The good, the bad and the ugly: What do we really do when we identify the best and the worst organisations?

Gary A Abel ,[1] Denis Agniel,[2] Marc N Elliott [2]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjqs-2023-017039).

[1]Department of Health and Community Science, University of Exeter, Exeter, UK
[2]RAND Health, RAND, Santa Monica, CA, USA

Correspondence to
Dr Marc N Elliott;
elliott@rand.org

Received 19 December 2023
Accepted 7 October 2024
Published Online First
18 October 2024

To cite: Abel GA, Agniel D, Elliott MN. *BMJ Qual Saf* 2025;**34**:53–61.

ABSTRACT

Identifying high and poorly performing organisations is common practice in healthcare. Often this is done within a frequentist inferential framework where statistical techniques are used that acknowledge that observed performance is an imperfect measure of underlying quality. Various methods are employed for this purpose, but the influence of chance on the degree of misclassification is often underappreciated. Using simulations, we show that the distribution of underlying performance of organisations flagged as the worst performers, using current best practices, was highly dependent on the reliability of the performance measure. When reliability was low, flagged organisations were likely to have an underlying performance that was near the population average. Reliability needs to reach at least 0.7 for 50% of flagged organisations to be correctly flagged and 0.9 to nearly eliminate incorrectly flagging organisations close to the overall mean. We conclude that despite their widespread use, techniques for identifying the best and worst performing organisations do not necessarily identify truly good and bad performers and even with the best techniques, reliable data are required.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Identifying high and poorly performing organisations is common practice in healthcare with established statistical methods employed for this purpose. Some methods are known to preferentially identify organisations with few data points or identify too many organisations, while others, considered the 'gold-standard' will select an equal proportion of organisations across a range of sample sizes.

WHAT THIS STUDY ADDS

⇒ We find that, despite widespread use, even the best techniques for identifying the best and worst performing organisations only identify truly good and bad performers when the underlying data have high statistical reliability.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Existing practice often focuses on which methods should be employed to identify high and poorly performing organisations rather than the statistical reliability of the data used. This work demonstrates that the latter is far more important and provides a sound theoretical basis for using reliability threshold of 0.7 for this purpose.

INTRODUCTION

Quality improvement (QI) efforts commonly identify the best and worst performing healthcare organisations. Sometimes, QI is linked with payment incentive schemes; for example, the Medicare Hospital Value-Based Purchasing programme awarded hospitals $1.4 billion in performance-based incentives in 2015.[1] Other schemes focus on public reporting and data availability for regulatory bodies. For example, the English General Practice Patient Survey collects patient experience data from approximately 7000 general practices and is publicly reported and used by the Care Quality Commission to inform inspection processes.[2 3] Relatedly, public reporting of best and worst performers can be used to inform patient choice. In some cases, processes are enacted to impede enrolment in poor-performing health plans and facilitate enrolment in high-performing ones.[4] QI research often selects participating organisations based on a quality indicator, either contrasting high and low performers to identify best practices,[5] or to target interventions.[6]

Poor-performing hospitals are also identified for safety-monitoring purposes. The Hospital Standardised Mortality Ratio is used widely, comparing the expected number of deaths at a hospital with the observed number.[7–9] While such data are not used as direct safety indicators, they can prompt further investigations.[10] Similarly, mortality statistics are often monitored for individual surgeons and other specialists.[11 12] Regardless, confidence is needed that organisations identified as being good or bad are indeed among the best and worst performers. Thus, there has been considerable academic investment into developing methods to perform these classifications. In this paper, we focus on simple single-measure indicators. Composite indicators are also often used, but, present numerous challenges that are not always addressed in practice, including often failing to recognise uncertainty.[13] The concepts discussed in this paper are broadly applicable to many indicators and are discussed in general terms.

While many statistical methods exist for identifying the best and worst performers, differences between methods largely concern adjustment for differences in the served populations. We do not address the issue of case-mix or population adjustment but focus on statistical methods used after adjustment. Most performance classification methods belong to one of three categories. The first is simple ranking, where organisations' data are taken at face value and performance classification is based on where an organisation sits in a ranked list of eligible organisations. Simple ranking will preferentially select organisations with a smaller sample size as being the best and worst performers and is suboptimal.[14 15] Best and worst performers based on simple ranking will differ according to the distribution of the number of observations used for each organisation (figure 1). Although this approach has been used historically, most contemporary examples are found in media reporting or local reporting.[16]

Frequentist statistical methods, including descriptive ranking, posit the existence of underlying organisational quality, an organisation's expected performance on an infinite sample of patients; this corresponds to the expected quality in the future under similar circumstances, when making statistical inference, including statements about whether an organisation's underlying performance differs from an overall average, or the construction of a CI about expected future performance.[17] Statistical inference is used to acknowledge that observed performance is an imperfect measure of underlying quality.

The second method is a statistical test of whether individual organisations differ statistically from a reference value, which is often the overall mean (although other quantities can be used, such as a target value or percentile). This may employ CIs, z-scores or other statistics and may reflect the type of data used (eg, the binomial distribution can be used for percentage indicators). These methods are commonplace[18 19] and are the default methods used in England's Public Health Outcomes Framework.[20] Because this method identifies the same organisations as the standard funnel plot method; we consider such methods the same.[21] The funnel plot method is often used when considering mortality associated with individual clinicians.[11 12] These methods typically identify many good and bad performers, preferentially selecting large organisations as the best and worst performers.[22] This occurs because there is more variability in organisational performance than would be expected by random variation alone. This additional variation is due to variation in underlying quality between organisations.[15] One way to conceive of this underlying quality is as the performance one would observe with infinite sample size.

The third method was developed recognising the issues associated with these two sources of variation, known as 'overdispersion', which is commonplace in organisation performance metrics. While 'overdispersion' has been used in different contexts,[23–25] here it refers to the additional variability in z-scores due to these two sources of variation. By calculating overdispersed z-scores, or factoring overdispersion into funnel plots, we can account for real/underlying variability between organisations.[22 26 27] Such methods identify best and worst performers independent of sample size (assuming sample size is not related to underlying performance). This method, recommended by The Committee of Presidents of Statistical Societies,[14] is generally considered the gold-standard and is used by many organisations, including NHS England.[28]

It is widely accepted that uncertainty due to finite sample sizes introduces uncertainty into scores, thus all three methods will be imperfect. The second and third methods have been compared with real-world data.[29] However, the true performance of flagged organisations measured without error has not been considered. It is impossible to address this question empirically with real-world data, as the true performance (ie, that measured without error) is not observed. Instead, theoretical or simulation approaches must be used. Earlier work started to address this question by considering the degree of misclassification that poor reliability introduces into grouping organisations.[30 31] Here, we develop these ideas by examining the performance of the two widely used z-score methods in the presence of differing amounts of chance using a simulation approach combined with a theoretical approach based on Bayes theorem.

Simulation

Each simulation represents a different level of reliability (see box 1), contains 1 million simulated organisations and can be thought of as a vertical slice through a funnel plot (figure 1). For simplicity, we restrict our

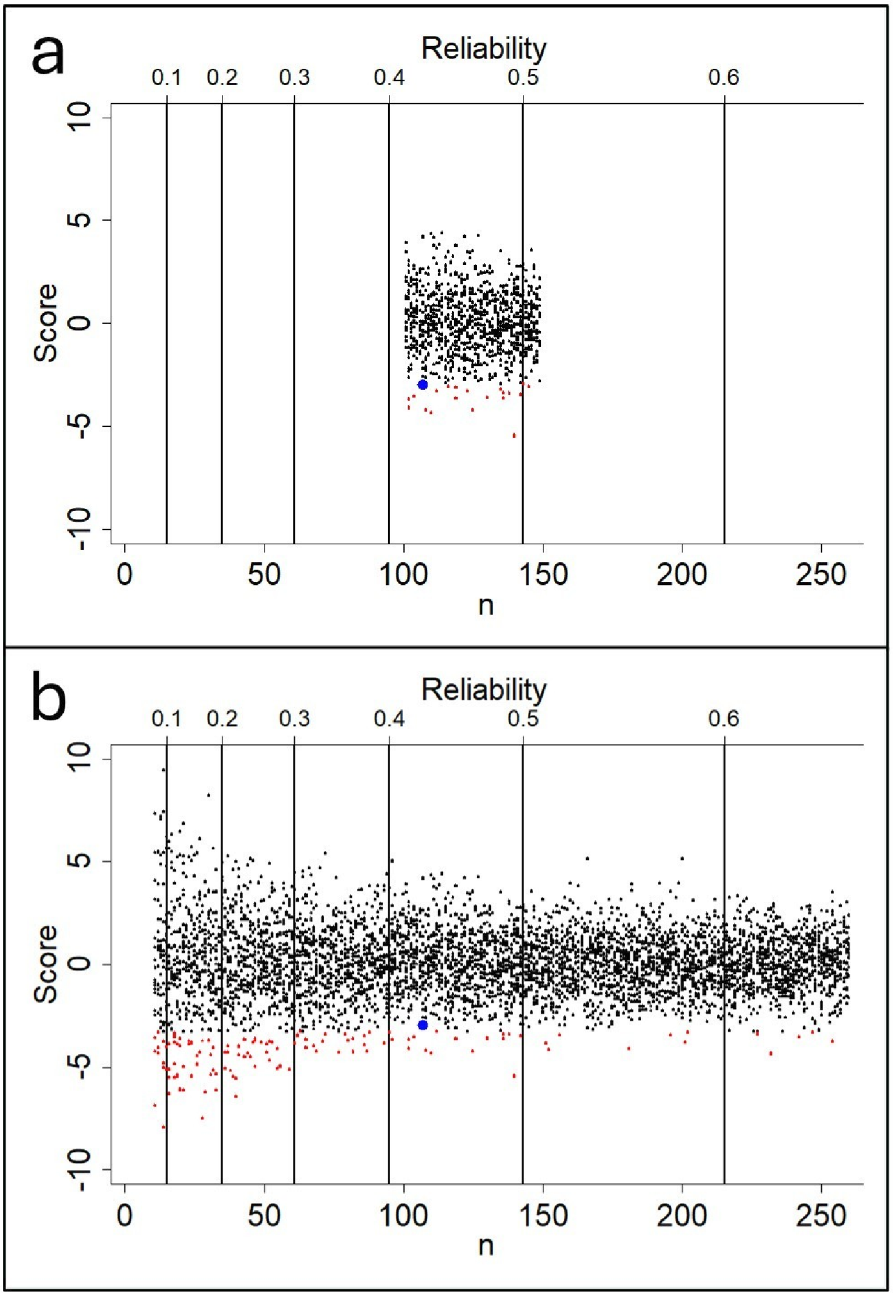

Figure 1 The influence of varying numbers of observations. The above figures show simulated organisational scores on an imagined metric plotted against the number of observations used in each organisation. As expected, there is more variation (scatter) when the number of observations is small, and the influence of chance is high. A wide range of numbers of observations is used in (a), and (b) a restricted range (100< n < 150) is considered. In each case, the lowest 2.5% of scores, based on simple ranking, are plotted in red. In (a), there is a tendency to identify organisations with fewer observations. Importantly, the organisations flagged as poor performers are not the same when we consider a restricted range of numbers of observations (b). The organisation highlighted with a large blue dot is not flagged in (a) but is flagged in (b). Thus, it is hard to make general statements about the outcome of using simple ranking, as it depends on the range of sample sizes involved. In contrast, z-score and overdispersed z-score-based methods depend only on the sample size of the organisation under consideration. In both panels, the organisation highlighted in blue is flagged as a poor performer using an overdispersed z-score method, but not flagged using a standard z-score based method. Such methods are consistent and can be shown to depend only on reliability (box 1). With a continuous metric, reliability can be thought of as a vertical slice through a funnel plot as indicated for the example above.

examination to a normally distributed continuous measure. The underlying distribution of true organisational performance (ie, that which would be observed if very large sample sizes per hospital were available, so that random error was small) is assumed to be normally distributed with a mean of zero and an SD, σ_u, = 1 . Each organisation has an underlying performance score drawn at random from this distribution. To replicate the impact of chance due to finite sample sizes, we add normally distributed noise to this score. To make findings applicable across settings, rather than specify sample sizes and distributions for individual patients, we define the noise relative to the between -organisation variability in terms of the reliability of the metric

Box 1 Reliability

When considering the reliability of an organisational quality metric we generally refer to Spearman Brown, or inter-unit reliability (also known as rankability). This reliability is a measure of how reliably different organisations can be distinguished, ranked, or classified based on this metric, and takes a value between 0 and 1. When reliability is low, noise due to finite sample sizes will dominate organisational scores, meaning it is difficult to distinguish among organisations. In contrast when reliability is high, the signal to noise ratio is high and observed scores better reflect true performance. Reliability, λ, is formally defined as the ratio between the true underlying organisation variance, σ_u^2, and the variance of observed organisational scores σ_o^2. It can also be expressed considering the variance of noise/chance, σ_n^2, or in terms of the within organisation (or patient level) variance σ_w^2, and the sample size for an organisation, n, ie,

$$\lambda = \frac{\sigma_u^2}{\sigma_o^2} = \frac{\sigma_u^2}{\sigma_u^2+\sigma_n^2} = \frac{\sigma_u^2}{\sigma_u^2+\sigma_w^2/n}$$

From this, we can ascertain the SD of a noise distribution, σ_n, such that the reliability (λ) of the score is a set value for each simulation and is given by$\sigma_n = \sqrt{(1-\lambda)/\lambda}$

(see figure 1). We refer to the sum of the underlying performance and noise as the observed performance. Nine simulations are performed for reliability values between 0.1 and 0.9 in steps of 0.1.[30 32–36]

To illustrate this, the distribution of the noise component and observed scores from the simulation is shown in figure 2. In the low-reliability case (reliability=0.3), the noise component (figure 2c) has much more dispersion than in the high reliability case (reliability=0.9 figure 2d). Where reliability <0.5, the variance of noise exceeds the variance of the underlying distribution, whereas when reliability >0.5, the opposite is true. Greater dispersion in the noise component causes greater dispersion in the observed distributions, since observed scores are the sum of the noise component and the underlying scores (which do not depend on reliability). When reliability is low, the observed-score variance is much greater than that of the underlying scores (figure 2e), and when it is very high (reliability=0.9), there is little increase in the variance from the underlying to observed scores (figure 2f).

Identifying the best and worst performing organisations

As we know the underlying performance of simulated organisations, we can identify the true best and worst performers. The simulated observed performance is used to flag the best and worst performers using the two methods. The first (standard z-score/funnel plots) identifies all organisations with observed scores greater than $\pm 1.96\sigma_n$, that is, with an observed score which would have a p value <0.05 if a statistical test was applied comparing it to the overall mean. The second method (overdispersed z-score/funnel plots) first calculates the SD of the observed scores (σ_o by adding the variance of the underlying scores and the variance of the noise $\left(\sigma_o = \sqrt{\sigma_u^2 + \sigma_n^2}\right)$ and identifies all organisations with observed scores greater than $\pm 1.96\sigma_o$. Importantly, the two methods seek to identify different sets of organisations. The standard z-score method aims to identify organisations different from the mean, whereas the overdispersed z-score aims to identify those at the edge of the distribution.

For each method (standard and overdispersed z-scores), we examine the distribution of the underlying performance of organisations flagged as being the worst performers: we produce histograms of the underlying performance of practices flagged as the worst performers and calculate the proportion of flagged worst performers falling into the following categories:

- Those which have an underlying score $<-1.96\sigma_u$, that is, those in the worst 2.5% of the distribution.
- Those which have an underlying score $<-1\sigma_u$, that is, those poor performers outside of the core of the distribution (~the worst 16%).
- Those which have an underlying score <0, that is, organisations performing worse than average.

While a simulation is used here for illustrative purposes, one can obtain the expected distribution of underlying scores for organisations flagged as the best or worst performers via Bayes theorem (see online supplemental Digital Content 1).

RESULTS

Figure 3 shows the results of the simulations (histograms) and predicted distributions from Bayes theorem (lines), demonstrating very good agreement between the simulations and predicted distributions. Characteristics of the organisations flagged as poor performers calculated from Bayes theorem are shown in figure 4, with similar results from the simulation shown in online supplemental Digital Content 2.

Standard z-score method

The number of flagged organisations depends on reliability (online supplemental Digital Content 2), with 3.1% (31 210/1 000 000) flagged as poor performers when reliability was 0.1 and 26.7% (267 130/1 000 000) flagged as poor performers when reliability was 0.9. Second, while the distribution of underlying performance for flagged organisations changes with reliability (figure 3a, c, e, g and i), it fails to consistently flag true best or worst performers at all levels of reliability (figure 4a). While at high reliabilities, nearly

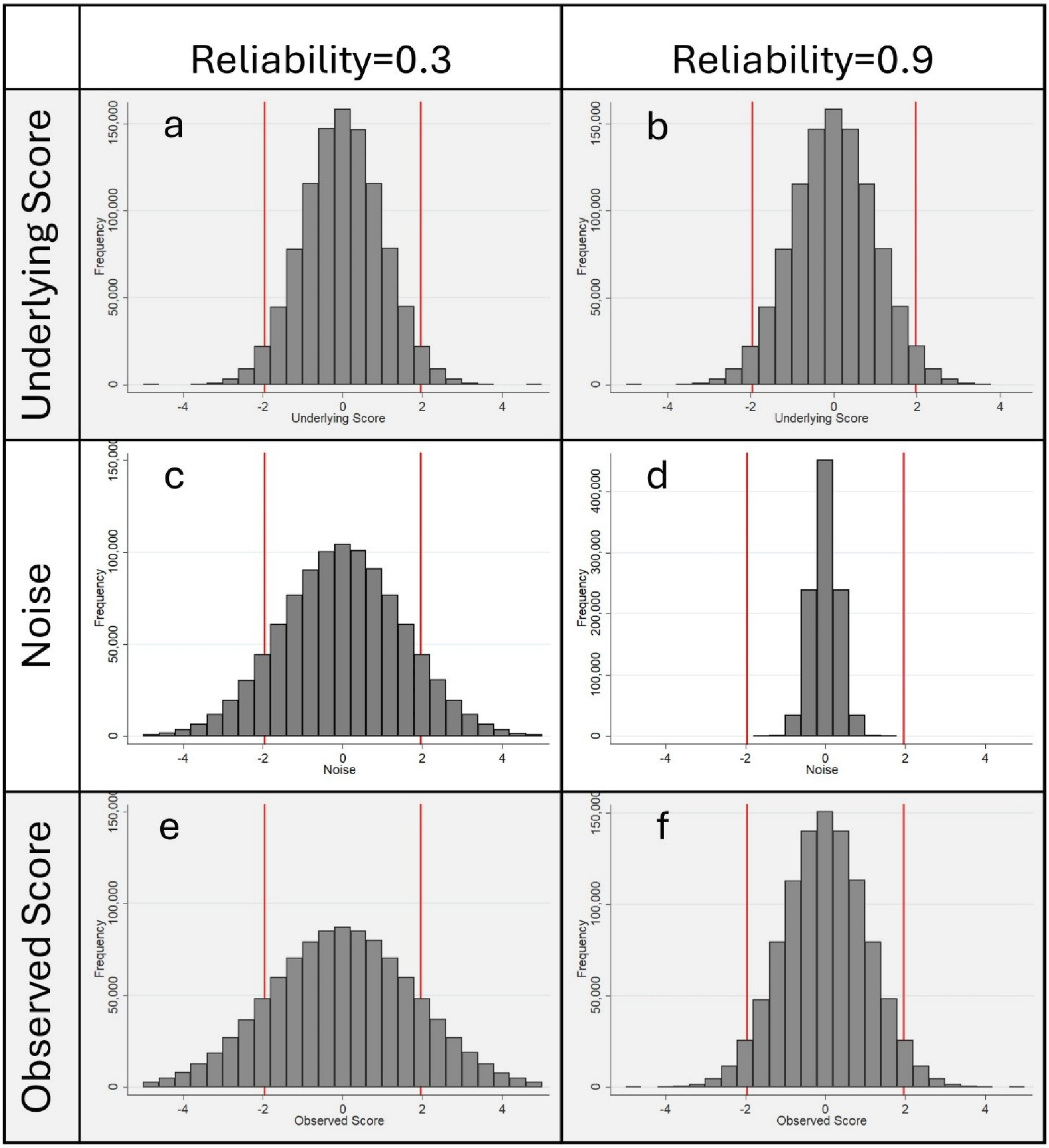

Figure 2 Distribution of simulated underlying score (a and b, noise component (c and d) and observed scores (e and f) for reliabilities of 0.3 (a, c and e) and 0.9 (b, d and f). Vertical lines indicate ±1.96 SD of the underlying score.

all organisations flagged as the worst performers have an underlying performance below the mean (eg, 99.6% for reliability of 0.9, figure 3i and figure 4a), a substantial proportion has an underlying score less than 1 SD from the mean (eg, 57.3% for reliability of 0.9), and very few are truly at the extremes of the distribution (eg, 9.3% for a reliability of 0.9). We see that the proportion of flagged organisations with an underlying score below the mean increases monotonically with increasing reliability (figure 4a). However, the same is not true when we consider organisations with an underlying score more than 1, or 1.96, SD below the mean. The highest proportion of flagged organisations having underlying scores more than 1 SD below the mean is identified when reliability is 0.68. This reliability reduces to 0.5 when considering organisations 1.96 SD below the mean.

Overdispersed z-score method

As expected, the overdispersed z-score method flags a consistent proportion (~2.5%) of organisations regardless of reliability (online supplemental Digital

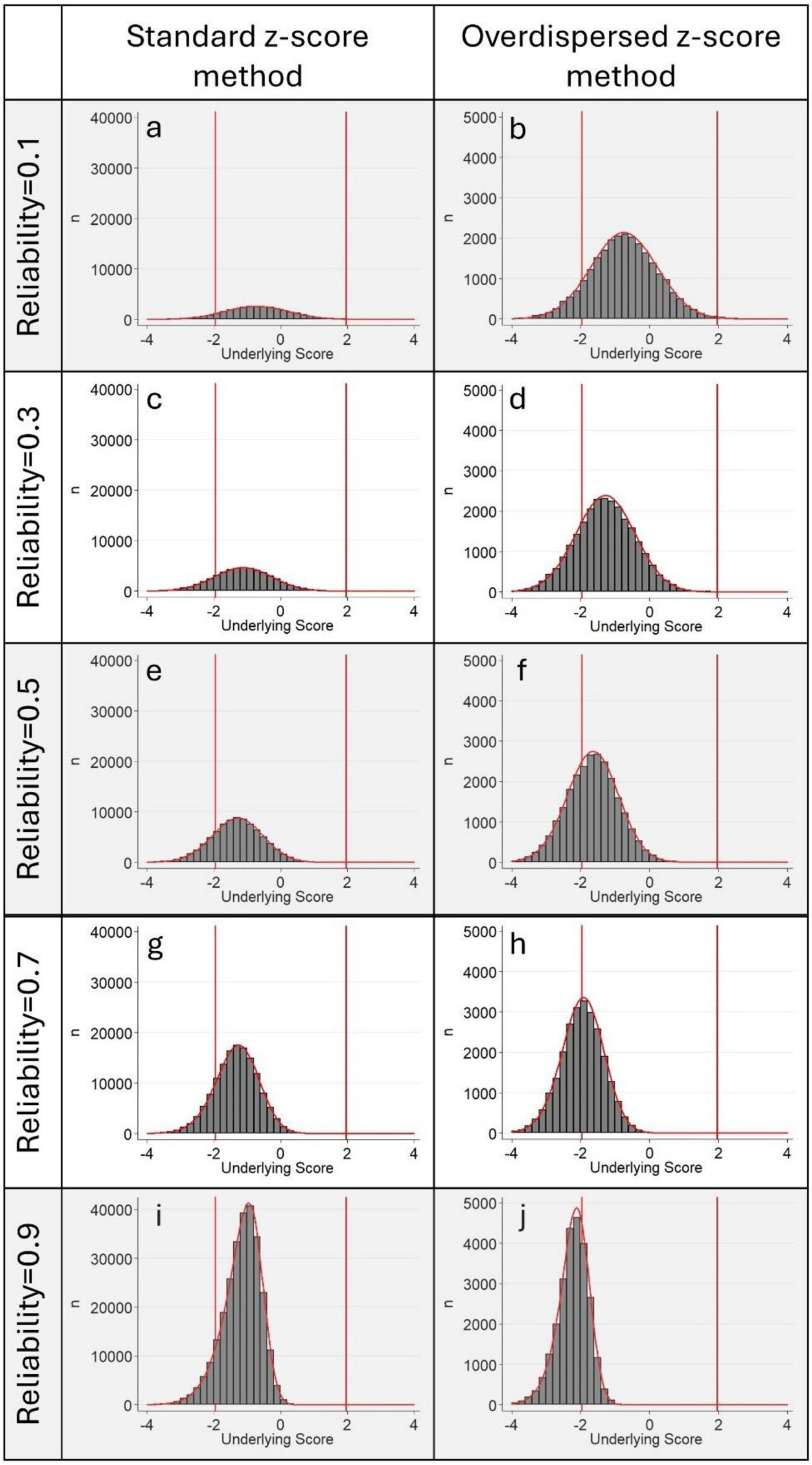

Figure 3 Distribution of simulated underlying scores (histograms) for organisations flagged as the worse performing organisations for a range of reliabilities using a standard z-score method (a, c, e, g and i) and an overdispersed method (b, d, f, h, j). Curves show the predicted distributions from Bayes theorem and the vertical lines indicate±1.96 SD of the underlying score distribution for all organisations.

Content 3, figure 3b, d, f, h and i). However, the distribution of underlying performance for flagged organisations changes considerably with reliability. At all reliabilities, more flagged organisations have an underlying performance below the global mean than above, increasing from 77.9% when reliability was 0.1 (figures 3b and 4b) to over 99% for reliabilities exceeding 0.6 (figure 3i,b). Unlike the standard z-score method, when incorporating overdispersion, the percentage of organisations flagged as poor performers with underlying scores either more than 1 SD below the global mean or with extremely low

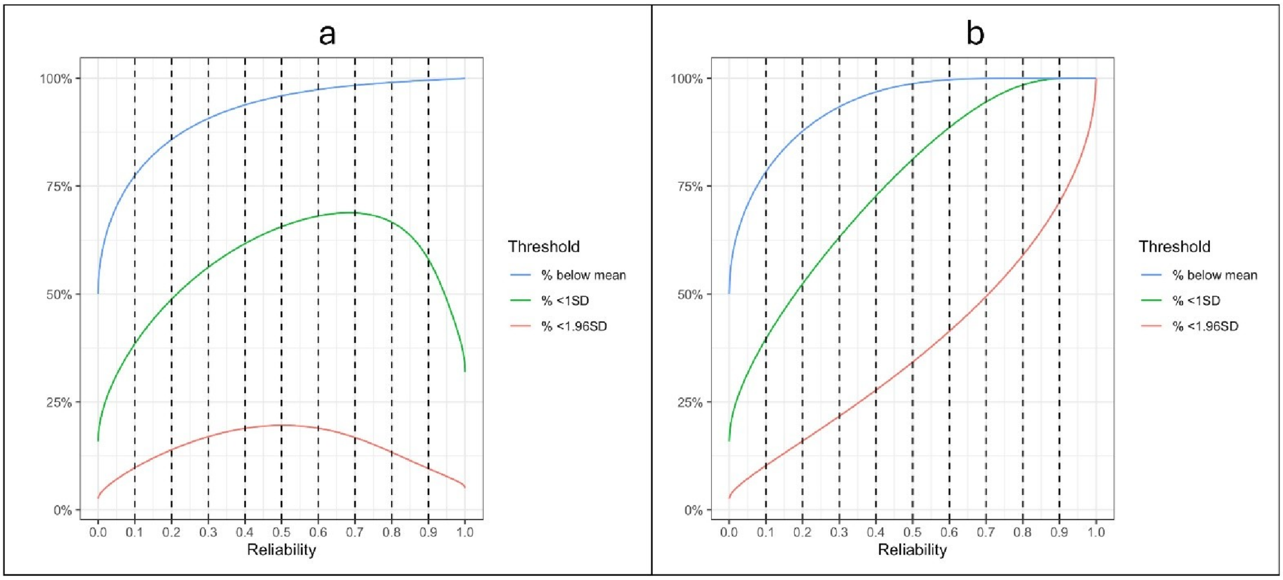

Figure 4 Characteristics of the distribution of underlying scores for organisations flagged as the worst performing organisations as a function of reliability in the situation where the worst performers are identified using a standard z-score method (a) and overdispersed z-score method (b).

underlying scores (<−1.96 SD) increases monotonically with increasing reliability (from 39.6% and 10.2%, respectively, for reliability of 0.1 to 99.9% and 71.4% for a reliability of 0.9 figure 4b). For around half of organisations flagged as poor performers to have an underlying performance more than 1.96 SD below the global mean reliability had to be at least 0.71 (figure 4b). Even at this level of reliability, ~4.9% of flagged organisations had underlying performances in the core of the underlying distribution (scores within 1 SD of global mean).

DISCUSSION

Summary of main findings

We have compared two methods for flagging best and worst performers, at different reliabilities of the quality metric when statistical adjustment is used to differentiate observed performance from underlying quality. The arguments we make here are based in the frequentist inferential framework. There are likely parallels to be drawn within some other frameworks (eg, Bayesian statistics) and consideration should be given to the extent to which they may apply in any given method.

When reliability is low, most flagged organisations have an underlying score in the core of the distribution, using either method. When reliability is very low (0.1), the distribution of underlying performance of flagged organisations is quite similar to that of all organisations—organisations are flagged almost at random. Under these conditions, the noise component dominates underlying performance (signal), and organisations are flagged when they have a good or bad score due to chance, regardless of the method used. When reliability is high, the standard z-score method flags a very large number of best or worse performers, ignoring only those with underlying scores close to the overall mean, consistent with previous work on these methods.[21 22] In contrast, the overdispersed z-score method flags the same proportion of organisations as best or worst performers, regardless of reliability. For reliabilities well over 0.7, the degree of misclassification is low, with most organisations having an underlying score at the extremes of the distribution.

Implications for the use of standard z-score methods

We have demonstrated that standard z-score methods lead to substantial misclassification, with many flagged organisations having an underlying performance within the core of the distribution regardless of the reliability of the indicator used. In other words, many flagged organisations have typical performance and are unlikely to be consistently flagged as good or poor performers from year to year. Given that these methods often flag many organisations, they are rarely useful in identifying exclusively good and poor performers.

Implications for the use of overdispersed z-score methods

Overdispersed z-score methods are considered the gold standard and are often applied when influence of chance is low or variable, as where sample size per organisation is highly variable. Despite this, we show high misclassification when reliability is low, with many flagged organisations having underlying scores within the core of the distribution, just as with the standard z-score method. When reliability is high, the overdispersed z-score method performs better, flagging organisations towards the extreme of the underlying distribution. In other words, when reliability is high, the overdispersed z-score method does identify organisations that really are performing well or badly;

overdispersed z-scores are appropriate when reliability exceeds 0.7.

Importance of reliability profiling

Profiling the reliability of organisational performance indicators is not routine. There are exceptions to this, some of which have underpinned changes in the data collection used to construct these indicators or the reporting conventions,[32 34 37] with some examples of unreliable scores being flagged.[38] Without assessing reliability, it is hard to know how well an indicator is performing and the degree of resultant misclassification. Currently, there is no universally agreed convention for the minimum required reliability to justify the use of an indicator. Authors have argued for thresholds of 0.7, 0.8 or 0.9 depending on the application, but with no real empirical basis for these thresholds. Here, we show clearly that when reliability is below 0.7, more than half of flagged organisations using the overdispersed z-score method are misclassified (using a threshold of 1.96) and that many organisations have an underlying score in the core of the distribution, supporting the use of the previously proposed thresholds.

Notably, for a given reliability, simple ranking and overdispersed z-scores are equivalent. Thus, if reliability is high, there will be little gained by the more complex overdispersed z-score methods. Rather than focusing on the method used, our findings suggest that it is more important that only reliable indicators are used, rather than employing complex methods for identifying the best and worst performers.

There are several potential consequences of using unreliable indicators. Improvement efforts may be misplaced, leading to both actual and opportunity costs of not addressing real quality deficits. There may also be financial implications, either directly related to a lack of appropriate performance-related pay and indirectly due to lower patient numbers. Comparisons of high-performers and low-performers may not reach useful conclusions if organisations are not meaningfully different. Similar issues may apply to qualitative research focused on apparently low or high performers, which may not elucidate relevant factors.

Cause of misclassification

We have demonstrated that Bayes theorem explains misclassification well. Although the probability of one organisation being flagged as a poor performer increases with poorer underlying performance, most organisations are in the core of the distribution; unless reliability is high, the same holds for flagged organisations.

We have focused on identifying best and worst performers, rather than outlier detection, treating organisations as drawn from a single distribution. If a subset of organisations come from a distinct distribution due to being inherently different, or outliers, the methods described above will perform differently, especially if a very strict threshold such as z-scores of 3 or 4 is applied and more organisations are identified than would be expected from a single normal distribution. If there truly is a separate population, then a low reliability may be less of an issue.

CONCLUSION

Frequentist statistical techniques are commonly applied to performance metrics that acknowledge that observed performance is an imperfect measure of underlying quality. Despite widespread use, the techniques commonly used for identifying best and worst performers do not necessarily identify true good and bad performers and reliable data are still required. Methods based on standard z-scores are unlikely to be useful in most scenarios, whereas methods which account for overdispersion and even simple ranking may be useful when reliability is high. This work provides support for commonly used thresholds of reliability of 0.7 and 0.9.

Contributors GA contributed to the conception and design, acquisition of data, analysis and interpretation of data, drafting of the manuscript, critical revision for important intellectual content and is the guarantor. DA contributed to the analysis and interpretation of data and critical revisions for important intellectual content. ME contributed to the conception and design, acquisition of data, analysis and interpretation of data and critical revision for important intellectual content.

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 applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer-reviewed.

Data availability statement There are no data in this work.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Gary A Abel http://orcid.org/0000-0003-2231-5161
Marc N Elliott http://orcid.org/0000-0002-7147-5535

REFERENCES

1 Elliott MN, Beckett MK, Lehrman WG, *et al*. Understanding The Role Played By Medicare's Patient Experience Points System In Hospital Reimbursement. *Health Aff (Millwood)* 2016;35:1673–80.

2 NHS England. GP patient survey: nhs england. 2024 Available: https://gp-patient.co.uk/surveysandreports

3 Campbell J. Patients' experience of primary care: James Mackenzie Lecture 2017. *Br J Gen Pract* 2019;69:38–9.

4 Haviland AM, Damberg CL, Mathews M, *et al*. Shifting From Passive Quality Reporting to Active Nudging to Influence Consumer Choice of Health Plan. *Med Care Res Rev* 2020;77:345–56.

5 Reeves D, Hann M, Rick J, *et al*. Care plans and care planning in the management of long-term conditions in the UK: a controlled prospective cohort study. *Br J Gen Pract* 2014;64:e568–75.

6 Vindrola-Padros C, Ledger J, Barbosa EC, *et al*. The Implementation of Improvement Interventions for "Low Performing" and "High Performing" Organisations in Health, Education and Local Government: A Phased Literature Review. *Int J Health Policy Manag* 2020.

7 Pouw ME, Peelen LM, Lingsma HF, *et al*. Hospital standardized mortality ratio: consequences of adjusting hospital mortality with indirect standardization. *PLoS One* 2013;8:e59160.

8 Brien SE, Ghali WA. Public reporting of the hospital standardized mortality ratio (HSMR): implications for the Canadian approach to safety and quality in health care. *Open Med* 2008;2:e70–3.

9 Jarman B, Gault S, Alves B, *et al*. Explaining differences in English hospital death rates using routinely collected data. *BMJ* 1999;318:1515–20.

10 Mohammed MA, Stevens AJ. A Simple Insightful Approach to Investigating a Hospital Standardised Mortality Ratio: An Illustrative Case-Study. *PLoS One* 2013;8:e57845.

11 Walker K, Neuburger J, Groene O, *et al*. Public reporting of surgeon outcomes: low numbers of procedures lead to false complacency. *The Lancet* 2013;382:1674–7.

12 Kunadian B, Dunning J, Roberts AP, *et al*. Funnel plots for comparing performance of PCI performing hospitals and cardiologists: demonstration of utility using the New York hospital mortality data. *Catheter Cardiovasc Interv* 2009;73:589–94.

13 Barclay M, Dixon-Woods M, Lyratzopoulos G. The problem with composite indicators. *BMJ Qual Saf* 2019;28:338–44.

14 Ash A, Fienberg SF, Louis T, *et al*. Statistical issues in assessing hospital performance. 2012.

15 Abel G, Elliott MN. Identifying and Quantifying Variation between Healthcare Organisations and Geographical Regions: Using Mixed-Effects Models. BMJ Quality, 2019.

16 ProPublica. Surgeon scorecard: propublica. 2015 Available: https://projects.propublica.org/surgeons

17 Elliott MN, Zaslavsky AM, Cleary PD. Are finite population corrections appropriate when profiling institutions? *Health Serv Outcomes Res Method* 2006;6:153–6.

18 New york state DoH. Cardiovascular disease data and statistics.

19 Minnesota Community Measurement. Minnesota health care disparities by race, hispanic ethnicity, language and country of origin. 2019.

20 Office for Health Improvement and Disparities. Public Health Profiles. 2024 Available: https://fingertips.phe.org.uk/profile/public-health-outcomes-framework/supporting-information/further-info

21 Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med* 2005;24:1185–202.

22 Spiegelhalter DJ. Handling over-dispersion of performance indicators. *Qual Saf Health Care* 2005;14:347–51.

23 Hinde J, Demétrio CGB. Overdispersion: Models and estimation. *Comput Stat Data Anal* 1998;27:151–70.

24 Cameron AC, Trivedi PK. Regression-based tests for overdispersion in the Poisson model. *J Econom* 1990;46:347–64.

25 Berk R, Overdispersion MJM, regression P. *J Quant Criminol* 2008;24:269–84.

26 Haneuse S, Dominici F, Normand S-L, *et al*. Assessment of Between-Hospital Variation in Readmission and Mortality After Cancer Surgical Procedures. *JAMA Netw Open* 2018;1:e183038.

27 Cohen ME, Ko CY, Bilimoria KY, *et al*. Optimizing ACS NSQIP modeling for evaluation of surgical quality and risk: patient risk adjustment, procedure mix adjustment, shrinkage adjustment, and surgical focus. *J Am Coll Surg* 2013;217:336–46.

28 NHS England and NHS Improvement. NHS Oversight Framework – CCG year end assessment 2019/20 - Methodology manual. 2020.

29 Alexandrescu R, Bottle A, Jarman B, *et al*. Classifying hospitals as mortality outliers: logistic versus hierarchical logistic models. *J Med Syst* 2014;38:29:7:.

30 Adams JL, Mehrotra A, Thomas JW, *et al*. Physician cost profiling--reliability and risk of misclassification. *N Engl J Med* 2010;362:1014–21.

31 Paddock SM, Adams JL, Hoces de la Guardia F. Better-than-average and worse-than-average hospitals may not significantly differ from average hospitals: an analysis of Medicare Hospital Compare ratings. *BMJ Qual Saf* 2015;24:128–34.

32 Abel G, Saunders CL, Mendonca SC, *et al*. Variation and statistical reliability of publicly reported primary care diagnostic activity indicators for cancer: a cross-sectional ecological study of routine data. *BMJ Qual Saf* 2018;27:21–30.

33 Roland M, Elliott M, Lyratzopoulos G, *et al*. Reliability of patient responses in pay for performance schemes: analysis of national General Practitioner Patient Survey data in England. *BMJ* 2009;339:b3851.

34 Lyratzopoulos G, Elliott MN, Barbiere JM, *et al*. How can health care organizations be reliably compared?: Lessons from a national survey of patient experience. *Med Care* 2011;49:724–33.

35 Verburg IWM, de Keizer NF, Holman R, *et al*. Individual and Clustered Rankability of ICUs According to Case-Mix-Adjusted Mortality. *Crit Care Med* 2016;44:901–9.

36 van Dishoeck A-M, Lingsma HF, Mackenbach JP, *et al*. Random variation and rankability of hospitals using outcome indicators. *BMJ Qual Saf* 2011;20:869–74.

37 Barclay ME, Lyratzopoulos G, Greenberg DC, *et al*. Missing data and chance variation in public reporting of cancer stage at diagnosis: Cross-sectional analysis of population-based data in England. *Cancer Epidemiol* 2018;52:28–42.

38 Centers for Medicare and Medicaid Services. Stratified Reporting. 2024 Available: https://www.cms.gov/About-CMS/Agency-Information/OMH/research-and-data/statistics-and-data/stratified-reporting