Reducing the number of avoidable deaths in hospital is the focus of many quality improvement initiatives worldwide.\textsuperscript{3} Comparing indicators of avoidable mortality between different hospitals could help to target improvement efforts, but optimally defining and measuring hospital deaths that could be deemed preventable remains a challenge.\textsuperscript{2} Unlike performance comparisons based on hospital standardised mortality ratio (HSMR), a new policy initiative announced by the UK Government will rank hospitals for avoidable mortality based on case reviews of 2000 deaths in English hospitals each year. Although this initiative aims to overcome limitations of current policies, two statistical properties of the proposed approach mean that it is unsuitable for classifying hospital performance.

The first issue relates to the ability to identify whether any one death really was avoidable on a case-by-case basis. It would appear\textsuperscript{3} that the planned process is based on work by Hogan et al\textsuperscript{4} using retrospective case record review (RCRR). In line with previous studies using RCRR, these investigators asked experienced clinicians to rate whether a death was preventable on a 6-point Likert scale.\textsuperscript{4,5} Their study recognised that the use of a semicontinuous scale better reflects ‘the probabilistic nature of reviewers’ decision making more closely than requiring a simple “yes” or “no” response’.\textsuperscript{4,5} However, in operationally defining an avoidable death, the probabilistic component of the instrument is lost because a fixed cut-off is used such that deaths where it is judged that there is more than a 50% chance that the death was preventable are classified as avoidable, and those below 50% are not. (It should be noted that the somewhat arbitrary choice of a 50% cut-off value is not the real issue here, but rather the dichotomisation itself is. However, hereafter, we assume a 50% cut-off value is used as proposed).\textsuperscript{4}

By dichotomising cases into being avoidable or not, the information about the distribution is lost. One might naively argue that the probabilities above 50% will average out with those below 50% to give the right answer. In fact, this is only true when the mean chance of a death being avoided (where the chance is greater than zero) is 50%. This is a strong assumption that will nearly always be untrue. To illustrate this further, we can consider two scenarios. First, a scenario where there were 100 deaths, each with a 60% chance of preventability, implying that 60 deaths would have been avoided if there were no problems in care (assuming independence between cases); and another scenario where there were 100 deaths, each with a 20% chance of preventability, implying that 20 deaths would have been avoided. By only focusing on deaths where the judged preventability is greater than 50% (ie, the proposed operational definition of an avoidable death), we would have estimated 100 ‘avoidable deaths’ in the first scenario and zero ‘avoidable deaths’ in the latter—both conclusions being evidently untrue. These errors arise as a result of ignoring that there will be a range of risks that deaths are preventable. In reality, the distribution of risk that
The first challenge is that small numbers can lead to large variation. For example, if each person flipped a fair coin 20 times, we would expect about 10 heads, but the actual number might vary greatly. A useful analogy here is one of flipping coins: even with a small number of flips, large variation can occur. Getting 10 heads is not guaranteed, even if the coin is fair. Similarly, even if a hospital has a high chance of being preventable, we cannot guarantee that it will always be so. The number of preventable deaths per hospital can be highly variable due to chance alone.

The second issue is that small numbers can lead to non-concordant evidence. For example, suppose we have two hospitals, A and B. Hospital A has 10 preventable deaths, while hospital B has 10 non-preventable deaths. If we review 10 cases per hospital, we might expect about 5 cases per hospital to be preventable. If we randomly select 10 cases per hospital, we might get 5 preventable cases in hospital A and 5 non-preventable cases in hospital B. However, if we review 50 cases per hospital, we might get 25 preventable cases in hospital A and 25 non-preventable cases in hospital B. This illustrates that small numbers can lead to non-concordant evidence.

In the scenario described above, 160 acute hospitals in England are reviewed every year, with an average of about 13 cases reviewed per hospital. The second issue with this plan is that it is very unlikely that most hospitals will be flagged as outliers even if there is no true variation between hospitals. The chance variation illustrated will be further exaggerated when extrapolating back to the entire hospital cohort using an appropriate weighting scheme. Previous simulation studies have also provided concordant evidence.

The two issues outlined here are not without potential cure. The use of a dichotomising cut-off when defining preventable deaths can be overcome by recognising the full range of probabilities and factoring them into calculations. For example, this can be done by preserving the case reviewers’ estimate of preventability. By doing so, the number of deaths that could be avoided can be estimated by multiplying the mean chance of a death being preventable by the total number of hospital deaths. In turn, the issue of small numbers may be addressed simply by reviewing more cases. In order to determine how many cases is enough, one should consider the statistical reliability (also known as rankability) of the resulting metrics. Reliability, in this context, is defined as the proportion of observed variance in hospital scores/metrics that is explained by the true (underlying) variance in hospital performance or, put another way, the proportion of the overall observed variation not due to noise. When sample sizes for individual hospitals are small, the uncertainty on those individual hospital scores will artificially inflate the apparent variation between hospitals. When the difference between true and observed variance is small, metrics may be considered reliable; however, there are times when the true variability accounts for only a minority of the observed variability, and metrics are highly unreliable. It should be noted that reliability depends on sample size and the degree of true variability between hospitals as it is easier to distinguish hospitals when there is larger variation between them. Of course, it may not be feasible or practical to review enough cases to produce reliable ranks.

A related concept worth consideration is the identification of outliers. There are many ways in which this is done, but often hospitals for which there is statistical evidence that their performance differs from the national average are flagged as outliers. Some hospitals are likely to be flagged as outliers even when
reliability is low. In such situations (and to a lesser degree when reliability is high), we can observe the apparently paradoxical situation when hospitals flagged as better or as worse than average are not statistically significantly different from some hospitals flagged as average. This apparent paradox is easily explained when one considers that the hypothesis considered when testing if a hospital is different from the national average is not the same as when testing if a hospital is different to another individual hospital. It is easier to distinguish a hospital from a national average than it is to distinguish from another hospital. When ranking hospitals, implicit comparisons between individual hospitals are being made, and thus, it is the latter distinction that matters rather than the former. It should also be borne in mind that when reliability is very high, nearly all hospitals will be statistically significantly different from the national average. For such reasons, it is often preferable to account for the real variation between hospitals, and the resulting overdispersion, when identifying outliers.

In conclusion, classifying the performance of English hospitals for avoidable hospital mortality based on a review of 2000 hospital deaths per year will result in both invalid and unreliable rankings. In spite of the limitations of HSMR and similar metrics, they do not suffer from the issues described here. For example, when estimating the HSMR value for a given hospital, no attempt is made to classify individual deaths as preventable or not; rather, the total number of deaths is simply compared with the expected number given the case-mix of patients. Further, given all deaths are counted, small number issues are minimised. Given known methodological limitations of HSMR use, this leaves the improvement community and policy-makers in an uncertain, in fact uncomfortable, place. It should be acknowledged that the very notion of case reviews may be beneficial in itself. For example, it can help to engage hospital leaders on reflecting about their own performance, and incentivise and motivate local improvements. It may also be useful for determining a useful national benchmark. Therefore, this new initiative may result in quality improvements; it will, however, remain a grossly inadequate measure for judging comparative hospital performance in respect of avoidable mortality.

Contributors GA conceived the original argument and presented and drafted the manuscript. GL helped develop the argument and contributed to the writing and editing of the manuscript.

Funding GA received no specific funding for this piece of work. GL is supported by a Cancer Research UK Clinician Scientist Fellowship award (A18180).

Competing interests None declared.

Provenance and peer review Not commissioned; internally peer reviewed.

Open Access This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: http://creativecommons.org/licenses/by/4.0/

REFERENCES

Ranking hospitals on avoidable death rates derived from retrospective case record review: methodological observations and limitations

Gary Abel and Georgios Lyratzopoulos

BMJ Qual Saf published online July 3, 2015

Updated information and services can be found at:
http://qualitysafety.bmj.com/content/early/2015/07/03/bmjqs-2015-004366

These include:

References
This article cites 14 articles, 7 of which you can access for free at:
http://qualitysafety.bmj.com/content/early/2015/07/03/bmjqs-2015-004366#BIBL

Open Access
This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See:
http://creativecommons.org/licenses/by/4.0/

Email alerting service
Receive free email alerts when new articles cite this article. Sign up in the box at the top right corner of the online article.

Topic Collections
Articles on similar topics can be found in the following collections
Open access (268)

Notes

To request permissions go to:
http://group.bmj.com/group/rights-licensing/permissions

To order reprints go to:
http://journals.bmj.com/cgi/reprintform

To subscribe to BMJ go to:
http://group.bmj.com/subscribe/