Composite performance indicators: bringing uncertainty out into the open

Robert Grant

Faculty of Health and Social Care Sciences, Kingston University & St George's, University of London

Background

Despite a sceptical public and unresolved academic debate, interest in rating and ranking public service providers continues. In healthcare, the UK coalition government has shifted emphasis from process to outcome[1] but familiar methodological problems remain, typified by the disparity in ratings given to Mid-Staffordshire NHS Foundation Trust by the Care Quality Commission and Dr Foster[2]. The problems inherent in composite performance indicators have not been resolved since a report for the Royal Statistical Society warned about subjectivity and uncertainty in their composition[3]. The extent of uncertainty needs to be openly discussed during construction; it arises from sampling error as well as the range of possible formulas to create the composite[4,5]. In the case of multilevel data such as patients within hospitals, two distinct designs have been defined[6]: the composite can be formed either by summarising across the indicators and then the patients, or vice versa. I contend that well-constructed and communicated composite indicators are a positive contribution in making official statistics accessible; done badly they can obscure or misrepresent the facts.

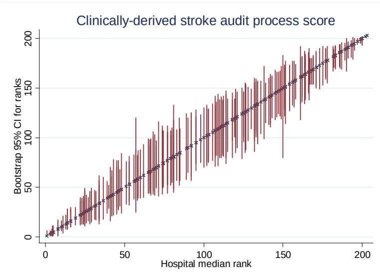
Data and methods

Anonymous data on 26 dichotomous measures of the quality of care received in 203 NHS hospitals by 10,617 people admitted following a stroke were supplied by the Royal College of Physicians of London from the national clinical audit of stroke 2008. Three composite indicators are compared:

- The "process score" used in the audit, derived from clinical consensus with almost equal weights
- An alternative summarising indicators then patients
- An alternative derived from a novel method[7] (multilevel principal components analysis) to capture maximum variance between hospitals.

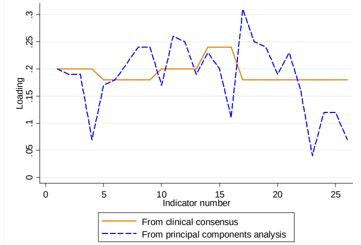
Sampling error

Sampling error can be addressed even in complex indicators by bootstrapping, stratified by hospital. Inference can be made for both raw scores and hospital ranks, but ranking is implicitly a discontinuous transformation of the scores, and this can lead to unstable confidence intervals. However, the extent and overall pattern of uncertainty can be shown by a graph such as this. The pattern of decreasing uncertainty at the extremes of high and low performance is characteristic. The extent of uncertainty in these data make it possible only to give broad classifications of hospital performance, and a league table would be highly misleading.



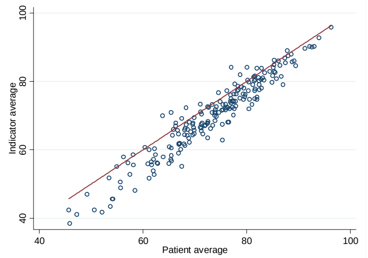
Choice of weights

Monte Carlo simulation can be used to assess the impact of changing weights (standardised to match PCA loadings so that their squares sum to one, see graph) between those focussed on clinical validity, those that capture maximum variance, and anything in between. Scores and ranks can be saved from each simulation producing a plot matching the bootstrap above. However, care needs to be taken over the choice of distribution and covariance structure for the "pseudo-loadings". This contributes far less uncertainty to the stroke audit than sampling error, with mid-rank hospitals typically moving by 10 ranks compared to 60 ranks (2.5 and 97.5 percentiles).



Order of averaging (multilevel data)

The two options are described by Reeves and colleagues as "patient average" and "indicator average". In the former, a score is given to each patient that summarises the indicators, and then the hospitals are each given the average patient score. In the latter, each hospital has its performance on each of the indicators calculated as a percentage of patients, and then these are averaged. In the stroke data, hospital scores and ranks differ notably between these, particularly among poorly-performing hospitals.



More than one dimension

The uncertainty in a uni-dimensional composite may be so large that retaining more dimensions, such as rotated principal components, is preferred. Bootstrapped replicates of these will differ by sampling error but also in the alignment of the axes, which is not of interest. Procrustes analysis re-aligns each hospital's scores as closely as possible through orthogonal rotations. The remaining unexplained variation is of interest. Residuals can be plotted to show the most sensitive hospitals, and an overall residual sum of squares allows for comparison of the size of effects.

If the number of clusters (hospitals) is small, plots of bootstrap replicate scores in two dimensions can be shown with convex hulls which "peel" away the outer 5% of points to give a non-parametric confidence polygon. Uncertainty in two-dimensional ranks can be shown clearly - given the usual caveats about overlapping confidence intervals.

Conclusions

These analyses show how it is possible to achieve an open representation of uncertainty and that graphs can aid discussion. These analyses provide insight into the potential weaknesses of any composite indicator but human judgement is still essential to make the final formula. This has potential to aid the construction of robust, valid summary measures that will be more widely accepted by experts and the public. The corollary is that where no adequate summary can be formed for particular data, this will be made apparent and misleading analyses and spuriously certain conclusions will be avoided.

In the stroke audit data, sampling error is the largest source of uncertainty. Few hospitals had no valid data and so there is little scope for imputing missing data, though this is another source of uncertainty. Adjusting for covariates is another, though here it is not useful for process measures, but can easily be incorporated in the bootstrap[7].

References

1. "Equity and excellence: liberating the new NHS". HMSO, London 2010. ISBN 9780101788120.
2. 2. Thompson, G. "Indicators of hospital performance published by the Care Quality Commission and Dr. Foster Research". House of Commons library, 2009
3. Bird SM, Cox D, Farewell VT, Goldstein H, Holt T, Smith PC. "Performance indicators: good, bad and ugly". *Journal of the Royal Statistical Society - Series A* (2005); **168**(1): 1-27.
4. 4. Jacobs R, Smith PC, Goddard M. "Measuring performance: An examination of composite performance indicators". *University of York Centre for Health Economics Technical Paper 29* (2004).
5. Nardo M, Saisana M, Saltelli A, Tarantola S, Hoffman A, Giovannini E. "Handbook on constructing composite indicators: methodology and user guide". *OECD Statistics Working Papers*, 2005/3, OECD.
6. Reeves D, Campbell SM, Adams J, Shekelle PG, Kontopantelis E, Roland MO. "Combining multiple indicators of clinical quality: an evaluation of different analytic approaches" *Medical Care* (2007); **45**(6): 489-496.
7. Grant, RL. "Using a multilevel principal components analysis to summarise inter-cluster variance, with application to composite performance indicators and quality of life scale" Submitted to *Statistics in Medicine* 2011.

Acknowledgments

I am grateful to the stroke team at the Royal College of Physicians' Clinical Standards department for providing anonymised data for these analyses, in particular Geoff Cloud, Alex Hoffman, and Tony Rudd. Thanks are also due to Nick Black and Chris Frost of the London School of Hygiene and Tropical Medicine for their advice.

E-mail: robert.grant@sgul.kingston.ac.uk
Website: staffnet.kingston.ac.uk/~ku45386

St George's University of London

Kingston University London