10% with the worst or best performance. The impact of the model is evaluated using the weighted kappa for rankings overall, percentage agreement on outlier designation, and changes in rankings over time.* ***Results.*** *Agreement among models was relatively high overall (kappa typically >0.85). Agreement on outlier designation was more variable and often below 80%, especially for high outliers. Rankings were more similar for processes than for outcomes and expenses. Agreement among annual rankings per model was low for all models.* ***Conclusions.*** *Differences among models were relatively small, but the choice of statistical model did affect the rankings. In addition, most measures appear to be driven largely by chance, regardless of the model that is used. Profilers should pay careful attention to the choice of both the statistical model and the performance measures.*
Key words: *profiling; risk adjustment; report cards; econometric methods; performance measures; managed care.* ***(Med Decis Making 2014;34:192–205)***

Purchasers and other actors in health care are increasingly interested in comparative information on the performance of health care providers. Variation in resource use and quality of care is well documented, and in many countries, purchasers increasingly use specific measurement methodologies to gain insight into providers' relative performance. The data derived from these measurements are often summarized in performance profiles, which may contain information on various aspects of providers' performance and can be used in various ways to spur improvement. For example, they may be used to provide feedback to providers,[1] to allocate incentive payments,[2] and to steer consumers to high-performing providers via public reporting[3] and/or creating selective and tiered provider networks.[4]

Evidently, profiling is useful for these purposes only if profiles reflect true provider performance. Random variation and differences in case mix may explain large portions of observed performance variation and can obscure the signal of providers' true performance.[5,6] Therefore, if they are to produce useful input for improvement efforts, profiles must take these factors into account. This is true especially for resource use and (clinical) outcome measures (e.g., HbA1c levels of patients with diabetes, hospital

Received 19 November 2012 from the Institute of Health Policy and Management, Erasmus University Rotterdam, Rotterdam, the Netherlands. This work was presented at the European Conference on Health Economics (ECHE), Zurich, Switzerland, July 2012. Revision accepted for publication 28 June 2013.

Supplementary material for this article is available on the *Medical Decision Making* Web site at http://mdm.sagepub.com/supplemental.

Address correspondence to Frank Eijkenaar, MSc, Institute of Health Policy and Management, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3000 DR Rotterdam, the Netherlands; telephone: +31 10 408 9183; fax: +31 10 408 9094; e-mail: eijkenaar@bmg.eur.nl.

DOI: 10.1177/0272989X13498825

readmissions) because they are particularly sensitive to random chance and relevant patient characteristics such as age and disease severity. To mitigate the role of random variation, measures should be used only when there is sufficient between-provider variation and when a sufficient number of patients can be sampled. To mitigate incentives for risk selection and to ensure fair comparisons, adequate risk adjustment must be applied.[7–12]

In profiling, comparing providers' observed performance to their expected performance (based on their case mix) has become standard.[7] In practice, purchasers typically calculate expected performance using model-derived (patient-level) predictions. Therefore, in addition to accurate data on relevant patient characteristics, risk adjustment requires an appropriate statistical model, the choice of which will depend on characteristics of the data[13] such as the type of data (binary, count, continuous) and the shape of the distribution (e.g., roughly normal or highly skewed). In practice, however, other considerations will likely play a role in this choice as well. Instead of relying on expensive external expertise, purchasers often perform these analyses themselves (typically on an annual basis) and will therefore prefer models that are easy to use and maintain. In addition, for risk adjustment to fulfill its purpose, it is important that providers whose performance is being profiled understand and support the method. If not, even when differences in case mix are appropriately taken into account, providers may still view the risk-adjustment method as a ''black box'' and be suspicious of its validity,[14] which could undermine the entire profiling system. Therefore, where possible, purchasers would opt for keeping the risk-adjustment method simple. An often-used method that can easily be applied to many types of performance data is ordinary least squares (OLS). However, performance data often have specific characteristics rendering OLS unsuitable. More sophisticated models, although more difficult to explain and maintain, will usually fit these data better. Nonetheless, despite often being the less suitable method, in practice OLS (at the patient level) could generate similar *profiling* results (at the provider level).

In this article, we use administrative data from a large Dutch health insurer to compare statistical models that can be used for analyzing and risk adjusting the performance of Dutch general practitioners (GPs) and health centers (HCs) on several measures of quality and resource use. The insurer has been implementing several performance-profiling programs in the Dutch primary care sector and, for the reasons mentioned above, wanted insight in the extent to which simple methods (OLS) yield similar profiling results compared with more appropriate sophisticated methods. Previous studies have looked at the impact on profiling results of varying the risk-adjustment methodology,[15–22] treatment of patients with extreme values,[23] definition of performance index,[20,24–26] and method for categorizing providers in different performance categories (e.g., high, average, low).[5,27] This study focuses on the impact of the statistical model, holding constant the set of risk adjusters and other factors. Although there have been some other studies that assessed the influence of the *statistical model* on performance-profiling results, these studies included only a few model types in their comparisons (e.g., 2 or 3). In addition, each of these studies evaluated the impact for only 1 performance measure: satisfaction with asthma care,[17] managed care pharmacy expenses,[28] or in-hospital mortality for patients undergoing coronary artery bypass grafting surgery.[21,29] Our study compares more statistical models and assesses their impact for 11 performance measures applicable to 3 different patient populations. In addition, by comparing annual provider rankings over 3 adjacent years, we also provide insight into the influence of the statistical model on the stability of profiling results over time, which has not been done in previous work. Large fluctuations would indicate that the risk-adjusted measures are mainly driven by random chance instead of true provider performance.

METHODS

Study Setting and Data

In the Dutch health care sector for curative care, private, risk-bearing insurers are expected to act as prudent purchasers of care on behalf of their members. To adequately fulfill this role, insurers can use several managed-care instruments, including selective contracting, financial incentives, and performance feedback to providers. Each of these instruments requires an adequate profiling system. In this study, performance profiles are constructed for GPs and HCs using administrative data for the years 2006 to 2008 obtained from a Dutch insurer. For each year, data on about 2.8 million members are available, including sociodemographic characteristics and proxies for health status. In the Netherlands, these data are routinely available in health insurers' files at no additional cost. For each member, it is

known with which GP he or she was registered in a particular year. In the Netherlands, GPs have fixed patient panels and act as gatekeepers to hospital care. Thus, GPs can influence the amount and type of hospital care their patients use. A small but increasing number of GPs hold practice in an HC, which is an entity in which multiple GPs (typically 4 or 5) and other primary care providers (e.g., physiotherapists, dietitians) provide and coordinate care, usually from the same building. In our data, a GP may or may not be affiliated with an HC. Thus, for each member, our data provide a link to his or her own GP, and, if this GP is affiliated with an HC, also a link to this HC. Approximately 10% of the GPs in our data set were affiliated with an HC, so the vast majority of members did not receive primary care from GPs working in an HC.

Dependent Variables (Performance Measures)

Using the administrative data, we constructed 3 types of performance measures: expenses (3 measures), utilization of hospital care (2 measures), and clinical quality (6 measures). The expenses measures are GP expenses (generated through visits and diagnostic tests/examinations), prescription medication expenses, and total expenses (the sum of GP, medication, and hospital expenses). Regarding utilization, the total number of inpatient admissions and outpatient visits are available. Both are indicated by diagnosis treatment combinations, which were implemented in the Dutch health care system to facilitate contracting for hospital services.[30] A diagnosis treatment combination is a predefined care product, selected by the medical specialist based on the patient's condition and representing all hospital procedures/services related to treating a patient with a specific diagnosis within a fixed period. It is similar to a diagnosis-related group used by, for example, Medicare in the United States, except that diagnosis treatment combinations are more broadly defined and also include the payment for medical specialists. Finally, providers are compared on clinical process and outcome quality for patients with diabetes mellitus (DM) and patients with chronic obstructive pulmonary disease (COPD). For DM, the percentage of patients on statins and the number of DM-related hospital admissions were available. For COPD, 3 process measures were defined: the percentage of patients using bronchodilators, the percentage of patients using prednisone, and the percentage of patients receiving physiotherapy. The number of COPD-related hospital admissions was used as outcome. The result is 3 types of dependent variables: continuous (expenses, lower is better), count (utilization, lower is better), and binary (clinical processes, higher is better).

A small number of extreme outlier members (109) were excluded to minimize distorting effects on coefficients and profiling results and to increase the chance of algorithm convergence for the more complex statistical models. In addition, because the distributions of medication and total expenses are highly skewed and we could not rule out the possibility that several extremely high values (between 50 and 100 patients per year) were erroneous (e.g., miscoded), based on a visual inspection these variables were top-coded at €25,000 and €125,000, respectively. Dependent variables for members enrolled for less than a year were annualized and weighted based on months of enrollment. Providers were included only if they had $\geq$100 patients in each year for the non–disease-specific variables. For the disease-specific variables, providers had to have $\geq$30 patients to be included (we chose these thresholds because they are commonly used in practice and in the literature). After applying these restrictions, 4396, 628, and 517 GPs were included for the non–disease-specific, DM, and COPD measures, respectively. For HCs, these numbers are 120, 45, and 35.

Independent Variables (Risk Adjusters)

The models adjust for various patient characteristics, all of which were derived from the administrative data (Table 1). In addition to age and sex, we included 5 indicators of socioeconomic status, 3 of which were measured at the member's ZIP-code level. For example, the 3 categories of educational level (low, medium, high) relate to the average educational level of people living in the member's ZIP-code area. The variable ethnicity is based on the percentage of persons in the ZIP-code area of whom at least 1 parent was born in Turkey, Africa, Latin America, or Asia (excluding Japan). This variable was included because different ethnic groups may exhibit different patterns of utilization[31] and may not be equally adherent to recommended treatment.[32,33] The variable urbanization is based on the number of adjacent addresses per square kilometer for 2006 and on the number of inhabitants in the member's town/city of residence for 2007–2008. We also included 2 proxies for health status: pharmacy-based cost groups and diagnosis cost groups. Both proxies have been developed in the context of the Dutch risk-equalization scheme (used to calculate risk-adjusted capitation

Table 1 Included risk adjusters

Age-sex interactions (38 categories)
Yes/no living in a deprived area
Monthly income (ZIP-code, 10 categories)
Educational level (ZIP-code, 3 categories)
Ethnicity (ZIP-code level, 6 categories in 2006, 5 in 2007–2008)
Urbanization (5 categories in 2006, 8 in 2007–2008)
Yes/no died in year of interest
Pharmacy-based cost groups (20 categories/comorbidities)
Diagnosis cost groups (13 categories)

payments for health insurers[34,35]) and are designed to identify patients with chronic conditions. Pharmacy-based cost groups are based on prior (outpatient) use of medication. A member is assigned to a certain pharmacy-based cost group if prescribed $\geq$181 defined daily doses of a particular disease-specific medication in the prior year. For example, if a member was prescribed $\geq$181 defined daily doses of insulin in year t, he or she will be classified in the pharmacy-based cost group for diabetes type 1 in year $t + 1$. Our data distinguishes 20 pharmacy-based cost groups (members can be classified in multiple groups), all of which relate to a certain chronic condition (e.g., DM, heart disease, rheumatoid arthritis, cancer, epilepsy). Members were identified as having DM if classified in the pharmacy-based cost group for DM. COPD patients were defined in a similar way (using the pharmacy-based cost group for chronic nonspecific respiratory conditions) among members 45 years of age or older. Diagnosis cost groups are based on the diagnoses of hospitalizations in the prior year. About 500 diagnosis treatment combinations for which high future expenses are likely were clustered on homogeneity of expenses, resulting in 13 diagnosis cost groups. If a member was admitted to the hospital and classified in one of these diagnosis treatment combinations in year t, this member will be classified in the associated diagnosis cost group in year $t + 1$. Members can be classified in only 1 diagnosis cost group (i.e., the most expensive one).

All risk adjusters were carefully developed for the purpose of explaining cost variation at the individual member level and are therefore appropriate for expenses measures.[36] Because the utilization measures are closely related to (total) expenses, the risk adjusters are also relevant for these measures. This was confirmed when we ran the models; all risk adjusters were typically significantly associated with the dependent variable. However, for the process measures, this was not always the case, especially regarding the diagnosis cost groups. But because the pattern of (lack of) significant associations with the dependent variables was not consistent across models and over time, we chose to include all variables in all models to ensure comparability. As a result, all models use the same risk adjusters.

Model Selection

Expenses and utilization data have specific features that complicate modeling of these data, including a large fraction of people without any consumption (i.e., a large zero mass), skewed distributions, and heteroskedasticity (i.e., nonconstant error variance). As modeling by OLS may lead to imprecise estimates, more robust methods have been proposed that recognize the distribution of the data and are less sensitive to the right tail. Another issue is that many methods assume independent observations. Yet it is likely that in our case, the data are not generated independently but in groups because patients with specific characteristics tend to choose and remain with physicians with specific characteristics.[37] Our procedure of selecting statistical models that can accommodate these features comprised 2 steps. First, we consulted key references in the field of health econometrics and profiling[13,38–41] to create a list of relevant *types* of models:

- OLS was applied to all performance measures, including the binary variables. A linear probability model is justified here because the individual expected probabilities are aggregated to the provider level typically yielding an expected probability between 0 and 1.
- Generalized linear models take into account heteroskedasticity while retaining the original scale, thus making retransformation methods superfluous.[13,42] They accommodate skewness via variance weighting and require specification of a distribution and a nonlinear link function of the dependent variable that can be modeled (by maximum likelihood) as a linear function of independent variables. Using the GENMOD procedure in SAS 9.2, we tested several distributions: normal and gamma for expenses; normal, gamma, Poisson, and negative binomial (negbin) for counts; and binomial for binary variables.
- Two-part models deal with dependent variables with many zeroes by splitting consumption in 2 parts: the probability of any consumption and the level of consumption conditional on having any.[38] Two-part models are estimated for medication expenses (30% zeroes), admissions (92%–95% zeroes), and

outpatient visits (75% zeroes). Parameters are estimated separately for each part (using the same covariates), and the prediction is obtained by multiplying the estimated probability from a probit or logit model by the conditional outcome.

- Multilevel models (MLMs; also known as random-effects models) explicitly model the hierarchical structure of the data, thereby recognizing that nested observations may be correlated. When this is the case, MLMs produce estimates that are more robust to small sample size and more precise as predictions.[13,17,21,43] Intervals around provider-specific performance estimates will also be wider, reflecting the uncertainty arising from both variation between patients within providers and variation between providers.[13,44] Using the GLIMMIX procedure in SAS, we employed 2-level models with a random provider intercept with mean zero and constant variance, adjusting for the fixed effects of patients' risk characteristics. We also considered the NLMIXED procedure but chose GLIMMIX because NLMIXED tends to have problems in achieving an accurate integral approximation in the log-likelihood in models with a relatively large number of random effects.[45] All MLMs were estimated by maximum pseudo-likelihood. We also tried estimating the models by Laplace approximation and adaptive quadrature, but as these techniques often resulted in computational (convergence) problems, we decided not to use them further.

We did not include models with provider fixed effects. The reason is that this would often result in unworkable models given the large number of providers. However, we acknowledge the controversy between fixed- and random-effects models and the fact that both types of models compute provider effects in different ways.[21,28,46,47] We ran OLS models with provider fixed effects for all measures for HCs and for the disease-specific measures for GPs. Results were nearly identical to models without these effects, as also found by others.[28,29]

In step 2 of our selection procedure, for each of the model types, we created a final set of model *specifications* with a comparable fit. Appropriate specifications (i.e., well-fitting links and distributions) were determined using the following criteria and tests:

- Percentage explained variance (R^2): 1 – [variance(residuals) / variance(dependent variable)]
- Mean absolute deviation (MAD): the average of the absolute value of the residuals
- Bayesian information criterion: [$-2 \times \ln$(likelihood)] + [number of parameters $\times \ln(n)$]
- Pregibon's link test[48]
- Modified Hosmer-Lemeshow test
- Calibration: the extent to which the mean expected value approximates the mean observed value (at the member level). If the mean expected value differs from the mean observed value, the model requires recalibration, which is achieved by multiplying each member's expected value by a factor obtained from dividing the overall mean observed value by the overall mean expected value. Model calibration was also assessed by performing an OLS regression with the observed outcome as the dependent variable and the expected outcome as the independent variable. If this yields an intercept of 0 and a slope of 1, recalibration is not necessary
- Adequate convergence of algorithm in all years

We included only converging models with a satisfactory fit in all years. As a result, models with a good fit in a particular year may still have been excluded. We followed this approach for 2 reasons: 1) although excluded model specifications sometimes performed better than some included specifications, differences were small, and 2) having the same models in all years enables calculations on the stability of profiling results over time.

Model Comparison

We calculated agreement among models on provider rankings based on *z* scores. The *z* score has widely been used in profiling and is preferred over other metrics.[10,13,18,49] Using the measure-specific patient-level observed and expected values, we calculated the mean observed and mean expected performance level for each provider in each year by summing the observed and expected patient-level values and dividing by the number of patients per provider. The provider-specific *z* score is then obtained by dividing the difference between these 2 means by the standard error of this difference.

Agreement is measured separately for each measure using the weighted kappa statistic, which measures agreement between rankings beyond agreement due to chance.[50] For each model, we ranked providers on *z* scores and recoded the ranking into 20 equally-sized groups. Next, for each pair of models, we calculated the weighted kappa by comparing both rankings. Finally, for each model, we calculated the average agreement with all the other models using the weighted kappas obtained from the pairwise comparisons with the other models. Models are also compared on the extent to which they agree on outlier designation. A provider is considered an outlier if belonging to the 10% of providers with the worst performance or to the 10% of providers with the best

performance. The average percentage agreement was calculated for each model for both low and high outliers. Finally, models are compared on stability of results over time using the average of the agreement between the rankings of 2006 and 2007, of 2006 and 2008, and of 2007 and 2008.

Agreement statistics are calculated separately for HCs, GPs in an HC, and GPs not in an HC. During 2006 to 2008, HCs participated in a pay-for-performance program in which most of the measures used in this article were included. Variation in profiling results over time for (GPs in) HCs could be a reflection of this program's having an effect. In that case, results will be more stable for GPs not in an HC.

RESULTS

Table 2 provides descriptive statistics for members and providers. Table 3 shows the models that were included for each performance measure as well as some fit statistics for 2008 (results for 2006–2007 can be found in the online appendix; the magnitude of values sometimes differ across years, but patterns are similar). As expected, OLS is often outperformed by several other models, although differences are generally quite small. Regarding the binary measures, OLS yields the lowest R^2 and highest MAD, whereas the MLMs yield the highest R^2 and lowest MAD. A similar pattern can be observed for GP expenses, whereas for other types of expenses, alternatives to OLS do not add much. Regarding the count variables, several models yield lower R^2 and higher MAD values than OLS, but there is always at least 1 model performing better on both statistics. Two-part models are among the models with the lowest R^2 and for admissions and visits also have the highest MAD. Finally, several models needed to be recalibrated (last column of Table 3). Models for which this was most necessary typically had the worst fit (e.g., log-normal for COPD-related admissions and gamma-power for medication expenses).

The R^2 values also provide insight into the importance of risk adjustment. As expected, the models explained a relatively large fraction (22%–38%) of total member-level variation in expenses. This is also true for outpatient visits (36%), whereas for hospital admissions, models explain only about 7% to 12% of the variation. As risk adjustment is undoubtedly important for these measures, the low R^2 values are probably a result of a combination of inadequate risk adjustment and the fact that hospital admissions are relatively rare. Even less variation is explained in 3 of the 4 process measures. The very high R^2 for statins can be explained by very strong associations with some pharmacy-based cost groups (e.g., heart disease).

Agreement among Models per Year

Table 4 presents average levels of agreement for 2008 (figures for 2006–2007 are similar for most measures, see the online appendix; exceptions are higher agreement for HCs for physiotherapy, lower agreement overall but higher agreement on outliers for disease-related admissions, and lower agreement for HCs for outpatient visits in 2006–2007 compared with 2008). Agreement on overall rankings is high, with kappa often greater than 0.90, typically greater than 0.85, and never below 0.74. Agreement on outlier designation is more variable but still quite high and tends to be higher for processes than for outcomes and expenses, for which agreement is often less than 80%. Overall, models tend to agree better on designation of low outliers than of high outliers, although there are exceptions (e.g., GP expenses for HCs). Models agree somewhat better for GPs than for HCs, especially for disease-related admissions and expenses. Finally, models with similar fit statistics may agree poorly on profiling results. For GP expenses, for example, the normal-power model agrees worse with the other model(s) than OLS.

Figure 1 shows the distribution of z scores for GPs for 2 measures: statins and GP expenses. Despite high agreement among models, differences may be large for individual providers. In addition, highly similar rankings do not preclude large differences, which become visible when an absolute threshold is used to discern providers. For example, for statins (Figure 1a), a threshold of (–)2 results in lower agreement on outlier designation between OLS and logit than presented in Table 4. Plots such as in Figure 1 also visualize differences between measures. For example, assuming an absolute threshold, many more GPs will be classified as outliers for GP expenses (Figure 1b) than for other measures for which z scores have a much smaller range.

Agreement among Years per Model

Table 4 also shows limited agreement among annual rankings per model, ranging from absent (DM-related admissions) to fair (statins, COPD-related admissions) to moderate (all other measures; see the online appendix for results for the other measures). Agreement on outlier designation is higher than agreement overall but still fairly low. No model consistently produces more or less stable results than other models.

Table 2 Descriptive statistics of the study population, by year

	2006	2007	2008
All members, independent variables	*n* = 2,809,250	*n* = 2,802,632	*n* = 2,808,838
Age, mean (SD)	40.1 (23.2)	40.2 (23.2)	40.4 (23.3)
Male, %	50.5	50.4	50.3
Living in a deprived area, %	6.7	6.5	6.5
Monthly income, mean (SD)[a]	5.3 (2.9)	5.3 (2.9)	5.3 (2.8)
Educational level, mean (SD)[b]	2.0 (0.8)	2.0 (0.8)	2.0 (0.8)
Ethnicity, mean (SD)[c]	2.1 (1.5)	1.3 (0.8)	1.3 (0.8)
Urbanization, mean (SD)[d]	2.9 (1.4)	4.7 (2.2)	4.7 (2.1)
Died, %	0.9	0.8	0.8
In a pharmacy-based cost group, %	16.2	16.9	17.5
In ≥2 pharmacy-based cost groups, %	4.5	4.8	5.1
In ≥3 pharmacy-based cost groups, %	1.2	1.3	1.4
In a diagnosis cost group, %	2.6	2.3	2.4
All members, dependent variables	*n* = 2,809,250	*n* = 2,802,632	*n* = 2,808,838
Inpatient admissions, mean (SD)	0.10 (0.46)	0.11 (0.49)	0.10 (0.46)
No inpatient admission, %	92.7	92.5	92.8
Outpatient visits, mean (SD)	0.52 (1.43)	0.53 (1.52)	0.53 (1.21)
No outpatient visit, %	72.4	73.0	74.1
GP expenses, mean (SD)	119 (112)	128 (122)	127 (119)
No GP expenses, %	2.1	2.1	2.2
Medication expenses, mean (SD)	275 (908)	310 (1024)	302 (1048)
No medication expenses, %	31.9	31.1	29.2
Total expenses, mean (SD)	1476 (5306)	1531 (5359)	1485 (4879)
No expenses, %	1.6	1.6	1.6
Members with DM, dependent variables	*n* = 86,208	*n* = 88,536	*n* = 89,320
On statins, %	59.4	63.0	63.6
Inpatient admissions, mean (SD)	0.08 (0.38)	0.08 (0.39)	0.07 (0.36)
No inpatient admission, %	94.0	94.1	94.9
Members with COPD, dependent variables	*n* = 65,315	*n* = 68,927	*n* = 69,892
Receiving physiotherapy, %	4.17	4.76	5.55
On prednisone, %	33.4	32.6	32.6
On bronchodilators, %	81.5	80.5	80.2
Inpatient admissions, mean (SD)	0.07 (0.36)	0.08 (0.37)	0.07 (0.36)
No inpatient admission, %	94.6	94.3	94.6
General practitioners	*n* = 7471	*n* = 5447	*n* = 5538
≥100 patients in all years, *n* (mean sample size)	4396 (529)	4396 (533)	4396 (529)
≥100 patients in all years and in an HC, *n* (mean sample size)	355 (688)	355 (692)	355 (653)
≥30 DM patients in all years, *n* (mean sample size)	628 (70)	628 (71)	628 (72)
≥30 DM patients in all years and in an HC, *n* (mean sample size)	79 (68)	79 (70)	79 (72)
≥30 COPD patients in all years, *n* (mean sample size)	517 (58)	517 (60)	517 (61)
≥30 COPD patients in all years and in an HC, *n* (mean sample size)	66 (55)	66 (59)	66 (59)
Health centers	*n* = 142	*n* = 179	*n* = 186
≥100 patients in all years, *n* (mean sample size)	120 (1791)	120 (1874)	120 (1841)
≥30 DM patients in all years, *n* (mean sample size)	45 (131)	45 (141)	45 (141)
≥30 COPD patients in all years, *n* (mean sample size)	35 (117)	35 (130)	35 (130)

Note: COPD = chronic obstructive pulmonary disease; DM = diabetes mellitus; GP = general practitioner; HC = health center; SD = standard deviation.
a. Ten categories (1 = lowest income decile, 10 = highest income decile).
b. Three categories (1 = low, 3 = high).
c. Six categories in 2006 (1–6) and 5 categories in 2007–2008 (0–4). A high score corresponds to a high percentage of non-Western immigrants living in the member's ZIP-code area.
d. Five categories in 2006 (1–5), 8 categories in 2007–2008 (1–8). A high score corresponds to a low level of urbanization of the member's living area.

Table 3 Selected fit statistics of included models, by performance measure, 2008

Measure (Applicable Population)	Model	R^{2a}	MAD	Calibration[b]
Yes/no physiotherapy (COPD patients)	OLS	0.037	0.101	$Y = \hat{Y}$
	GLM binomial-probit	0.039	0.100	$Y = 1.000\hat{Y}$
	MLM normal-id (HCs)	0.042	0.099	$Y = \hat{Y}$
	MLM normal-id (GPs)	0.063	0.099	$Y = \hat{Y}$
Yes/no prednisone (COPD patients)	OLS	0.061	0.411	$Y = \hat{Y}$
	GLM binomial-logit	0.061	0.411	$Y = \hat{Y}$
	MLM normal-id (GPs)	0.085	0.411	$Y = \hat{Y}$
Yes/no bronchodilators (COPD patients)	OLS	0.026	0.312	$Y = \hat{Y}$
	GLM binomial-logit	0.028	0.310	$Y = \hat{Y}$
	MLM normal-id (GPs)	0.048	0.293	$Y = \hat{Y}$
Yes/no statins (DM patients)	OLS	0.701	0.136	$Y = \hat{Y}$
	GLM binomial-logit	0.707	0.136	$Y = \hat{Y}$
	MLM normal-id (GPs)	0.713	0.134	$Y = \hat{Y}$
No. of inpatient admissions (COPD patients)	OLS	0.114	0.124	$Y = \hat{Y}$
	GLM normal-log	0.109	0.126	$Y = 0.946\hat{Y}$
	GLM Poisson-power	0.113	0.123	$Y = \hat{Y}$
	GLM negbin-power	0.111	0.123	$Y = 0.960\hat{Y}$
	GLM gamma-log	0.115	0.124	$Y = 1.000\hat{Y}$
	MLM normal-power (HCs)	0.118	0.123	$Y = 1.024\hat{Y}$
	2-part logit-OLS	0.105	0.123	$Y = \hat{Y}$
	2-part logit-normal power	0.105	0.123	$Y = 1.000\hat{Y}$
	2-part logit-Poisson power	0.105	0.123	$Y = 1.000\hat{Y}$
No. of inpatient admissions (DM patients)	OLS	0.070	0.125	$Y = \hat{Y}$
	GLM normal-log	0.077	0.125	$Y = 0.992\hat{Y}$
	GLM Poisson-power	0.072	0.124	$Y = \hat{Y}$
	GLM negbin-power	0.071	0.124	$Y = 0.971\hat{Y}$
	GLM gamma-log	0.071	0.124	$Y = 1.001\hat{Y}$
	MLM normal-power (HCs)	0.073	0.124	$Y = 1.013\hat{Y}$
	2-part logit-OLS	0.069	0.124	$Y = \hat{Y}$
	2-part logit-normal log	0.069	0.124	$Y = 1.001\hat{Y}$
	2-part logit-Poisson power	0.068	0.124	$Y = 1.000\hat{Y}$
No. of inpatient admissions (all members)	OLS	0.109	0.166	$Y = \hat{Y}$
	GLM Poisson-power	0.104	0.166	$Y = \hat{Y}$
	GLM negbin-power	0.101	0.166	$Y = 0.965\hat{Y}$
	Gamma power	0.108	0.166	$Y = 1.000\hat{Y}$
	MLM normal-id (GPs)	0.109	0.166	$Y = \hat{Y}$
	2-part logit-OLS	0.103	0.167	$Y = \hat{Y}$
	2-part logit-normal power	0.103	0.167	$Y = 1.000\hat{Y}$
	2-part logit-Poisson power	0.102	0.167	$Y = 1.000\hat{Y}$
No. of outpatient visits (all members)	OLS	0.366	0.477	$Y = \hat{Y}$
	GLM Poisson-power	0.365	0.477	$Y = \hat{Y}$
	GLM negbin-id	0.363	0.476	$Y = 0.998\hat{Y}$
	GLM gamma-id	0.363	0.476	$Y = 1.000\hat{Y}$
	MLM normal-id (GPs)	0.369	0.474	$Y = \hat{Y}$
	2-part logit-OLS	0.365	0.478	$Y = \hat{Y}$
	2-part logit-normal power	0.367	0.479	$Y = 1.002\hat{Y}$
	2-part logit-Poisson power	0.366	0.478	$Y = 1.000\hat{Y}$
GP expenses (all members)[c]	OLS	0.288	52.326	$Y = \hat{Y}$
	GLM normal-power	0.288	52.439	$Y = 0.999\hat{Y}$
	MLM normal-id (HCs)	0.290	52.121	$Y = \hat{Y}$
	MLM normal-id (GPs)	0.318	50.701	$Y = \hat{Y}$
	MLM gamma-power (GPs)	0.319	50.442	$Y = 0.993\hat{Y}$

(continued)

Table 3 (continued)

Measure (Applicable Population)	Model	R^{2a}	MAD	Calibration[b]
Medication expenses (all members)[d]	OLS	0.375	222.676	$Y = \hat{Y}$
	GLM gamma-power	0.335	224.940	$Y = 0.929\hat{Y}$
	MLM normal-id (GPs)	0.376	222.966	$Y = \hat{Y}$
	2-part probit-OLS	0.375	222.165	$Y = 1.000\hat{Y}$
	2-part probit-normal power	0.362	237.793	$Y = 0.957\hat{Y}$
	2-part probit-gamma power	0.350	228.838	$Y = 0.958\hat{Y}$
Total expenses (all members)[e]	OLS	0.226	1452.191	$Y = \hat{Y}$
	MLM normal-id (HCs)	0.226	1449.624	$Y = \hat{Y}$
	MLM normal-id (GPs)	0.226	1450.547	$Y = \hat{Y}$

Note: COPD = chronic obstructive pulmonary disease; DM = diabetes mellitus; GP = general practitioner; HC = health center; MAD = mean absolute deviation; GLM = generalized linear model; MLM = multilevel model; negbin = negative binomial; OLS = ordinary least squares.
a. Percentage explained variation at the individual member level.
b. Extent to which the mean of expected values ($\hat{Y}$) approximates the mean of observed values (Y).
c. Expenses generated by GPs through office visits, home visits, and (diagnostic) tests.
d. Expenses related to the use of prescription medication, regardless of prescriber.
e. Sum of GP expenses, medication expenses, and inpatient expenses generated by medical specialists.

Our hypothesis that results would be less stable for (GPs in) HCs than for GPs not in an HC is not confirmed: there appears to be no relationship between type of provider and the stability of profiling results. Limiting the analysis to providers with $\geq$100 patients for disease-specific variables and $\geq$1000 patients for non–disease-specific variables did not change these results, although for some measures, agreement increased by up to 15 percentage points (data not shown).

DISCUSSION

This study has investigated the influence of the statistical model on performance-profiling results for primary care providers. Our main goal was to determine whether different statistical methods selected based on statistical as well as relevant practical criteria (from a purchaser's perspective) generate different profiling results. Our results showed that profiling results are sensitive to the statistical model that is used and that the choice of model does indeed seem to matter, especially for clinical outcome measures and expenses.

However, differences were relatively small, and the choice of model may not be as important as other choices such as the set of risk adjusters, definition of performance index, and method for categorizing provider performance.[5,15-22,24-27] In addition, simple methods have important practical advantages. For example, OLS can be applied to all measures and data, a feature not shared by many other models that may work well for 1 year and fail to converge in the next. For purchasers, these might be sufficient reasons to choose OLS (or a logit model for binary variables). Nonetheless, caution is clearly warranted. Agreement of 75% to 95% suggests that the models still relatively often classify providers in different performance categories, which, depending on the purpose for which the rankings are used (e.g., performance feedback, pay for performance, public reporting), may have far-reaching (financial) consequences for providers. In addition, compared with agreement overall, agreement on outlier designation was lower and more variable. For example, for non–disease-specific measures and using 10% cutoffs for both tails to determine outliers status, even 5% disagreement means that the choice of model alone determines for 44 GPs (4396 GPs $\times$ 20% $\times$ 5%) whether they will be classified as outlier or not, which may be hard to justify. Thus, for each individual measure selected for profiling, decision makers are faced with a difficult tradeoff between identifying the best-fitting model each year (a cumbersome task) and simply using a well-known method that is easy to apply, maintain, and explain but that may also result in somewhat different provider classifications.

The first option will be time-consuming and can be expensive, especially when providers are profiled on many measures and if the modeling is outsourced to an external (commercial) party. In addition, it may result in mixed signals toward providers. For example, it may be confusing for providers if the purchaser tries to convince them about a new sophisticated

Table 4 Average agreement among models (2008) and per model among years for HCs, GPs not in an HC, and GPs in an HC, by selected performance measure

		Per Year among Models (2008)									Per Model among Years (2006–2008)								
		HCs			GPs, no HC			GPs, yes HC			HCs			GPs, no HC			GPs, yes HC		
Measure	Model[a]	All[b]	Low[c]	High[d]	All[b]	Low[c]	High[d]	All[b]	Low[c]	High[d]	All[b]	Low[c]	High[d]	All[b]	Low[c]	High[d]	All[b]	Low[c]	High[d]
Bronchodilators	OLS	0.96	1.0	1.0	0.97	0.98	0.97	0.96	0.86	0.93	0.52	0.50	0.42	0.49	0.54	0.34	0.49	0.67	0.48
	GLM binomial-logit	0.96	1.0	1.0	0.96	0.98	0.96	0.95	0.86	0.86	0.52	0.50	0.58	0.49	0.55	0.35	0.49	0.57	0.57
	MLM normal-id	—	—	—	0.96	0.98	0.97	0.96	0.86	0.93	—	—	—	0.49	0.54	0.34	0.49	0.57	0.52
Statins	OLS	0.97	1.0	1.0	0.98	0.97	0.96	0.98	1.0	1.0	0.20	0.33	0.40	0.31	0.36	0.30	0.27	0.25	0.29
	GLM binomial-logit	0.97	1.0	1.0	0.97	0.95	0.96	0.97	1.0	1.0	0.20	0.33	0.40	0.31	0.35	0.29	0.27	0.25	0.29
	MLM normal-id	—	—	—	0.97	0.97	0.96	0.97	1.0	1.0	—	—	—	0.31	0.34	0.27	0.27	0.25	0.29
COPD-related admissions	OLS	0.87	0.85	0.50	0.92	0.90	0.85	0.92	0.93	0.79	0.24	0.67	0.10	0.04	0.18	0.11	0.15	0.33	0.29
	GLM normal-log	0.84	0.80	0.60	0.86	0.89	0.77	0.86	0.82	0.61	0.37	0.42	0.17	0.04	0.18	0.12	0.16	0.24	0.14
	GLM Poisson-power	0.87	0.85	0.70	0.92	0.92	0.85	0.93	0.93	0.75	0.26	0.67	0.17	0.04	0.17	0.11	0.16	0.33	0.19
	GLM negbin-power	0.86	0.75	0.70	0.90	0.83	0.77	0.90	0.82	0.71	0.27	0.58	0.10	0.04	0.16	0.10	0.16	0.10	0.19
	GLM gamma-log	0.87	0.85	0.70	0.92	0.92	0.86	0.93	0.93	0.79	0.25	0.67	0.17	0.04	0.18	0.09	0.16	0.33	0.29
	MLM normal-id	0.82	0.80	0.60	—	—	—	—	—	—	0.20	0.42	0.17	—	—	—	—	—	—
DM-related admissions	OLS	0.90	0.84	0.80	0.92	0.95	0.88	0.91	0.97	0.84	–0.03	0.07	0.20	0.04	0.12	0.12	0.03	0.08	0.21
	GLM normal-log	0.84	0.84	0.68	0.83	0.88	0.78	0.82	0.88	0.66	–0.01	0.07	0.20	0.04	0.13	0.12	0.01	0.08	0.21
	GLM Poisson-power	0.89	0.80	0.80	0.92	0.95	0.87	0.91	0.97	0.84	–0.03	0.07	0.13	0.03	0.12	0.10	0.02	0.08	0.21
	GLM negbin-power	0.87	0.84	0.72	0.91	0.90	0.83	0.89	0.97	0.75	–0.05	0.07	0.07	0.04	0.14	0.13	0.00	0.08	0.13
	GLM gamma-log	0.90	0.84	0.80	0.93	0.95	0.88	0.92	0.97	0.84	–0.03	0.07	0.20	0.04	0.12	0.11	0.02	0.08	0.21
	MLM normal-id	0.84	0.80	0.68	—	—	—	—	—	—	–0.00	0.07	0.13	—	—	—	—	—	—
GP expenses[e]	OLS	0.85	0.75	0.83	0.92	0.93	0.88	0.93	0.92	0.90	0.51	0.64	0.36	0.54	0.64	0.54	0.51	0.57	0.44
	GLM normal-power	0.74	0.58	0.88	0.88	0.88	0.81	0.89	0.83	0.84	0.47	0.44	0.42	0.53	0.60	0.51	0.50	0.54	0.43
	MLM normal-id	0.85	0.75	0.88	0.92	0.92	0.87	0.93	0.92	0.88	0.52	0.64	0.42	0.53	0.63	0.53	0.51	0.58	0.42
	MLM gamma-power	—	—	—	0.90	0.91	0.84	0.92	0.85	0.84	—	—	—	0.53	0.64	0.50	0.51	0.63	0.46
Medication expenses[f]	OLS	0.77	1.0	0.67	0.90	0.92	0.87	0.91	0.89	0.86	0.47	0.78	0.33	0.47	0.52	0.49	0.47	0.57	0.38
	GLM gamma-power	0.77	1.0	0.67	0.82	0.86	0.76	0.83	0.81	0.76	0.55	0.72	0.36	0.46	0.50	0.52	0.47	0.52	0.39
	MLM normal-id	—	—	—	0.90	0.93	0.87	0.90	0.89	0.88	—	—	—	0.48	0.53	0.50	0.48	0.58	0.38

Note: A dash (—) indicates that the relevant model did not converge or that the covariance parameter for the random provider intercept was not significant. COPD = chronic obstructive pulmonary disease; DM = diabetes mellitus; GP = general practitioner; OLS = ordinary least squares; GLM = generalized linear model; MLM = multilevel model; negbin = negative binomial; HC = health center; GP = general practitioner.

a. Two-part models were excluded from these comparisons, for 3 reasons: expected values are very similar to those from the 1-part models (the correlation coefficient was typically >0.99), they do not appear to fit the data better, and calculation of standard errors is relatively complex.

b. Average pairwise agreement (weighted kappa) on overall rankings on *z* scores.

c. Average pairwise percentage agreement on classification of providers to the 10% providers with the worst performance based on *z* scores.

d. Average pairwise percentage agreement on classification of providers to the 10% providers with the best performance based on *z* scores.

e. Expenses generated by general practitioners through office visits, home visits, and performing (diagnostic) tests.

f. Expenses related to the use of prescription medication, regardless of prescriber.

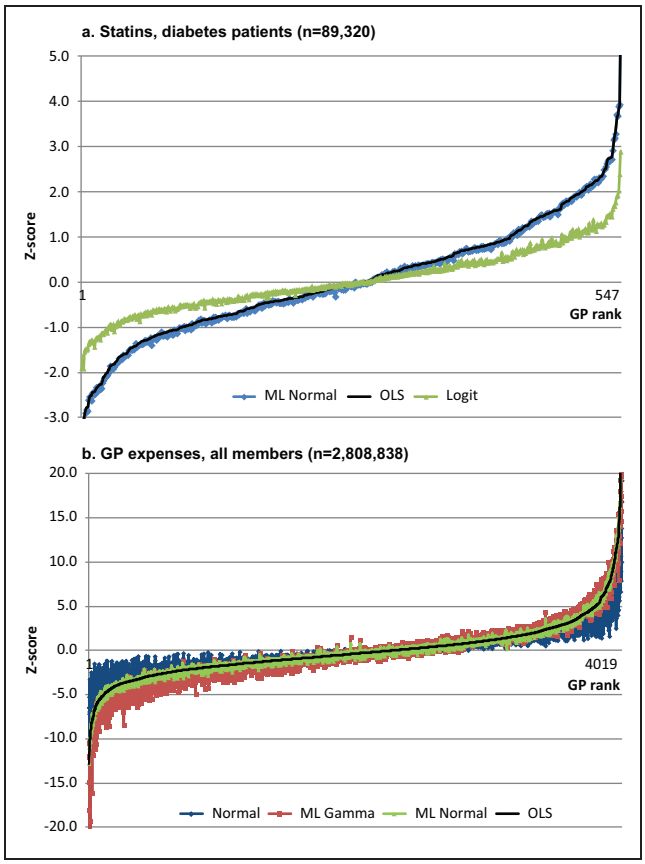

Figure 1 Distributions of z *scores for general practitioners (GPs; not in a health center) for 2 measures, 2008. ML = multilevel; OLS = ordinary least squares. The figure displays GPs' z scores produced by the different models for 2 different measures: the percentage of diabetes patients on statins (a process quality measure, 3 models, relevant for 547 GPs) and total GP expenses (a resource use measure, 4 models, relevant for 4019 GPs). GPs are ranked on their z scores (derived from OLS) from highest performance (rank 1) to lowest performance (rank 547 or 4019, depending on the measure). The figure illustrates that differences may be large for individual providers despite high agreement among models. In addition, as shown in panel a, highly similar rankings do not preclude large differences among models.*

method for a specific (type of) measure (''this specification is best suited to account for your specific patient mix for this measure!'') whereas in the year before they had just been convinced about the merits of another method for the same measure. A practical solution may be to use a simple and easy to apply, maintain, and explain model (e.g., OLS) and to compare the results with the results of a relatively simple 2-level MLM (e.g., assuming a normal distribution and identity link). In our data, such MLM specifications had little convergence problems and were often (but not always) among the models with the best fit statistics. In addition, as noted above, these models have advantages that may be appealing to providers, although these advantages may be difficult to explain.[51]

There were several other notable findings. First, models agreed more for processes than for outcomes and expenses. Yet we did observe differences for processes as well, and because even small differences can be important, the conclusion that the choice of model does not matter for processes cannot be justified. Second, models tended to agree more on designation of low outliers than of high outliers, especially for utilization and expenses. The explanation is that for high outliers, expected utilization and expenses are high compared with what is observed. Because generalized linear (mixed) models can better predict high expenses and utilization than OLS, agreement on high-outlier designation will be lower than on low-outlier designation. This is an important finding, as most pay-for-performance programs reward only high performance.[2] Third, agreement was higher for GPs than for HCs, especially for disease-related admissions and expenses. It thus seems that the choice of model matters more for HCs than for GPs. Fourth, profiling results varied substantially over time. As this is unlikely to be a result of a specific intervention, it probably reflects random variation and low reliability.[6,49] Results were most unstable for hospital admissions and total expenses, which is not surprising as these are more difficult to influence by providers than other types of measures such as processes.[52,53] Measures will be more reliable when sample size and the intraclass correlation coefficient (i.e., the proportion of total variation that can be attributed to between-provider variation) are large.[54–56] Limiting the analysis to providers with more patients increased agreement, but much variation remained, implying relatively low between-provider variances. Results were most stable for GP expenses. Given the large range in *z* scores (Figure 1b), this may be a particularly useful measure for profiling. However, in view of GPs' gatekeeping role, purchasers should then be cautious that GPs are not penalized for keeping patients out of the hospital and/or not rewarded for unwarranted referrals.

This study has limitations. First, we identified COPD patients using the pharmacy-based cost group for chronic nonspecific respiratory conditions and the patients' age ($\geq$45 years). As a result, we may have overestimated the number of COPD patients in our data. Second, outlier providers were arbitrarily defined. We chose a relative threshold for determining

outlier status because this is common in practice. However, an absolute threshold based on, for example, conventional levels for statistical testing may yield different results. A provider will then be an outlier when the absolute value of the z score is larger than 1.96 ($P = 0.05$) or 2.58 ($P = 0.01$). Using a large P value mitigates the risk of incorrectly classifying outliers as average but also increases the risk of classifying average providers as outliers. Absolute thresholds have the advantage that they are transparent and that they may provide stronger incentives for providers to improve.[57] Conversely, with a relative threshold, purchasers know exactly how many providers will be designated as outlier, which, when the profiling results are used for allocating incentive payments, provides budgetary certainty for the purchaser. The sensitivity of our results to the method of categorizing provider performance merits further study.

Third, we analyzed all inpatient admissions and outpatient visits (i.e., grouped together) and not, for example, only ambulatory care–sensitive ones,[58] which might have been the preferred approach in profiling primary care providers. Although separate analysis of different types of admissions/visits could have been valuable, admission/visit type could not be derived from our data. But even if ambulatory care–sensitive admissions could have been identified, they may well have been too rare to model reliably.[7]

A fourth limitation is that our set of risk adjusters was based on administrative information that is routinely available in insurers' files. This information was not generated for profiling purposes but for explaining variation in costs for the purpose of calculating risk-adjusted capitation payments for insurers. Ideally, risk adjustment for performance profiling would use detailed (clinical) information from medical records and patient surveys, especially regarding clinical quality. However, collecting such data on a routine basis is expensive, and we expect that in practice, insurers will often mainly use data already available in their files. In addition, although we would have preferred to use individual-level information on socioeconomic status, ZIP-code–level variables have been shown to discern broadly similar patterns compared with the corresponding individual-level variables.[59,60] As expected, our risk adjusters were best suited for expenses (and outpatient visits), whereas for other measures, 90% or more of the variation remained unexplained. Although a high R^2 indicates that risk adjustment is important, a low R^2 does not necessarily imply that risk adjustment is unimportant; the risk adjusters may simply not be adequate. In agreement with Ash and Ellis,[7] we would argue that risk adjustment is necessary for all measures until it is convincingly demonstrated that patient characteristics cannot predict these measures. Another concern is that some of our risk adjusters can be directly influenced by providers and thus provide opportunities for gaming when used for risk-adjusting performance profiles. For example, a GP could increase the number of defined daily doses such that more patients are classified in a pharmacy-based cost group, effectively making his or her patient population look sicker than it is in reality. Incentives for such behavior will be larger when results are used as input for high-stakes improvement efforts such as public reporting and pay for performance. Although we have no reason to believe our conclusions would be substantially different using better risk adjustment, development of tailored risk-adjustment methods merits high priority.

Finally, our results may not generalize to other settings and measures. We looked at a specific group of providers (Dutch GPs with fixed patient panels and acting as gatekeepers) using administrative data from 1 insurer. Given the widespread use of performance profiling, future research should investigate whether our results are confirmed in other settings for reliable and commonly used performance measures.

In summary, although simple methods such as OLS have advantages from a practical viewpoint, they may produce different profiling results compared with more suitable methods. Therefore, the choice of statistical model for performance profiling should be made with care, especially when results are used as input for high-stakes improvement efforts. In addition, regardless of the model, performance comparisons should preferably be conducted over multiple time periods to gain insight in the extent to which the measures are driven by chance and thus if they are potentially suitable for profiling. Even for process measures, over which providers supposedly have much control, random chance may determine providers' relative positions to a large extent, which, depending on how and by whom the profiles are used, can have far-reaching (financial) consequences for providers.

ACKNOWLEDGMENTS

We would like to thank Wynand van de Ven for comments on earlier drafts, participants of the iBMG seminar series for discussing preliminary results, the associate editor and 3 anonymous referees, and Geert Groenenboom (Achmea Zorg & Gezondheid) and Matthijs Hagenaars (Achmea Zorg & Gezondheid) for providing the data.

REFERENCES

1. van der Veer SN, De Keizer NF, Ravelli ACJ, Tenkink S, Jager KJ. Improving quality of care: a systematic review on how medical registries provide information feedback to health care providers. Int J Med Inform. 2010;79(5):305–23.

2. Eijkenaar F. Pay for performance in health care: an international overview of initiatives. Med Care Res Rev. 2012;69:251–76.

3. Fung CH, Lim YW, Mattke S, Damberg C, Shekelle PG. Systematic review: the evidence that publishing patient care performance data improves quality of care. Ann Intern Med. 2008;148(2):111–23.

4. Brennan TA, Spettell CM, Fernandes J, Downey RL, Carrara LM. Do managed care plans' tiered networks lead to inequities in care for minority patients? Health Aff. 2008;27(4):1160–6.

5. Adams JL, McGlynn E, Thomas JW, Mehrotra A. Incorporating statistical uncertainty in the use of physician cost profiles. BMC Health Serv Res. 2010;10:57.

6. Friedberg MW, Damberg CL. A five-point checklist to help performance reports incentivize improvement and effectively guide patients. Health Aff. 2012;31(3):612–8.

7. Ash AS, Ellis RP. Risk-adjusted payment and performance assessment for primary care. Med Care. 2012;50(8):643–53.

8. Chang RE, Lin S, Aron D. A P4P program in Taiwan improved care for some diabetes patients, but doctors may have excluded sicker ones. Health Aff. 2012;31:93–102.

9. Chen TT, Chung KP, Lin IC, Lai M. Unintended consequence of diabetes P4P program in Taiwan: are patients with more comorbidities or more severe conditions likely to be excluded from the P4P program? Health Serv Res. 2011;46:47–60.

10. Rosen A, Wu J, Chang BH, et al. Risk-adjustment for measuring health outcomes: an application in VA long-term care. Am J Med Qual. 2001;16(4):118–27.

11. Pope GC, Kautter J. Profiling efficiency and quality of physician organizations in Medicare. Health Care Financ Rev. 2007;29(1):31–43.

12. Tucker AM, Weiner JP, Honigfeld S, Parton RA. Profiling primary care physician resource use: examining the application of case mix adjustment. J Ambul Care Manage. 1996;19(1):60–80.

13. Iezzoni LI, ed. Risk-Adjustment for Measuring Health Care Outcomes. 3rd ed. Chicago: Health Administration Press; 2003.

14. Christianson JB, Leatherman S, Sutherland K. Lessons from evaluations of purchaser pay-for-performance programs: a review of the evidence. Med Care Res Rev. 2008;65(6):5S–35S.

15. Mukamel DB, Glance LG, Li Y, et al. Does risk-adjustment of the CMS quality measures for nursing homes matter? Med Care. 2008;46(5):532–41.

16. Mukamel DB, Brower CA. The influence of risk-adjustment methods on conclusions about quality of care in nursing homes based on outcome measures. Gerontologist. 1998;38(6):695–703.

17. Huang IC, Dominici F, Frangakis C, Diette GB, Damberg CL, Wu AW. Is risk-adjustor selection more important than statistical approach for provider profiling? Asthma as an example. Med Decis Making. 2005;25(1):20–34.

18. Thomas JW, Grazier KL, Ward K. Economic profiling of primary care physicians: consistency among risk-adjusted measures. Health Serv Res. 2004;39(4 pt 1):985–1003.

19. Rosen A, Loveland S, Rakovski C, Christiansen C, Berlowitz D. Do different case-mix measures affect assessments of provider efficiency? Lessons from the department of Veterans Affairs. J Ambul Care Manage. 2003;26:229–42.

20. Kang HC, Hong JS. Do differences in profiling criteria bias performance measurements? Economic profiling of medical clinics under the Korea national health insurance program: an observational study using claims data. BMC Health Serv Res. 2011; 11:189.

21. DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. Comparing risk-adjustment methods for provider profiling. Stat Med. 1997;16(23):2645–64.

22. Iezzoni LI, Ash AS, Shwartz M, Daley J, Hughes JS, Mackiernan YD. Judging hospitals by severity-adjusted mortality rates: the influence of the severity-adjustment method. Am J Public Health. 1996;86(10):1379–87.

23. Thomas JW, Ward K. Economic profiling of physician specialists: use of outlier treatment and episode attribution rules. Inquiry 2006;43(3):271–82.

24. Thomas JW. Economic profiling of physicians: does omission of pharmacy claims bias performance measurement? Am J Manage Care. 2006;12(6):341–51.

25. Rosen AK, Rakovski CC, Loveland SA, Anderson JJ, Berlowitz DR. Profiling resource use: do different outcomes affect assessments of provider efficiency? Am J Manage Care. 2002;8(12):1105–15.

26. Metfessel BA, Greene RA. A nonparametric statistical method that improves physician cost of care analysis. Health Serv. Res. 2012;47(6):2398–417.

27. Austin PC, Alter D, Anderson G, Tu JV. Impact of the choice of benchmark on the conclusions of hospital report cards. Am Heart J. 2004;148(6):1041–6.

28. Cowen ME, Strawderman RL. Quantifying the physician contribution to managed care pharmacy expenses: a random effects approach. Med Care. 2002;40(8):650–61.

29. Glance LG, Dick A, Osler TM, Li Y, Mukamel DB. Impact of changing the statistical methodology on hospital and surgeon ranking: the case of the New York State Cardiac Surgery Report Card. Medical Care. 2006;44(4):311–9.

30. van de Ven WPMM, Schut FT. Managed competition in the Netherlands: still work-in-progress. Health Econ. 2009;18(3):253–5.

31. Van der Lucht F, Verweij A. Etniciteit en zorggebruik. In: Volksgezondheid Toekomst Verkenning, Nationaal Kompas Volksgezondheid. Bilthoven, the Netherlands: RIVM; 2010. http://www.nationaalkompas.nl/bevolking/etniciteit/allochtonen-en-zorggebruik/.

32. Peeters B, van Tongelen I, Boussery K, Mehuys E, Remon JP, Willems S. Factors associated with medication adherence to oral hypoglycaemic agents in different ethnic groups suffering from type 2 diabetes: a systematic literature review and suggestions for further research. Diabet Med. 2011;28(3):262–75.

33. Bailey CJ, Kodack M. Patient adherence to medication requirements for therapy of type 2 diabetes. Int J Clin Pract. 2011;65(3): 314–22.

34. van de Ven WPMM, van Vliet RCJA, Lamers LM. Health-adjusted premium subsidies in the Netherlands. Health Aff. 2004;23(3):45–55.

35. Prinsze F, van Vliet RCJA. Health-based risk-adjustment: improving the pharmacy-based cost group model by adding diagnostic cost groups. Inquiry. 2007;44:469–80.

36. van Kleef RC, van Vliet RCJA. Prior use of durable medical equipment as a risk adjuster for health-based capitation. Inquiry. 2011;47(4):343–58.

37. Greenfield S, Kaplan SH, Kahn R, Ninomiya J, Griffith JL. Profiling care provided by different groups of physicians: effects of patient case-mix (bias) and physician-level clustering on quality assessment results. Ann Intern Med. 2002;136(2):111–21.

38. Jones AM. Health econometrics. Handbook of Health Economics. 2000;1:265–344.

39. Manning WG, Mullahy J. Estimating log models: to transform or not to transform? J Health Econ. 2001;20:461–94.

40. Deb P, Manning WG, Norton EC. Modeling health care costs and counts. IHEA World Congress on Health Economics, Toronto, Canada, July 2011.

41. Mihaylova B, Briggs A, O'Hagan A, Thompson SG. Review of statistical methods for analyzing healthcare resources and costs. Health Econ. 2011;20:897–916.

42. McCullagh P, Nelder JA. Generalized Linear Models. Boca Raton, FL: Chapman & Hall/CRC; 1989.

43. Goldstein H, Spiegelhalter DJ. League tables and their limitations: statistical issues in comparisons of institutional performance. J R Stat Soc A. 1996;190(pt 3):385–443.

44. Rice N, Jones A. Multilevel models and health economics. Health Econ. 1997;6:561–75.

45. Zhang H, Lu N, Feng C, et al. On fitting generalized linear mixed-effects models for binary responses using different statistical packages. Stat Med. 2011;30:2562–72.

46. Racz MJ, Sedransk J. Bayesian and frequentist methods for profiling using risk-adjusted assessments of medical outcomes. J Am Stat Assoc. 2010;105(489):48–58.

47. Jones HE, Spiegelhalter DJ. The identification of ''unusual'' health care providers from a hierarchical model. Am Statistician. 2011;65(2):154–63.

48. Pregibon D. Goodness of link tests for generalized linear models. Appl Stat. 1980;29:15–24.

49. Berlowitz DR, Anderson JJ, Ash AS, Brandeis GH, Brand HK, Moskowitz MA. Reducing random variation in reported rates of pressure ulcer development. Med Care. 1998;36(6):818–25.

50. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74.

51. Friedberg MW, Damberg CL. Methodological Considerations in Generating Provider Performance Scores for Use in Public Reporting: A Guide for Community Quality Collaboratives. AHRQ Publication No. 11-0093. Rockville, MD: Agency for Healthcare Research and Quality; 2011.

52. Hofer TP, Hayward RA, Greenfield S, Wagner EH, Kaplan SH, Manning WG. The unreliability of individual physician'' report cards'' for assessing the costs and quality of care of a chronic disease. JAMA. 1999;281(22):2098–105.

53. Krein SL, Hofer TP, Kerr EA, Hayward RA. Whom should we profile? Examining diabetes care practice variation among primary care providers, provider groups, and health care facilities. Health Serv Res. 2002;37(5):1159–80.

54. Nyweide DJ, Weeks WB, Gottlieb DJ, Casalino LP, Fisher ES. Relationship of primary care physicians' patient caseload with measurement of quality and cost performance. JAMA. 2009; 302(22):2444–50.

55. Scholle SH, Roski J, Adams JL, et al. Benchmarking physician performance: reliability of individual and composite measures. Am J Manage Care. 2008;14(12):833–8.

56. Adams J, Mehrotra A, Thomas J, McGlynn E. Physician cost profiling—reliability and risk of misclassification. N Engl J Med. 2010;362:1014–21.

57. Eijkenaar F. Key issues in the design of pay for performance programs. Eur J Health Econ. 2013;14(1):117–31.

58. Agency for Healthcare Research and Quality. AHRQ Quality Indicators: Guide to Prevention Quality Indicators: Hospital Admission for Ambulatory Care Sensitive Conditions. Publication No. 02-R0203. Rockville, MD: Agency for Healthcare Research and Quality; 2001.

59. Zaslavsky AM, Epstein AM. How patients' sociodemographic characteristics affect comparisons of competing health plans in California on HEDIS quality measures. Int J Qual Health Care. 2005;17(1):67–74.

60. Krieger N. Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures—the public health disparities geocoding project. Am J Public Health. 2003;93:1655–71.