View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Apollo

Beyond the ecological fallacy: potential problems when studying healthcare organisations

Catherine L Saunders[1,2]

Marc N Elliott[3]

Georgios Lyratzopoulos[1,4]

Gary A Abel[1]

1. Cambridge Centre for Health Services Research, University of Cambridge, Cambridge, UK

2. RAND Europe, Cambridge, UK

3. RAND Corporation, Santa Monica, CA, USA.

4. Department of Epidemiology & Public Health, Health Behaviour Research Centre, University College London, London, UK.

Main text: 1994

Abstract: 69

Figures: 2

Tables: 1

Authors for correspondence:

Dr Catherine L Saunders

Email: csaunder@rand.org

Funding statement: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. GL is supported by a Cancer Research UK Clinician Scientist Fellowship (A18180). The views expressed in this publication are those of the authors and not necessarily those of any funder or any other organisation or institution.

Contributorship statement: All authors (CS, ME, YL, GA) contributed to the initial planning of this paper. CS wrote the initial draft and all authors contributed to the development and revision of the arguments and the discussion. CS is the guarantor for the work.

Competing interest statement: The authors have no competing interests to declare.

Abstract

Ecological studies, which consider patient groups rather than individuals, are common in health policy research. The 'ecological fallacy' is a well-recognised methodological concern, but in this perspectives paper, we focus on less often appreciated but equally important limitations of such studies. In particular, we consider reliability and power as they apply to ecological studies, and make recommendations to inform the appropriate design and interpretation of these increasingly popular studies.

Introduction

Ecological studies, which consider patient groups rather than individuals, are popular in American and European health services and policy research.(1-5) Such studies often correlate aspects of quality, or characteristics of health care organisations, examining associations such as those of patient experience with care quality,(4,5) or hospital mortality with competition.(6,7) Unfortunately, these quick, inexpensive studies may influence policy or clinical practice disproportionately to their methodological rigour.

The 'ecological fallacy' (the fact that associations between measures at the person level may differ from associations at an aggregated level, e.g. hospitals, providers) is a well-known concern about such studies.(8) Here we argue that the 'ecological fallacy' is only one of several important methodological concerns, some of which are less often appreciated. These concerns may stem from a failure to recognise the distinction between healthcare organisations and individual patients as different levels of analysis.(9) In particular, we consider measurement reliability and power as they apply to the design and interpretation of findings from studies of healthcare organisations.

Measurement reliability of organisational characteristics

The statistical concept of reliability can be applied in many situations where we are concerned with data reproducibility, and the degree to which observations are influenced by measurement error. It is increasingly recognised that the observed performance of healthcare organisations (or providers) is subject to chance variation. Often this chance variation is visualised with the aid of funnel plots, which provide a graphical summary of the 'scatter' of performance estimates for different organisations and the degree to which they reflect chance variation.(10) It can be quantified by estimating the unit-level, or Spearman-Brown reliability; in this contexts 'units' can be hospitals, health centres or individual physicians (e.g. general practitioners or surgeons) within which different individual patients are clustered. Unit-level reliability is a specific form of reliability suited to organisational characteristics that must be inferred by sampling and measurement within organisations.(11) We provide formulas for unit-level reliability in Box 1.

Some measures, particularly administratively-measured organisational characteristics such as the number of beds in a hospital, are likely to have minimal error. However, for measures that aggregate sampled patient-level data to the organisation-level, as many clinical quality measures do, the unit-level reliability may be substantially lower. Because these quality measures are interpreted as informing the likely experiences of a future patient at a given hospital, their precision is limited by the number of responding patients. This concern is clear when only some patients are sampled (e.g. for a patient survey); however even if all patients were measured for a given hospital, a degree of error will be present in these organisation-level measures due to statistical noise (chance) alone.(12) Unit-level reliability can be defined as the proportion of the total variance in measured organisational-level scores attributable to true variation among organisations.(11)

The unit-level reliability of an aggregated measure is determined by three factors. First, reliability increases as true organisational-level variability in the measure increases, measured by the intraclass correlation coefficient. Second, measurement error decreases and reliability increases with greater sample sizes/organisation. Third, measurement error decreases and reliability increases when measures have a lower patient-level variance. For binary (yes/no) performance indicators, this variance is highest and reliability is lowest for indicators with a frequency near 50%.

Because reliability depends on the degree of variability between organisations, it may be very context-specific. In England, most NHS hospitals are relatively similar to each other, in terms of size, spectrum of clinical services and specialties provided, staff training, recruitment and remuneration policies, patient case-mix, etc. This contrasts sharply with the US, where much greater variability is observed between different hospitals in respect of all these factors. Measures that may be reliably measured in a heterogeneous setting, e.g. all US hospitals, may not be reliable in a more homogeneous setting, such as English hospitals, even when similar numbers of observations are used per organisation.

The most important consequence of using measures with imperfect reliability is the attenuation of estimated effect sizes. An observed correlation coefficient between two measures will be attenuated (biased towards0) if either has reliability <1. Specifically, it will be attenuated by the product of the square roots of their reliabilities.(13) For example, weaker effect sizes typically observed for organisation-level correlations for binary performance indicators can be attributed to their lower reliability, rather than truly lower correlations.

A related issue is that ‘proxy’ measures may not in fact reliably measure the intended construct. When this is the case correlations of the underlying constructs will be underestimated. As an example, the presence of a CT scanner is a hospital characteristic which can be measured directly with little error. Nonetheless, correlations between the presence of CT scanning facilities and health outcomes might understate the correlation between CT access and health outcomes if the presence of a scanner is not a reliable and valid measure of the intended construct ease of patient access to CT investigations.

Power considerations

In studies considering the correlation between healthcare organisation characteristics, although the number of individuals included in each organisation determines the unit-level reliability of the organisation-level measure, it is the number of organisations, rather than the number of individuals which provides the most relevant sample size for the correlation. For example, the English General Practice Patient Survey currently comprises ~1 million patient responses but when using Clinical Commissioning Group average scores, the relevant sample size is the number of organisations, i.e. ~200.

In some instances, even organisation-level sample sizes are large. Studies of US hospitals will have a possible maximum sample size of up to 4800;(14) and studies of all English general practices have a sample size of about 8000.(15) However, in studies of English NHS acute hospitals, the organisational sample size upper bound is about 160,(15) and for hospitals in the Netherlands, less than 100.(16) Unlike studies of individual patients, where recruitment can be increased, the sample sizes of healthcare organisations in a geographical region or country are fixed.

As well as sample size we must also consider the true magnitude of associations being examined. For example, organisation-level correlations between two different measures of the same underlying construct (e.g. two measures of patient experience), are typically moderate, with correlation coefficients ≤0.45.(2) However, organisation-level correlations between different dimensions of healthcare quality can be much weaker. In a large English primary care study correlations between clinical quality measures and patient experience were positive but small, often ≤0.1.(5)

We use standard sample size calculations(17) to illustrate the relationship between power and effect size (here magnitude of correlations) in ecological studies, initially without further considering the role of measurement error. Figure 1 illustrates, in studies of 100 healthcare organisations, there is 80% power to detect (positive or negative) correlations of ≥0.28, in studies of 160 there is 80% power for correlations of ≥0.22 and in studies of sample size 1000 there is 80% power to detect correlations of ≥0.09. Where both the sample size and expected effect sizes are small, studies may be markedly underpowered. For example, studies of correlations near 0.1 in English NHS hospitals will be very underpowered (~50% power, i.e. a 50% type II false negative error rate).

These calculations assume all variables are measured without error. With less than perfect reliability, the observed correlation is attenuated. Such a situation requires a larger sample size of organisations to be adequately powered to detect the same true (as opposed to observed) association, than would have been needed if the organisation-level score had been measured without error (Figure 2). The number of required organisations approximately doubles for correlations between two measures, if both have reliability of 0.7 (often considered a level adequate for evaluating healthcare organisation performance, and so might typically be seen in studies using performance indicators) (Figure 2). In practice this would mean 320, rather than 160 hospitals being needed for a study to have 80% power to detect a true correlation of 0.22, or in 160 hospitals a study would have 80% power only for a true correlation of 0.31 or higher.

Consequences of underpowered studies

Statistically significant results from underpowered correlational studies of healthcare organisations will underrepresent small associations. For example, correlations between hospital characteristics and mortality seen in a large sample of US hospitals (1) are not replicated when translated to a sample of <100 hospitals in the Netherlands.(3) There is uncertainty around non-significant results from small studies and we do not know whether weak, non-significant findings in such studies are true negative findings or simply reflect inadequate power. In a study of 160 hospitals an observed correlation coefficient of zero (i.e. no association) is compatible with a true correlation between -0.16 and 0.16 (based on 95% confidence intervals, Table 1) when both measure have perfect reliability, and with a range of -0.23,0.23 if each measure has reliability of 0.7. That is, finding a correlation of zero in a study of English NHS hospitals does not rule out correlations of sizes typically seen in these studies. In a study of 1000 healthcare organisations observed correlations greater than 0.07 will be significant at $p<0.05$, but in a study of sample size 100 only observed correlations greater than 0.20 will be significant (Table 1). Apparent inconsistencies across settings may simply reflect failure to detect similar true associations in a less-powerful study.

Ease of access to data, multiple testing, and publication bias

The availability of publicly-reported data on health organisations is rapidly expanding.(14,15) For researchers, this organisation-level data is often free of the data governance and confidentiality issues that apply to using person-level data, and does not require lengthy and expensive primary data collection.

Under these circumstances it is easy to explore large numbers of hypotheses and correlations, raising concerns about multiple testing. For example, with ~1600 indicators available on the NHS indicator portal(15) there are >2.5 million possible hypotheses about correlations among them. If all indicators were independent, with a nominal p-value of 0.05, we would expect 125,000 false positive (type I error) findings from this data source alone. Even if an individual researcher is not performing multiple hypothesis tests or adequately accounts for doing so,(18) the public availability of data means that among all those performing research many tests will be carried out.

Further, selective reporting (when investigators opt not to report negative findings), and publication bias (the selection of papers with positive findings during the publication process) result in over-representation of papers with statistically significant results in the published literature, independent of the work's methodological strength. Large effect sizes can be eye-catching, but because statistically significant findings from small studies can only have very large effect sizes (table 1), false positive findings from small studies may be particularly over-represented. Both publication bias and selective reporting are more likely where true typical effect sizes are small, where there is a large number and less pre-selection of relationships to be tested, and where many teams are involved in similar studies;(19) all these conditions are present in studies of healthcare organisation-level correlations using publicly available data. The development and use of guidelines for the design and reporting of ecological studies could help improve research and editorial practice in this area, such as under the auspices of the RECORD initiative.(20)

Conclusions

Policy makers should exercise caution when making decisions based on the results of correlation studies of healthcare organisation performance or characteristics. Where the number of healthcare organisations is small, where expected associations are weak (such as for correlations between different dimensions of quality), or where measures are not reliable, studies are often underpowered and small associations may be undetectable. Null findings need to be carefully interpreted; the failure to replicate findings of large studies in smaller sample sizes and/or less diverse settings may reflect lack of power in the replication study, rather than an incorrect finding in the original. Further, undue emphasis should not be given to those results that are statistically significant, particularly large effect sizes from studies of small numbers of organisations. Following some simple recommendations for best practice (Box 2) will improve the translation of appropriate and robust research findings into healthcare policy and practice.

Box 1: Spearman-Brown reliability

Unit-level, or Spearman-Brown reliability is defined as the proportion of the total variance in measured organisational-level scores which is attributable to true variation between organisations (organisation-level variance), and can be estimated using the formula below. The term within-organisation variance used here is also sometimes known as the residual variance; “n” is the mean achieved sample size per organisation.

$$\text{Reliability} = \frac{\text{between organisation variance in measured scores}}{\text{between organisation variance} + \left(\frac{\text{within organisation variance}}{\text{n}}\right)}$$

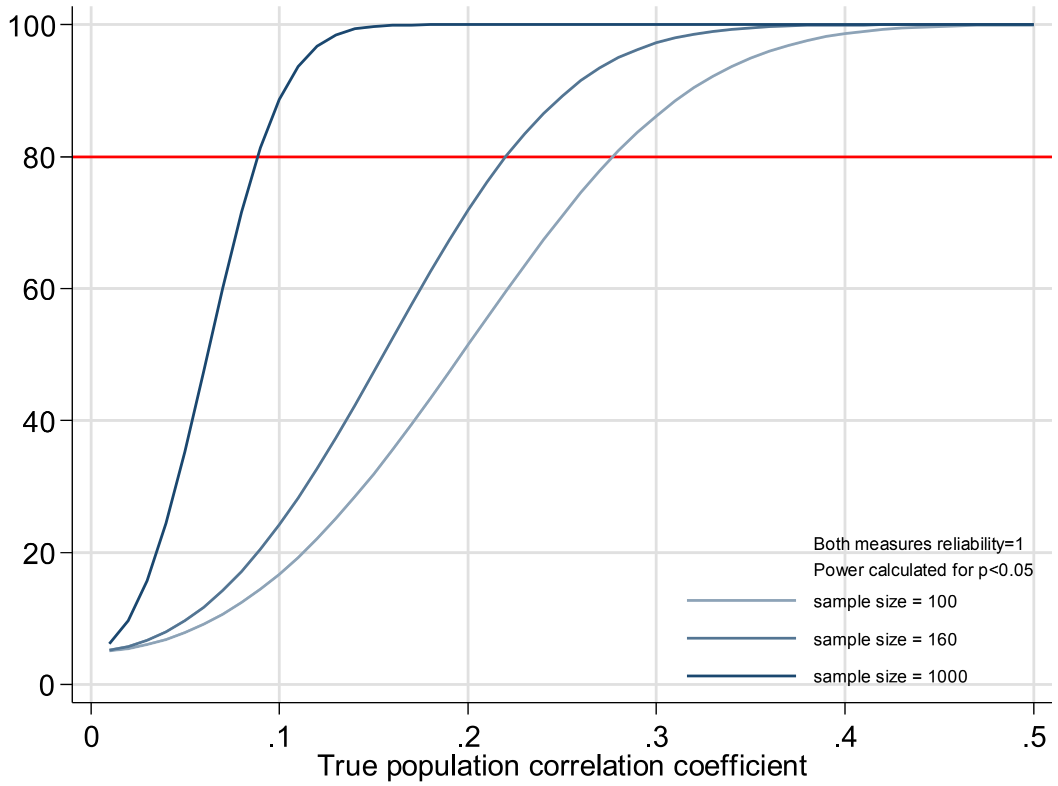
Spearman-Brown reliability is the intraclass correlation coefficient (ICC) when the number of observation or the sample size within each organisation is 1

$$\text{ICC} = \frac{\text{between organisation variance in measured scores}}{\text{between organisation variance} + \text{within organisation variance}}$$

By rearranging the above formula, unit-level reliability can therefore be estimated from the ICC and sample size as follows:

$$\text{Reliability} = \frac{ICC \times n}{1 + (n-1) \times ICC}$$

Figure 1. Power to detect correlations in samples of 100, 160 and 1000 healthcare organisations



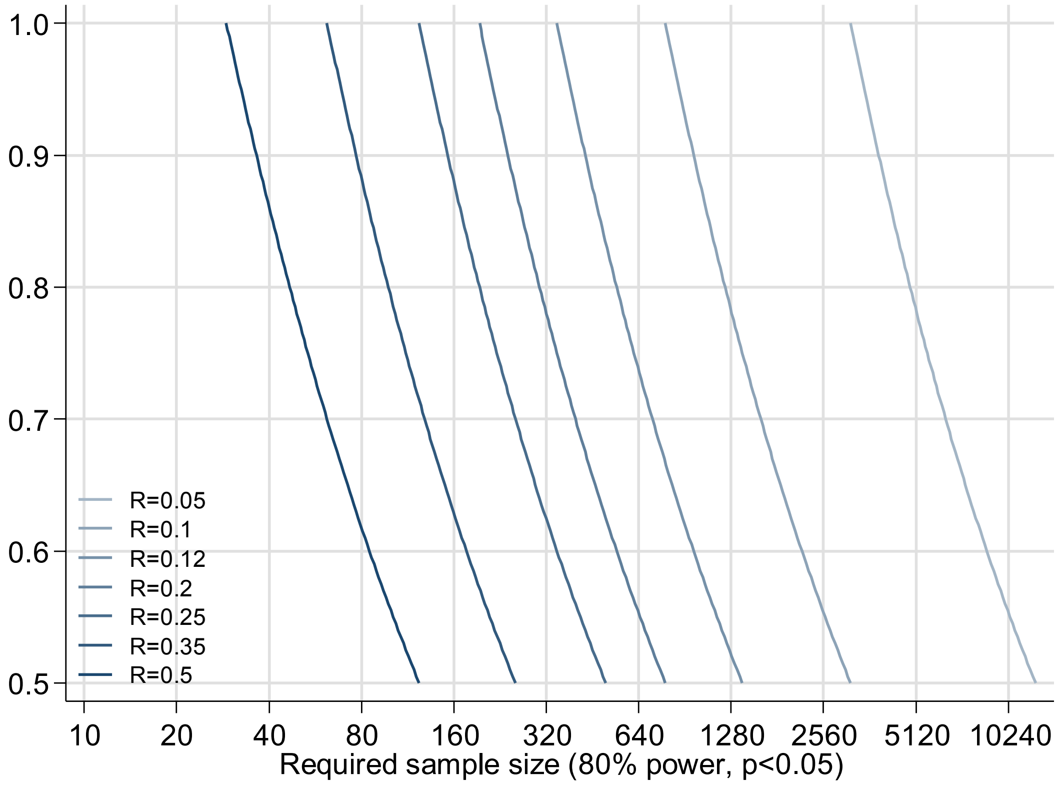
As the strength of the correlation to be detected increases, the power at all sample sizes increases, but 80% power is only possible at a higher true correlations when sample sizes are smaller. Illustrations are provided for sample sizes of 100 organisations (the approximate number of hospitals in the Netherlands) 160 organisations (applicable to the number of acute NHS hospitals in England) and 1000 (a possible sample size in studies of subsamples of general practices in England or hospitals in the US) is calculated for a range of true correlations.

Table 1. Margin of error for an observed correlation coefficient of zero

Sample size*	Margin of error (95% confidence interval for an observed correlation coefficient of zero)	Minimum observed correlation coefficient at which $p<0.05$
2000	- 0.04 to 0.04	0.04
1000	- 0.07 to 0.07	0.07
160	- 0.16 to 0.16	0.16
100	- 0.20 to 0.20	0.20
50	- 0.28 to 0.28	0.28

*for an ecological study of healthcare organisations, this would be the number of organisations, not the number of individuals

Figure 2. Required sample size (number of organisations) to detect a true population correlation where measurement (Spearman-Brown) reliability is less than 100%



In this figure, ‘R’ is the true correlation in the population. Where reliability is less than 1.0, required sample sizes are calculated assuming that the measurement reliability of both characteristics and / or performance measures considered for the correlation study is at the stated level. This calculation was made by multiplying the true correlation by the square root of the reliability of each measure and calculating the required sample size for this attenuated correlation. Note that the required sample size axis is on a log scale.

For measures of healthcare organisation quality which are used for ‘high stakes’ applications (such as for pay-for-performance or public reporting schemes) the measurement reliability is typically required to be between 0.7-0.9.(11)

Box 2: Recommendations for best practice

- Ecological or organisation-level correlation studies often provide only weak research evidence, and only those with adequate reliability, power, and validity should substantially influence policy

- Where patient-level associations are of interest, organisation-level analyses may not be the right approach. Where organisation-level analysis is performed it is important that this is stated or identified directly

- Regarding unit-level (Spearman-Brown) reliability, there need to be enough patients included at each organisation, and enough variation between organisations, for a reliable measure.

- The total number of organisations rather than individual patients is the most relevant sample size for the analysis of organisation-level associations. The data illustrated in figures 1 and 2 can be used as ‘ready reckoners’ by researchers and policy makers when designing or interpreting the findings of such studies

- The way that the organisation characteristics or performance are measured is also important. The construct validity and unit-level reliability of what is being measured are important to consider, and can also influence the power of the study.

- It is important to report any multiple testing and exploratory analyses as well as significant findings, particularly when using publicly available data for which multiple possible correlations could be considered

Reference

1. Fisher ES, Wennberg JE, Stukel TA, Skinner JS, Sharp SM, Freeman JL, et al. Associations among hospital capacity, utilization, and mortality of US Medicare beneficiaries, controlling for sociodemographic factors. Health Serv Res. 2000;34(6):1351-62.

2. Greaves F, Laverty AA, Millett C. Friends and family test results only moderately associated with conventional measures of hospital quality. BMJ. 2013;347:f4986.

3. Heijink R, Koolman X, Pieter D, van der Veen A, Jarman B, Westert G. Measuring and explaining mortality in Dutch hospitals; the hospital standardized mortality rate between 2003 and 2005. BMC Health Serv Res. 2008;8:73.

4. Lehrman WG, Elliott MN, Goldstein E, Beckett MK, Klein DJ, Giordano LA. Characteristics of hospitals demonstrating superior performance in patient experience and clinical process measures of care. Med Care Res Rev 2010;67(1):38-55.

5. Llanwarne NR, Abel GA, Elliott MN, Paddison CA, Lyratzopoulos G, Campbell JL, et al. Relationship between clinical quality and patient experience: Analysis of data from the English Quality and Outcomes Framework and the National GP Patient Survey. Ann Fam Med. 2013;11(5):467-72.

6. Bloom N, Cooper Z, Gaynor M, Gibbons S, Jones S, McGuire A, et al. In defence of our research on competition in England's National Health Service. Lancet. 2011;378(9809):2064-5.

7. Pollock A, Macfarlane A, Kirkwood G, Majeed FA, Greener I, Morelli C, et al. No evidence that patient choice in the NHS saves lives. Lancet. 2011;378(9809):2057-60.

8. Finney JW, Humphreys K, Kivlahan DR, Harris AH. Why health care process performance measures can have different relationships to outcomes for patients and hospitals: understanding the ecological fallacy. Am J Pub Health 2011;101(9):1635-42.

9. Anhang Price R, Elliott MN, Zaslavsky AM, Hays RD, Lehrman WG, Rybowski L, et al. Examining the role of patient experience surveys in measuring health care quality. Medical care research and review : MCRR. 2014;71(5):522-54.

10. Spiegelhalter DJ. Funnel plots for comparing institutional performance. Statistics in medicine. 2005;24(8):1185-202.

11. Lyratzopoulos G, Elliott MN, Barbiere JM, Staetsky L, Paddison CA, Campbell J, et al. How can health care organizations be reliably compared?: Lessons from a national survey of patient experience. Med Care 2011;49(8):724-33.

12. Elliott MN, Zaslavsky AM, Cleary PD. Are Finite Population Corrections Appropriate when Profiling Institutions? Health Serv Outcomes Res Met. 2006;6(304):153-6.

13. Muchinsky PM. The correction for attenuation. Educ Psychol Meas. 1996;56(1):63-75.

14. The Medicare Quality Compare Database http://www.medicare.gov/hospitalcompare/search.html Accessed 19/12/2014.

15. The Health and Social Care Information Centre Indicator Portal https://indicators.ic.nhs.uk/ Accessed 19/12/2014.

16. 94 acute care hospitals in the Netherlands www.ihf-fih.org from the international hospital federation in the Netherlands Accessed 19/12/2014.

17. Mander AP. SAMPSI_RHO: Stata module to compute sample size for a Pearson correlation, revised 12 Apr 2011. 2006.

18. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc B. 1995;57(1):289-300.

19. Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2(8):e124.

20. Nicholls SG, Quach P, von Elm E, Guttmann A, Moher D, Petersen I, Sørensen HT, Smeeth L, Langan SM, Benchimol EI. The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement: Methods for Arriving at Consensus and Developing Reporting Guidelines. PLoS One. 2015;10(5):e0125620.