Social Science & Medicine 162 (2016) 59–67

Contents lists available at ScienceDirect

Social Science & Medicine

journal homepage: www.elsevier.com/locate/socscimed

Developing robust composite measures of healthcare quality – Ranking intervals and dominance relations for Scottish Health Boards

Laura Schang [a, b, *], Yrjänä Hynninen [c], Alec Morton [d], Ahti Salo [c]

[a] Department of Management, London School of Economics and Political Science, London, United Kingdom
[b] Department of Health Services Management, Ludwig-Maximilians-Universität München, Schackstraße 4, 80539 Munich, Germany
[c] Department of Mathematics and Systems Analysis, Systems Analysis Laboratory, Aalto University School of Science, Aalto, Finland
[d] Department of Management Science, Strathclyde Business School, University of Strathclyde, Glasgow, United Kingdom

ARTICLE INFO

Article history:
Received 24 August 2015
Received in revised form
24 May 2016
Accepted 15 June 2016
Available online 16 June 2016

Keywords:
Performance comparison
Composite indicator
Weight
Denominator
Ranking interval
Dominance relation

ABSTRACT

Although composite indicators are widely used to inform health system performance comparisons, such measures typically embed contentious assumptions, for instance about the weights assigned to constituent indicators. Moreover, although many comparative measures are constructed as ratios, the choice of denominator is not always straightforward. The conventional approach is to determine a single set of weights and to choose a single denominator, even though this involves considerable methodological difficulties.

This study proposes an alternative approach to handle incomplete information about an appropriate set of weights and about a defensible denominator in composite indicators which considers all feasible weights and can incorporate multiple denominators. We illustrate this approach for comparative quality assessments of Scottish Health Boards. The results (displayed as ranking intervals and dominance relations) help identify Boards which cannot be ranked, say, worse than 4th or better than 7th.

Such rankings give policy-makers a sense of the uncertainty around ranks, indicating the extent to which action is warranted. By identifying the full range of rankings that the organizations under comparison may attain, the approach proposed here acknowledges imperfect information about the "correct" set of weights and the appropriate denominator and may thus help to increase transparency of and confidence in health system performance comparisons.

1. Introduction

The increasing complexity of health systems and the multidimensionality of health system performance have reinforced calls for the production of composite measures of performance (WHO, 2000; Healthcare Commission, 2005; Carinci et al., 2015). Summarizing the information contained in diverse indicators in a single index and ranking organisations or countries on that basis has the potential to present the "big picture", by highlighting in a unified way to what extent the objectives of health systems related to health outcomes, treatment appropriateness, and other dimensions have been met. Thus, composite measures may seem an attractive approach to strengthen accountability, facilitate communication with the public, and focus improvement efforts on poorly performing organisations (Goddard and Jacobs, 2009).

However, composite indicators also have important disadvantages. In contrast to assessing performance based on a range of separate indicators, rankings based on aggregate measures may disguise the sources of poor performance and thus obscure the best focus for remedial action (Smith, 2002). Composite indicators are also highly sensitive to methodological choices, in particular to the weights attached to constituent indicators (see e.g. Jacobs et al., 2005; Reeves et al., 2007; OECD, 2008). In their analysis of hospital performance based on star ratings in the English NHS, Jacobs et al. (2005) show, for instance, how subtle changes in the weighting system lead some hospitals to jump almost half of the league table. However, the techniques by which weights are determined are not straightforward. In addition, although many comparative quality measures are constructed as ratios, it is not necessarily obvious which indicators should be employed as denominators (Schlaud et al., 1998). In the context of low-birthweight

* Corresponding author. Department of Health Services Management, Ludwig-Maximilians-Universität München, Schackstraße 4, 80539 DE Munich, Germany.
E-mail address: schang@bwl.lmu.de (L. Schang).

http://dx.doi.org/10.1016/j.socscimed.2016.06.026
0277-9536/

survival rates, Guillen et al. (2011) illustrate how the choice of population denominator results in considerable variation depending on whether survival is reported relative to all births; live births; or neonatal intensive care unit admissions.

These concerns are critical especially when rankings have serious consequences for the rankees. For example, six of the Chief Executives of the twelve lowest ranked hospitals in England's star rating system (the so-called "dirty dozen") lost their jobs as a result (Bevan and Hamblin, 2009). It has been argued that France and Spain's apparently high ranking in the WHO's 2000 assessment of health systems substantially diminished pressure for reform in these countries (Navarro, 2000). In Medicare's Premier Hospital Quality Incentive Demonstration, a pay-for-performance scheme based on a composite quality score, hospitals below the ninth decile faced a 2% deduction in their Medicare payment (CMS, 2009). With such high stakes, understanding whether ranks are robust to alternative assumptions seems critical.

This study proposes an alternative approach to handle the lack of information about an appropriate set of weights and about a defensible denominator in composite indicators. We make two main contributions. First, we demonstrate the use of an approach to ranking organisations based on ranking intervals and dominance relations which accounts for the full set of feasible weights. This avoids the need to settle on a single, potentially controversial set of weights as it is required for instance in data envelopment analysis (DEA), in which weights are chosen such that each organisation appears in its best possible light (Cherchye et al., 2007). Feasible weights are less restrictive and thus potentially better able to increase transparency and to acknowledge imperfect information about the "correct" set of weights. The ranking intervals obtained with this approach can be said to be robust in the sense that they reflect the full range of rankings that the organizations under comparison may attain when weights are selected from their respective feasible weight sets. Second, we address the problem of choice of denominator in ratio-based measures of performance.

2. Challenges in developing composite indicators of healthcare quality

A composite indicator is commonly expressed as an additive model based on a weighted sum of a set of performance indicators

$$C_k = \sum_{j=1}^{J} w_j \, x_{jk}, \tag{1}$$

where J is the number of constituent indicators, w_i is the weight attached to indicator j, and x_{jk} the score on indicator j for organisation k. Composite measures of this form require choices about (i) the indicators included; (ii) the methods used to transform indicators (to achieve a common unit of measurement); (iii) the weights applied; (iv) any aggregation rules used; and (v) adjustments for environmental influences on performance. In addition (vi), although many quality indicators are reported as ratios, the choice of denominator is not always straightforward.

The focus of this study is on problems (iii) and (vi), how to handle incomplete information about weights and about the choice of denominator. Below we review the conceptual background and problems with conventional strategies to address these challenges. In the empirical application, we explain the approaches taken to problems (i), (ii), (iv) and (v).

2.1. Valuation of multiple healthcare quality measures

Healthcare performance is multidimensional. However, without a functioning market, there is no price mechanism for comparison. To aggregate heterogeneous indicators into a summary measure of performance, weights are required which – analogous to prices – should represent the opportunity cost of achieving improvements on each individual measure by capturing the relative value attached to an extra unit of it (Smith, 2002).

In practice, arriving at explicit trade-offs between different healthcare quality measures – and thus exact specifications of weights – is highly contentious. First, it is often unclear *whose* preferences should be elicited. Weights used often reflect a single set of preferences, although the evidence suggests substantial heterogeneity in preferences between and within groups of policymakers, patients and the public (Smith, 2002; Decancq and Lugo, 2012). Making precise judgments about the relative value of subindicators to the composite is typically both politically controversial and cognitively demanding, thus triggering reluctance among respondents to agree on a set of weights.

Second, there is no consensus on a single best method *how* to elicit weights. Different techniques for valuing health(care) outcomes – from simpler trade-off methods including ranking from most to least desired indicator and voting techniques to elaborate multi-attribute approaches such as conjoint analysis and the analytic hierarchy process – tend to produce different results. Each method has distinct advantages and disadvantages in terms of feasibility, consistency and validity (Dolan, 1997; OECD, 2008; Appleby and Mulligan, 2000).

To circumvent perceived difficulties with normative approaches to set weights, data-driven weighting systems are frequently used. For example, in data envelopment analysis (DEA) – a widespread method to compare organisations with multiple outputs and inputs (Hollingsworth and Street, 2006) – weights are derived from the data so as to maximise each organisation's performance (Cherchye et al., 2007). Each organisation receives a different set of weights which casts it in the best possible light. However, data-driven weights do not necessarily reflect meaningful trade-offs between performance domains (Decancq and Lugo, 2012). There is no logical reason why an organisation values most some performance domain because it performs relatively well on it: data-driven approaches thus purport to solve a deep philosophical problem of how to derive values from facts (Hume, 1739).

The conventional recommendation to address incomplete information about weights, and about the best method to elicit weights, is to conduct extensive sensitivity analysis on the chosen weights (Jacobs et al., 2005). However, traditional sensitivity analysis is problematic insofar as the choice of ranges of weights depends on the analyst. This form of sensitivity analysis thus corresponds to a "blind search" which is not explicitly oriented towards changes in ranks and the maximum and minimum plausible ranks an organisation can attain.

2.2. Choice of denominators

Healthcare quality measures are often reported as ratio measures where a specific quality measure is divided by some measure of population. Not all comparative assessments of healthcare quality require a denominator. So-called "never events", events which are deemed to be entirely preventable, are reported as absolute numbers without reference to a denominator (NHS England, 2015). However, typically a ratio-based measure is used in order to make entities of different sizes comparable and to establish a common "currency unit" in which performance is assessed as "good" or "poor" relative to other organisations.

To construct ratio-based quality measures, the denominator should represent the best available proxy for the population at risk (PAR) (Romano et al., 2010). However, the PAR of experiencing a