BMJ Open Quality

Healthcare provider profiling: fixing observation period or fixing sample size?

Werner Vach ,[1,2] Sonja Wehberg,[3] Bernhard Güntert,[4] Marcel Jakob,[5,6] George Luta[7,8]

To cite: Vach W, Wehberg S, Güntert B, *et al*. Healthcare provider profiling: fixing observation period or fixing sample size? *BMJ Open Quality* 2022;**11**:e001588. doi:10.1136/bmjoq-2021-001588

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjoq-2021-001588).

Received 17 June 2021
Accepted 8 March 2022

[1]Basel Academy for Quality and Research in Medicine, Basel, Switzerland
[2]Department of Environmental Sciences, University of Basel, Basel, Switzerland
[3]Research Unit of General Practice, University of Southern Denmark, Odense, Denmark
[4]Private University in the Principality of Liechtenstein, Triesen, Liechtenstein
[5]Medical Faculty, University of Basel, Basel, Switzerland
[6]Crossklinik, Basel, Switzerland
[7]Department of Biostatistics, Bioinformatics and Biomathematics, Georgetown University, Washington, DC, USA
[8]The Parker Institute, Copenhagen University Hospital, Copenhagen, Denmark

Correspondence to
Dr Werner Vach;
werner.vach@basel-academy.ch

BACKGROUND

Reporting information on the quality of healthcare providers is a popular strategy attempting to improve the overall healthcare quality at the national level.[1 2] The basic idea is to force providers to compare against each other and to stimulate poorly performing providers to improve their quality. In addition, patients may get the opportunity to select a well-performing provider when looking for a specific treatment.

A variety of quality indicators have been established over the last decades, and many countries have established a nationwide assessment of an indicator set.[3–10] Publishing an annual report listing the results of each provider from the previous year is a common practice, and the listings may be available to the providers, a central body or even to the public (eg, ref 11–13). Using a fixed observation period of 1 year implies that the number of patients per provider (ie, the annual volume) is varying across the providers. An alternative approach would be to include the same fixed number of patients from each provider, going back in time as long as necessary for each provider. In other words, the sample size is fixed instead of the observation period. This alternative approach seems feasible today, as many countries have collected data on the quality indicators in a rather stable fashion over the last years. It is the purpose of this paper to discuss potential advantages and challenges associated with this alternative approach.

We start with a look at the current practice of presenting profile data in annual reports and of identifying providers with poor or good performance. We then present three advantages of the fixed sample size approach and five challenges in implementing such an approach. Finally, we discuss the perspective to start supplementing the current reporting practice with reports based on fixed sample sizes.

A short outline of the current methodology for provider profiling

Considering the case of a binary quality indicator (eg, 30-day mortality rate, readmission rate, wound infection rate, etc) and ignoring the need for case-mix adjustment, the data used for provider profiling simply consist of the observed relative frequencies $\hat{p}_j$ and the volume v_j of each provider j. To judge and compare the providers, the values $\hat{p}_j$ are inspected and compared with the overall level $\bar{p}$, that is, the relative frequency over all providers. It is widely accepted that the stochastic imprecision of each relative frequency should also be taken into account. This reflects the desire to base statements and comparisons on the true underlying probability p_j. This probability reflects the quality future patients can expect if the provider j continues to manage patients at the current quality level. Two approaches to visualise the uncertainty are popular and illustrated in figure 1: (1) Each estimate $\hat{p}_j$ is surrounded by an α-CI covering the true value p_j with probability α. (2) In a funnel plot,[14 15] the estimates are contrasted with so-called control limits, such that the estimates should be within the control limits with probability α, if the true value p_j is identical to the overall level $\bar{p}$. Popular choices for α are 95% or 99.8%.

However, both ways of visualisation do not directly identify any provider as poorly or well performing. This can be approached by applying the corresponding rules. For example, a statistically significant deviation from the overall level may be required, that is, the CI does not cover the overall level, or the estimate $\hat{p}_j$ is outside of the control limits. It is also possible to define a threshold Δ^* for the true deviation from the overall level $\Delta_j = p_j - \bar{p}$ and to aim at identifying the providers above (or below) this threshold. Such a threshold should reflect that even under ideal circumstances some variation in the true

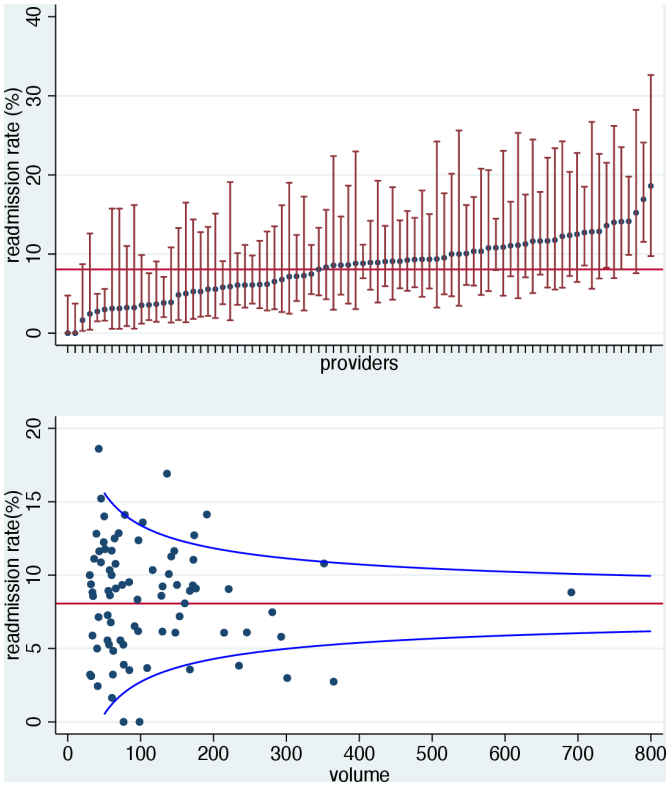

Figure 1 Presenting the results from a hypothetical provider profiling analysis. Upper half: relative frequencies with 95% CIs. The horizontal line refers to the overall level. The providers are ordered by the relative frequency. Lower half: funnel plot—relative frequencies and 95% control limits. The providers are ordered by the volume. The horizontal line refers again to the overall level.

values p_j is acceptable, for example, due to staff fluctuations and corresponding learning curve effects.

It has been recommended[16 17] to also take the overall variation of the estimated values $\hat{p}_j$ into account. If the overall variation is close to being explainable by random fluctuation (ie, most estimates are within the control limits), we may hesitate to call any provider a poorly performing provider. If the overall variation is high, we have good reasons to call at least some providers poorly performers. Formally, this idea is typically approached by considering the so-called posterior distribution of p_j. This distribution reflects the knowledge about p_j given the observed relative frequency $\hat{p}_j$, the volume v_j and the overall variation. This distribution is located closer to $\bar{p}$ in the case of a low overall variation than in the case of a high overall variation. The degree of shrinkage towards $\bar{p}$ depends on the volume v_j. The smaller the volume, the larger is the degree of shrinkage, reflecting the limited knowledge about p_j. The posterior distribution (of p_j or Δ_j) can be computed analytically (or at least approximated) and can serve as the basis for alternative rules. Popular choices are to compare the posterior mean of Δ_j with the threshold Δ^*, to require a certain posterior probability of Δ_j to be above 0 or to require a certain posterior probability of Δ_j to be above the threshold.[16 18 19] It is also possible to go one step further and

Table 1 The simulation scenario considered in this paper

- Eighty providers.

v_j takes the value	40	80	160	320	640
With probability	0.35	0.30	0.20	0.10	0.05

- p_j is drawn from a normal distribution with mean 8% and SD 3%. This implies that two-thirds of the providers have a true value p_j between 5% and 11%; p_j and v_j are drawn independently.
- The threshold Δ^* is set to 2 percentage points.

Simulations are always based on 10 000 repetitions. Details about the statistical computations are provided in the online supplemental material.

to consider the posterior distribution of the true rank R_j of provider j among all providers based on the true values p_j.[20 21]

Advantage 1: no dependence on volume

As mentioned above, it is desirable that classifying providers as poorly performing should depend on the true value p_j. Two providers with the same value of p_j should have equal probability to be labelled as a poor performer. Unfortunately, this is not the case for the commonly used rules. The probability to be labelled as a poor performer depends on the volume of the provider. Such probabilities can explicitly be determined if assumptions are made about the distributions of p_j and v_j across the providers. Table 1 specifies such a scenario, and the corresponding probabilities are shown in figure 2. Poorly performing providers with a true value of Δ_j above the threshold tend to have an increasing probability to be marked as poorly performing with increasing volume. The variation can be quite substantial. For example, when using as a rule a significant deviation from 0 (SIG), a provider with a true value Δ_j of about 3% has a probability of 13% to be labelled as a poorly performing provider in case of a volume of 40 patients, but a probability of 68% in case of a volume of 640 patients. When using the rule of a posterior mean being above the threshold (PMAT), the probabilities are 29% and 81%, respectively.

Such a variation can be regarded as unfair—at least from the perspective of high-volume providers. It can be also seen as unfair from a patient perspective: why should a patient be at risk to overlook that her or his personal provider is a poor performer, just because it is a low-volume provider? From a societal perspective the situation is less clear: detecting (and removing) poor quality in high-volume providers has a higher impact than in low-volume providers, as more patients would benefit.

Advantage 2: simplified presentation and interpretation of results

Presenting and interpreting results simplifies in the case of using a fixed sample size. The upper half of figure 3 illustrates this point by replicating the upper half of figure 1 in the case of equal sample sizes for all providers. Considering the estimates themselves or the lower or upper boundary of the CIs always gives the same ordering

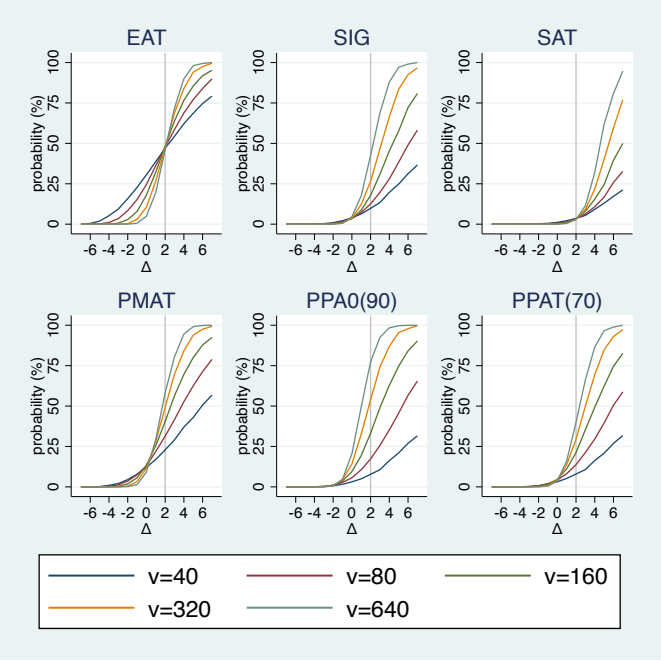

Figure 2 The probability to be labelled as a poor performer in relation to the true value of Δ_j and the volume v_j for the scenario defined in table 1. Six rules to identify poorly performing providers are considered. EAT: estimate Δ_j is above the threshold; SIG: estimate Δ_j is significantly above 0; SAT: estimate Δ_j is significantly above the threshold; PMAT: posterior mean of Δ_j is above the threshold; PPA0(90): posterior probability of Δ_j to be above the threshold is above 90%; PPAT(70): posterior probability of Δ_j to be above the threshold is above 70%. The threshold Δ^* is shown as a vertical line.

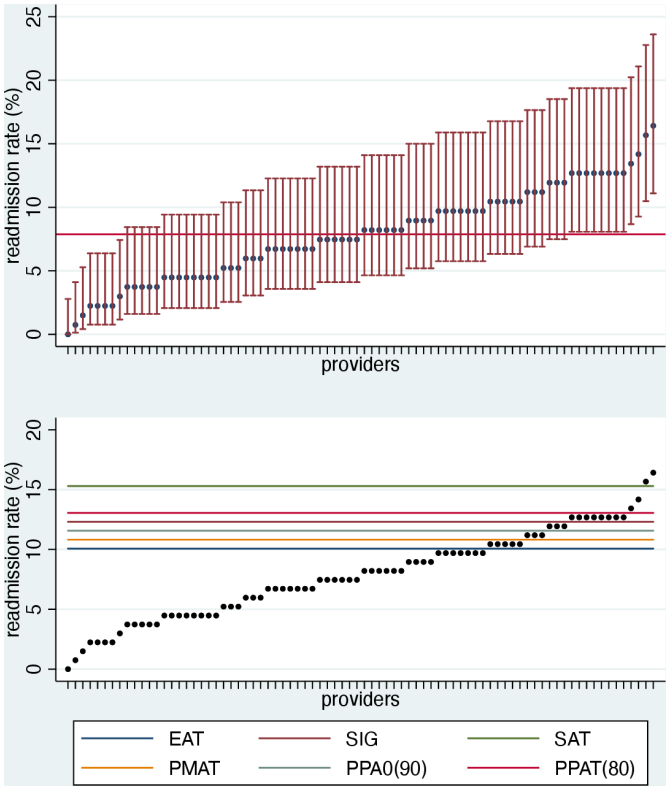

Figure 3 Presenting the results from a hypothetical provider profiling analysis with fixed sample size. Upper half: relative frequencies with 95% CIs. The horizontal line refers to the overall level. Lower half: relative frequencies with reference lines. Each line refers to a rule to identify poorly performing providers. Providers above the line are poorly performing according to the rules. The following rules are considered for a threshold of 2 percentage points. EAT: estimate Δ_j is above the threshold; SIG: estimate Δ_j is significantly above 0; SAT: estimate Δ_j is significantly above the threshold; PMAT: posterior mean of Δ_j is above the threshold; PPA0(90): posterior probability of Δ_j to be above 0 is above 90%; PPAT(80): posterior probability of Δ_j to be above the threshold is above 80%. The providers are ordered by the relative frequency in both plots.

of the providers. This reflects the simple fact that if all providers contribute with the same sample size, there is only one piece of information about the true underlying value p_j, namely the relative frequency $\hat{p}_j$ itself. Hence, any reasonable rule to order the providers will give the same ordering. Consequently, there remains little doubt about which provider has the worst *observed* quality.

There still remains the question which providers should be marked as poor performers, and the rules mentioned above can still be applied. However, there is the advantage that any rule results into a horizontal line, as illustrated in the lower half of figure 3.

Advantage 3: better decisions

Various types of decisions can be made based on provider profiling. Two examples are considered: (1) identifying the best local provider; (2) identifying poorly performing providers above the threshold. For each example, we consider various decision rules and compare their performance between using a fixed observation period and a fixed sample size. For the first, we consider again the scenario described in table 1; for the second, we fix the sample size to 134, such that the overall number of patients included in an annual analysis is identical.

If patients use provider profiling to select the best provider, they often focus on a preselection of local providers. Hence, we consider the decision to identify the best out of five randomly chosen providers. As performance criterion for the decision rule we consider the probability to identify the best local provider and the probability to identify a provider with a performance p_j at most 1 percentage point above the best local provider. According to table 2, a fixed sample size always implies a higher probability. Moreover, all three decision rules lead to identical decisions in the fixed sample size case, and hence perform identically.

Regulators may be interested in identifying all poorly performing providers with a value Δ_j above the threshold, for example, if they want to invite poorly performing providers for a review of their quality management. Here, a decision rule has to generate a list of providers. The performance of such a rule can be assessed by its sensitivity and specificity, that is, the probability of a provider above the threshold to be included in the list and the

Table 2 Performance of decision rules to identify the best local provider

Decision rule	Probability to identify the best local provider		Probability to come close to the best local provider	
Selecting the local provider with the lowest:	Fixed observation period	Fixed sample size	Fixed observation period	Fixed sample size
Estimate $\hat{p}_j$	0.61	0.70	0.73	0.81
Upper bound of 95% CI for p_j	0.55	0.70	0.66	0.81
Posterior mean of p_j	0.58	0.70	0.72	0.81

'Coming close to the best local provider' is defined as selecting a provider with a true value p_j at most 1 percentage point above the best local provider.

probability of a provider below the threshold not to be included, respectively. According to figure 4, using a fixed sample size moves the point given by sensitivity and specificity closer to the optimal value (1.0, 1.0) in the right upper corner, independent of the decision rule used. It depends, however, on the decision rule, whether mainly sensitivity or mainly specificity is improving.

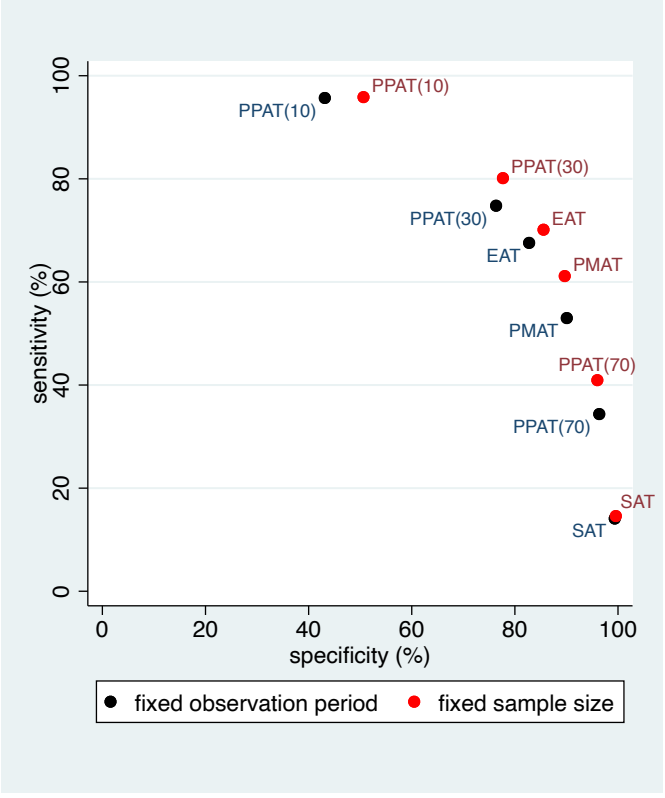

Figure 4 Sensitivity and specificity of different rules to identify the providers with a true value Δ_j above the threshold. The following identification rules are considered. EAT: estimate Δ_j is above the threshold; SAT: estimate Δ_j is significantly above the threshold; PMAT: posterior mean of Δ_j is above the threshold; PPAT(x): posterior probability of Δ_j to be above the threshold is above x%. PPAT(50) is not considered as this rule is identical to PMAT.

Challenge 1: scheduling of analyses

When fixing the observation period to 1 year, an obvious choice for scheduling analyses is an annual scheduling. When fixing the sample size, there is no natural choice for the scheduling such as 'after the next 100 patients', as this is reached for each provider at different time points. Consequently, there is a need for criteria to decide when a fixed sample size analysis should be performed and which sample size should be used.

A starting point may be to stick to annual reporting and aim at including the same overall number of patients as before—as we did in our considerations above when discussing advantage 3. This would imply observation periods shorter than 1 year for high-volume providers and observation periods longer than 1 year for low-volume providers (cf scenario II in figure 5). However, annual reporting implies now to use only a part of the patients available within each year from high-volume providers, that is, to throw away information. Consequently, it would be natural to schedule the analyses more frequently to ensure that each patient contributes at least once to the analysis (scenario III). Then providers with a very high volume and hence very short observation periods will imply very frequent analyses. To avoid this, prolonged observation periods have to be allowed for these providers (scenario IV). On the other hand, providers with a very low volume will have very long observation periods, and the connection to the actual situation for these providers may be lost. Consequently, a maximal value for the observation period may also be set, and some providers may be included with a smaller sample size (scenario V). A realistic choice may be to aim at three analyses per year and a maximal observation period of 3 years, implying that many providers would be included with the desired sample size even if they differ in annual volume up to a factor of 9.

Challenge 2: time-varying quality

The quality of a provider may vary over time. Consequently, fixing the observation period or fixing the sample size can lead to different conclusions about a single provider. The most crucial issue is a sudden change from good to poor performance. In a high-volume provider, this can be detected rather quickly when allowing short observation

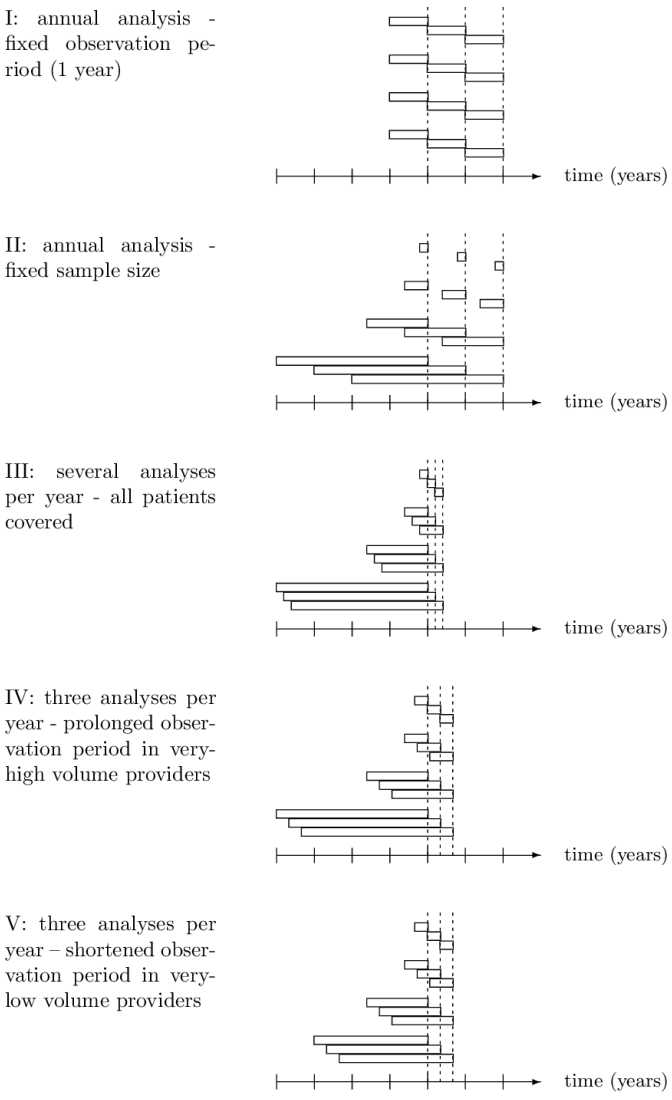

Figure 5 Five scenarios for the scheduling of analyses and observation periods. Shown are three subsequent analyses for four providers reflecting (from top to bottom) a provider with very high annual volume, a provider with annual volume above the average, a provider with annual volume below the average and a provider with very low annual volume. Each box reflects the observation period included in the analysis. The time points of scheduling are indicated by dashed lines.

periods, in particular if several analyses per year are made (scenarios III–V in figure 5). On the other hand, for a low-volume provider a long observation period may mask a sudden change. However, even when fixing the observation period to a shorter interval, it is not likely to detect this change due to the limited sample size the low-volume provider is contributing.

In case a quality indicator may be affected by seasonal variation (eg, due to typical changes in the patient population or working conditions during a calendar year), it might be necessary to round the provider-specific observation periods to full years.

Challenge 3: overlap between analyses

Scheduling fixed sample size analyses regularly in time implies that there is overlap between patient populations for different analyses, especially for low-volume providers (cf figure 5). This has at least two consequences. First, the results for low-volume providers from one analysis are highly predictive of the results from the subsequent analyses. This may encourage well-performing low-volume providers to de-emphasise their efforts in maintaining high quality. On the other side, poorly performing low-volume providers may want to include only new patients in the next analysis in order to have a chance that their efforts to improve quality become visible. Second, high-volume providers may interpret the higher fluctuations from analysis to analysis (compared with low-volume providers) as a higher risk to be marked as poorly performing due to random fluctuations, and hence as unfair. Hence, it is essential to inform them that they have in the long run the same risk to be marked as any low-volume provider of the same quality.

Challenge 4: introducing new indicators

Fixing the sample size implies that observation periods are defined retrospectively: starting at the current time point of the intended evaluation, we go backwards into the past until the sample size is reached. Introducing new indicators or a new assessment procedure for an existing indicator may imply that for some providers the intended sample size cannot be reached at the first evaluation after the introduction, requiring to accept the available number of patients as sample size. This issue does not appear when using fixed observation periods if the introduction happens at an evaluation time point.

Challenge 5: choosing the sample size

Fixing the sample size requires the choice of a sample size. Using statistical power considerations for sample size determination for provider profiling is not straightforward, as provider profiling can be used for different purposes—this is illustrated by our two examples. Nevertheless, there exist practical suggestions for sample size calculations.[15 22 23] The essential point is that sample size considerations can take into account knowledge about the expected prevalence and the expected spread in performance and volume across providers based on the data from previous years. By contrast, when fixing the observation period to 1 year, the sample size is determined completely by the annual volume of patients.

Challenge 6: reorganisation of the data flow

Reporting provider profiles requires that some reporting body has access to all necessary information. Often, providers enter the individual patient data into a central database accessible to the body. The body can then directly implement changes in the reporting practice. However, the data flow between the providers and the central body is typically a complex process involving checking and cleaning procedures to ensure completeness and high data quality. The annual reporting defines typically a corresponding cycle in the data flow. When increasing the frequency of reporting there is a need for

a more continuous quality control, increasing the burden for the providers. This is even more the case if providers are already required to provide aggregated data.

DISCUSSION

Fixing the sample size for provider profiling analyses has some clear advantages compared with fixing the observation period: a dependence of decisions on the volume is avoided, the visualisation of the results and the ranking with respect to the observed quality becomes much simpler and the quality of decisions is improved in the long run. Practical challenges in implementing this idea may make it necessary to allow some deviation from a fixed sample size for some providers. However, even in that case the advantages shown above remain, in the sense that the results are still easier to compare across providers and decisions tend to be better.

One practical obstacle against fixing the sample size might be the necessity to measure quality under stable conditions for more than 1 year. However, this seems to be the case today in many countries for many indicators. On the other side, the introduction of new indicators or major changes to the current assessment procedures will always make it necessary to deviate from a strict fixed sample size approach for some time periods. Practical obstacles may also arise from the need to adapt the data flow to a more continuous reporting.

In spite of these obstacles, we regard these advantages as sufficiently relevant to consider alternative reporting strategies aiming at more comparable sample sizes across providers. A first simple step would be to present reports based on a fixed sample size in addition to reports based on a fixed observation period. If this type of reporting becomes popular, further refinements may be considered. The long-term aim should be an 'optimal strategy' informed by knowledge about the expected magnitude of fluctuations in quality across providers and over time, while taking simultaneously the organisational needs into account.

Moving into this direction will make provider profiling a more complex task than the current practice of annual reporting. However, it should be kept in mind that the current practice of annual reporting has just emerged over time and does not involve any considerations about optimal design or sample sizes. Hence, moving towards fixed sample sizes may make a significant contribution to improving the field, in particular if this happens simultaneously with additional methodological improvements.[24]

We have not considered in this paper the need for case-mix adjustment.[17 25] Addressing this issue requires the use of standardised prevalence values instead of raw prevalence values. However, most considerations presented in this paper apply equally to standardised prevalence values, except that the precision of these estimates depends also on the distribution of the patient characteristics for each provider. However, this variation in precision will be usually small. In our simulations, we also ignored a potential dependence of the deviations from the overall level on the volume of the providers.

Finally, we note that data on quality indicators can and should be also used for other purposes than comparing providers, for example, for internal quality monitoring or creating annual administrative reports. These purposes imply specific ways of reporting and should be handled separately.

CONCLUSION

Fixing the sample size instead of fixing the observation period is a valuable alternative in performing provider profiling, and we recommend supplementing the current practice accordingly. This should be regarded as a step to place the design of the analysis and reporting strategy for provider profiling data on a more rationale base.

Contributors WV developed the idea for this project and conducted all computations. All authors participated in the interpretation and phrasing of the results. All authors approved the final version of the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

Werner Vach http://orcid.org/0000-0003-1865-8399

REFERENCES

1 Normand S-LT, Glickman ME, Gatsonis CA. Statistical methods for profiling providers of medical care: issues and applications. *J Am Stat Assoc* 1997;92:803–14.
2 Bird SM, Sir David C, Farewell VT, *et al*. Performance indicators: good, bad, and ugly. *J R Stat Soc Ser A Stat Soc* 2005;168:1–27.
3 Hospital quality alliance (HQA). Available: https://www.allhealthpolicy.org/glossary/hospital-quality-alliance/ [Accessed 30 Mar 2021].
4 Centers for Medicare & Medicaid Services (CMS). Available: https://www.cms.gov/ [Accessed 30 Mar 2021].
5 Intensive Care National Audit & Research Centre (ICNARC). Available: https://www.icnarc.org/Our-Audit/Audits/Cmp/About [Accessed 30 Mar 2021].
6 Institut für Qualitätssicherung und Transparenz Im Gesundheitswesen (IQTIG). Available: https://iqtig.org/ [Accessed 30 Mar 2012].
7 Scope Santé – Haute Autorité Santé. Available: https://www.scopesante.fr/#/ [Accessed 30 Mar 2021].
8 Swiss national association for quality development in hospitals and clinics (ANQ). Available: https://www.anq.ch/en/ [Accessed 30 Mar 2021].
9 Nationale intensive care Evaluatie (NICE). Available: https://stichting-nice.nl/ [Accessed 30 Mar 2021].

10 Sundhedsdatastyrelsen – Kliniske kvalitesdatabaser. Available: https://sundhedsdatastyrelsen.dk/da/registre-og-services/om-de-kliniske-kvalitetsdatabaser [Accessed 30 Mar 2021].
11 CMS Hospital Chartbook. Available: https://www.CMSHospitalChartbook.com [Accessed 30 Mar 2021].
12 ANQ Fachbereiche Messergebnisse. Available: https://www.anq.ch/de/messergebnisse/ [Accessed 30 Mar 2021].
13 Gemeinsamer Bundesausschuss – Refernzdatenbank. Available: https://www.g-ba-qualitaetsberichte.de/#/search [Accessed 30 Mar 2021].
14 Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med* 2005;24:1185–202.
15 Verburg IW, Holman R, Peek N, *et al*. Guidelines on constructing funnel plots for quality indicators: a case study on mortality in intensive care unit patients. *Stat Methods Med Res* 2018;27:3350–66.
16 Normand S-LT, Ash AS, Fienberg SE, *et al*. League tables for hospital comparisons. *Annu Rev Stat Appl* 2016;3:21–50.
17 Varewyck M, Goetghebeur E, Eriksson M, *et al*. On shrinkage and model extrapolation in the evaluation of clinical center performance. *Biostatistics* 2014;15:651–64.
18 Austin PC. Bayes rules for optimally using Bayesian hierarchical regression models in provider profiling to identify high-mortality hospitals. *BMC Med Res Methodol* 2008;8:30.
19 Jones HE, Spiegelhalter DJ. The identification of "unusual" health-care providers from a hierarchical model. *Am Stat* 2011.
20 Laird NM, Louis TA. Empirical Bayes ranking methods. *J Stat Educ* 1989;14:29–46.
21 Jewett PI, Zhu L, Huang B, *et al*. Optimal Bayesian point estimates and credible intervals for ranking with application to County health indices. *Stat Methods Med Res* 2019;28:2876–91.
22 Seaton SE, Manktelow BN. The probability of being identified as an outlier with commonly used funnel plot control limits for the standardised mortality ratio. *BMC Med Res Methodol* 2012;12:98.
23 Seaton SE, Barker L, Lingsma HF, *et al*. What is the probability of detecting poorly performing hospitals using funnel plots? *BMJ Qual Saf* 2013;22:870–6.
24 Pasquali SK, Banerjee M, Romano JC, *et al*. Hospital performance assessment in congenital heart surgery: where do we go from here? *Ann Thorac Surg* 2020;109:621–6.
25 Ash AS, Ellis RP. Risk-adjusted payment and performance assessment for primary care. *Med Care* 2012;50:643–53.