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

Laura K Schang, Yrjänä Hynninen, Alec Morton, Ahti Salo

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 the lack of 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.

Key words: performance comparison; composite indicator; weight; denominator; ranking interval; dominance relation.

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, CMS, 2009, 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. As such, 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 unlikely to be 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 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 set of indicators included; (ii) the methods used to transform the constituent indicators (in order to achieve a common unit of measurement); (iii) the weights applied; (iv) any specific aggregation rules used; and (v) potential adjustments for environmental or other uncontrollable influences on performance. In addition (vi), although many healthcare quality indicators that are used to construct a composite indicator 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 a lack of information about the appropriate set of weights and about the choice of denominator. Below we set out 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 measures are 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 policy-makers, patients and the public (Smith, 2002, Decancq and Lugo, 2012). Making precise judgments about the relative value of sub-indicators 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 more 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, data envelopment analysis (DEA) – one of the most widespread methods to compare organisations with multiple outputs and inputs (Hollingsworth and Street, 2006) – uses empirically derived, flexible weights, following a “benefit of the doubt“ approach. It is however questionable whether data-driven weights reflect meaningful trade-offs between performance domains (Decancq and Lugo, 2012). There is no logical reason why an organisation necessarily 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 the lack of clarity 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 typically 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 necessarily 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 (Romano et al., 2010). However, the population at risk of experiencing a specific event is not always obvious. Suppose a national government wants to assess performance on health-care associated infections (HAIs) among local health authorities which are responsible for protecting the health of their local populations. To measure health authority performance on HAIs, two measures of the PAR have been proposed: hospital occupied bed days (OBDs) and total population living in the health authority area (Health Protection Scotland, 2007).

Using OBDs as the denominator implies that each day spent in the hospital puts patients at risk of acquiring an infection there. However, OBDs ignore that some infections are not acquired in hospital but in the community (Health Protection Scotland, 2014). Using OBDs as the denominator might thus underestimate the actual number of exposed individuals. Total population as a measure of the PAR, in contrast, implies the view that every person could acquire an infection, independent of hospital activity (Health Protection Scotland, 2007). Nevertheless, total population might overestimate the population at risk by including individuals facing no or a negligible risk of experiencing the event (Marlow, 1995).

Ideally, one would specify a numerator that is unambiguously linked to one single denominator (McKibben et al., 2005); for example, by excluding community-acquired infections that are present on admission to hospital from the numerator. In practice, it is however often difficult to distinguish between infections that were present on admission and those acquired during a hospital stay (Naessens and Huschka, 2004, Zhan et al., 2007).

If the “correct” population at risk is not obvious, then Guillen et al. (2011) recommend to consider different denominators to acquire a more complete perspective on the outcome of interest. To do this, one could produce multiple ratios between all reasonable numerator and denominator combinations. However, the manual comparison of multiple performance ratios quickly becomes unwieldy. In a situation with, say, four numerators and three denominators, one would obtain 12 performance ratios for each entity under scrutiny.

3 METHODS

3.1 Ranking intervals and dominance relations for all feasible weights

We here examine the use of an alternative approach to handle the lack of knowledge about appropriate weights and about a defensible denominator. Rather than specifying explicit weights, this approach consists in developing ranking intervals and dominance relations based on the full set of feasible weights. The approach is also able to handle different choices of denominator variables.

We use the ratio-based efficiency analysis (REA) technique (Salo and Punkka, 2011). Suppose there are K Decision-Making Units (DMUs – the entities to be evaluated) that have N different measures for the numerator of a ratio and M measures for the denominator of a ratio. The values of the nth numerator and the mth denominator of the kth DMU are $y_{nk} \geq 0$ and $x_{mk} \geq 0$, respectively. Thus, the possible performance ratios of the DMU k are y_{nk}/x_{mk}, where $n = 1, \ldots, N$ and $m = 1, \ldots, M$.

REA enables the aggregation of different numerators and denominators in a summary measure of performance. The relative importance of the nth numerator and the mth denominator is captured by nonnegative weights u_n and v_m, respectively. The aggregated performance ratio of DMU k is defined as

$$E_k(u, v) = \frac{\sum_n u_n y_{nk}}{\sum_m v_m x_{mk}}. \tag{2}$$

To examine the pairwise relations between DMUs, REA uses the concept of dominance: DMU k dominates DMU l if the performance ratio of DMU k is at least as high as that of DMU l for all feasible weights and there exist some weights for which its performance ratio is strictly higher. If a dominance relation exists between two DMUs, one can be confident that for any set of assumption, one DMU outperforms the other. The dominance relation between DMUs k and l is determined by the pairwise performance ratio

$$D_{k,l}(u,v) = \frac{E_k(u,v)}{E_l(u,v)}. \quad (3)$$

The maximum and the minimum of $D_{k,l}(u,v)$ over all feasible weights provide upper and lower interval bounds on how well DMU k performs relative to DMU l. Thus, if the minimum of $D_{k,l}$ is greater than one, DMU k dominates DMU l.

The ranking interval indicates the best and worst performance rankings a DMU k can attain relative to other DMUs over all feasible weights. The best ranking is determined by the minimum number of other DMUs with a strictly higher performance ratio. For instance, the best ranking as third for a given DMU means that, no matter how the weights are selected, there are at least two other DMUs with a strictly higher performance ratio. If for some feasible weights the performance ratio of a DMU is higher than or equal to the ratio of any other DMU, then its best ranking will be one. The worst ranking is computed similarly.

The results of REA (ratio and ranking intervals and dominance graphs) are computed using general programming methods such as linear programming and mixed integer programming.

3.2 Method strengths and limitations

There are several innovative characteristics, and advantages, to this approach. First, the aggregation of numerators and the denominators is achieved without fixing a single set of weights for each DMU. While weights are derived analytically (as in DEA), the key innovation of REA is that one compares the relative magnitude of the performance ratios between DMUs for all feasible weights (rather than applying the most favourable weighting of variables to each organisation as in DEA (Cherchye et al., 2007)). Although one can obtain ranking intervals with DEA (by applying different sets of weight restrictions), these intervals still represent the highest possible performance for each set of weight restrictions. In REA, the upper limit of the performance ratio interval is identical to the performance score of DEA. In addition, however, the lower limit of intervals in REA shows organizational performance for the least advantageous weighting. Thus, one can produce robust information about organizational performance in the sense that the resulting intervals reflect the full range of rankings that DMUs may attain for all feasible weights.

Second, REA calculates pairwise comparisons between DMUs rather than comparing each DMU to an efficient frontier as in DEA or stochastic frontier analysis. This makes REA results more robust than frontier-based results, since the introduction or removal of an outlier DMU can substantially change the location of the efficiency frontier (Banker et al., 1986). In contrast, already established pairwise dominance relations obtained from REA cannot change if a new DMU is added; and the end points of any DMU's ranking interval can shift towards lower performance by at most one ranking.

Third, because the REA is based on pairwise comparisons, it requires a minimum of only two DMUs. In contrast, frontier-based methods typically require a larger number of DMUs to construct the frontier. For DEA, for instance, Banker et al. (1986) proposed the simple rule of thumb that the number of DMUs should be at least three times the number of variables. This is problematic because the number of indicators typically far outstrips the number of organisations.

It is important to point out that, where the choice of denominator is relatively straightforward, ratio-based analysis is not necessary. One can calculate individual performance rates for the respective indicators and aggregate them as a weighted sum as in equation (1). This is akin to evaluating the numerator of the performance ratio (2).

We here use a ratio-based analysis in order to illustrate robustness to different choices of denominator. Ratio-based measures have certain limitations. In particular, the use of a ratio function does not account for structural differences (such as a higher share of fixed costs) between organisations. This assumption implies that, in evaluating organisational performance, one does for instance not "allow" an organisation a comparatively higher number of healthcare-associated infections (in ratio terms, e.g. per 100,000 population) only because it is relatively small in size. However, in the context we examine here – Scottish Health Boards, as outlined below – this assumption seems justified since these Boards are allocated resources in line with a formula which seeks to compensate for structural differences so as to ensure a level playing field across organisations.

Ratio measures may be preferred when there is primarily a concern with evaluation (examining which organisations perform comparatively better or worse) rather than explanation (examining why organisations achieve particular performance outcomes, as in regression analysis). This paper is limited to the problem of comparative evaluation.

3.3 System context and data

Selection of indicators. We illustrate the robust ranking interval approach in the context of comparative quality assessments of Scottish Health Boards. In Scotland, responsibility for the allocation of resources is decentralized to 14 territorial Boards. The ultimate objectives of these Boards are to protect and improve the health of their populations through planning for and delivering health services (Scottish Government, 2014). To construct a composite indicator of the quality of care provided by Boards, we confined ourselves to indicators used in the HEAT target system. This existing performance management system is used by the Scottish Government to assess Health Board performance. All indicators used here (Table 1) come from the official performance measurement system, but are not meant to represent an exhaustive set of health system objectives. To address the two problems examined in this study, we use two data sets:

Part I: To examine robustness to choices of weights, we analyse six indicators from the HEAT target system which are intended to measure Boards' relative degree of achievement in ensuring appropriate treatment. This analysis is based on an additive model which is akin to analyzing the numerator of the performance ratio in equation (2).

Part II: To examine robustness to alternative choices of denominator (here, the population at risk of experiencing an infection), we relate the number of two types of HAIs (MRSA/MSSA and C.difficile infections) to OBDs and total population. This analysis relies on the more complex ratio-based model in equation (2). We focus on HAIs because there is a good justification for two alternative denominators, bed days and total population (as set out in section 2.2). The REA-based analysis with two numerators and two denominators thus shows the full strength of the ratio-based approach. However, our focus on HAIs does not mean that for the other four quality indicators, no alternative denominators might be possible.

Data transformation. To avoid mixing different units of measurement and to achieve scale invariance, data were normalized to the [0;1] range by dividing each value by the maximum value for a given indicator.

Environmental adjustment. The 14 Health Boards differ in terms of demographic, epidemiological and regional factors which are beyond their control but might influence observed performance. However, in Scotland, Health Boards are allocated resources based on a formula that takes account of variations in healthcare needs which arise from differences in age and sex composition, morbidity, life circumstances, and excess costs of delivering services in some (especially rural) regions which are deemed unavoidable (ISD Scotland, 2010). Thus, Boards have already been compensated for structural differences so that they can ensure the same level of quality. We acknowledge that the risk adjustment provided by this formula is not perfect. However, following this argument, it is not unreasonable to assume that Boards are comparable with respect to the performance indicators analysed here.

Tables 1 and 2 about here

3.4 Weight restrictions on quality measures

An advantage of REA is its ability to address incomplete information about weight specifications by using the full set of feasible weights. This can be an attractive option when one assumes complete ignorance about the relative value of averting particular events. However, while an elicitation of cardinal preferences over "how much" worse a, say, MRSA infection is compared to, say, an emergency admission may not feasible (e.g. due to high cognitive demands) or desirable (e.g. due to biases introduced by specific elicitation methods), it may be possible to obtain statements about which events are worse than others.

Introducing plausible weight restrictions based on ordinal preferences can be useful because this recognises people's ability to provide limited preference information about the relative badness of particular events without imposing implausibly exact weights. Restrictions on weights can be used to prevent inconsistencies with accepted views on the relative importance of measures analysed (Allen et al., 1997, Pedraja-Chaparro et al., 1997).

Based on their own subjective assessment, the research team arrived at a set of ordinal weights through pairwise comparisons of any two quality measures, along the lines "*If you* *could avoid either an emergency admission to hospital or an MRSA infection, which event would you rather avoid"*. Corresponding to their relative badness, events were ranked as follows (from worst=1 to least bad=6):

1. an MRSA/MSSA infection;
2. an emergency admission;
3. a C.difficile infection;
4. having to wait longer than 18 weeks from referral to treatment;
5. having to wait more than 4 hours in A&E (we assumed a condition where patients are in mild to moderate discomfort);
6. a delayed discharge.

In flexible weighting systems, the composite score may be heavily influenced by a sub-indicator that is marginally important in the wider health system context (Goddard and Jacobs, 2009). To address this problem, for Part I we made the (illustrative but reasonable) assumption that avoiding a particular event can at most have half of the overall value attached to avoiding an event of each of the six quality measures. This resulted in the following proportional weight restrictions: avoiding an event of the worst healthcare quality measure cannot be more than ten times as valuable as avoiding an event of the least bad quality measure (since with six indicators, a ratio of 1/10 means that one quality measure can have at most half of the weight mass).

For part II, we made the (illustrative but reasonable) assumption that avoiding one C.difficile infection must be at least 1/4 as valuable as avoiding one MRSA/MSSA infection.

No weight restrictions for denominator variables were used. In efficiency analysis, denominator weights have a clear interpretation, because they indicate the substitutability of different types of inputs (labor, capital, intermediate inputs). In quality comparisons, denominators represent different populations at risk. However, denominator weights lack a clear interpretation as in efficiency analysis since it is hard to think about trade-offs between different populations at risk.

4 RESULTS

4.1 Robustness to choices of weights: Unrestricted and restricted ranking intervals for feasible weight sets

The ranking intervals (Figures 1-3) show the possible rankings that Boards can attain for different assumptions about weight sets. If one uses all feasible weights (Figure 1), then one obtains wide and overlapping ranking intervals spanning 9 to 14 ranks for a given Board. With ordinal weight restrictions, the width of ranking intervals decreases to 3 to 11 ranks (Figure 2). Thus, uncertainty about relative performance decreases as weight restrictions are applied.

However, the impact of weight restrictions on reductions in uncertainty differs across Boards. For Boards *L* and *H,* ordinal weight restrictions narrow the ranking interval from 11 respectively 12 ranks (Figure 1) to 3 possible ranks (Figure 2), thus clarifying Board performance. In contrast, for Boards *N, E, M* and *A,* ranking intervals remain wider, because these Boards perform comparatively well on some indicators, but comparatively worse on others (Table 2). Hence, the remaining flexibility to set weights influences the ranks these Boards may attain. For 7 out of 14 Boards (*K, F, B, E, C, A, J*), the additional use of proportional weight restrictions (Figure 3) further decreases uncertainty about relative ranks.

The width of the ranking interval reflects the impact of changes in weights. A narrow interval suggests that a Board's performance is robust to alternative modelling assumptions. For example, Board *L* (Figure 2) is ranked 3rd or higher no matter which assumptions are used. The interval bounds show the impact of modelling assumptions on relative ranks. Thus, one can be confident that Board *F,* for example, cannot be ranked worse than 7th and not better than 3nd.

Figures 1 to 3 about here

4.2 Dominance relations and comparative scope for improvement

Based on pairwise comparisons, the REA results can be displayed in a unified way as a dominance relation (Figure 4): insofar as Boards are more superordinate or "higher up", their relative performance is more robust to changes in the weights attached to the constituent indicators. Orkney (K), Shetland (L) and Western Isles (N) are top performers since they are not dominated by any other Board. Ayrshire and Arran (A), Fife (D), Greater Glasgow and Clyde (G), Lothian (J) and Tayside (M) are dominated by the other Boards.

There are two main reasons for this differentiation status. First, a Board's performance on the constituent indicators plays a role (Table 2). For instance, all three island Boards perform comparatively better than the rest of Scotland on MRSA/MSSA infections, 4-hour A&E waiting times and 18WRTT. Second, the ordinal weight restrictions used influence the dominance relations. In this example, performance on MRSA/MSSA infections is weighted more highly than performance on emergency admissions, which in turn receives a higher weight than performance on C.difficile, etc. Inspection of the underlying data (Table 2) suggests that the five Boards at the bottom of the dominance graph perform comparatively worse on MRSA/MSSA infections and emergency admissions. Nevertheless, their overall performance results from poor performance on several (up to four) indicators and thus not exclusively from the weighting scheme.

In Table 3, the value in row i and column j represents the minimal proportional improvement which Board i needs to reach Board j (by decreasing its rates, since these are "lower is better" indicators). Thus, if a value on row i and column j is presented, Board j performs better than Board i with all feasible weights and thus dominates Board i. For instance, Board A needs to reduce its rates on all the indicators by 8% so as not to be dominated by Board B. Non-dominated Boards are identified by rows without any values (Boards K, L, and N).

Multiple values on the same row mean that a Board is dominated by several Boards and would be situated on lower levels of the dominance graph. Looking horizontally, one can see the improvements needed for the five worst performing Boards J, G, D, M, A to become non-dominated by the better-performing Boards. Looking vertically, one can identify the distance that differentiates each Board from the national leaders, Boards K, L and N.

Figure 4 about here

Table 3 about here

4.3 Ratio-based analysis: Robustness to choice of denominator

Table 4 examines robustness to different choices of denominator. Although relative performance of seven Boards is similar for both denominators, the other seven Boards jump three to eight ranks up or down the ranking depending on whether total population or OBDs is used as the denominator (for C.difficile infections). For MSSA/MSSA, three Boards jump four or five ranks for different choices of denominator. Thus, the choice of denominator will make a difference to measured performance of these Boards on HAIs.

The REA-based ranking interval, which shows composite performance on MRSA/MSSA and C.difficile relative to OBDs and population, reveals seven Boards (marked in bold in Table 4) with a ranking interval spanning seven or more ranks. This uncertainty in ranking reflects, first, sensitivity to choice of denominator (e.g. Borders jumps up four ranks when MRSA/MSSA and C.difficile are measured relative to total population). Second, this may show differences in performance on MRSA/MSSA as opposed to C.difficile (e.g. Forth Valley is ranked 13th on the former but 2nd on the latter relative to OBDs).

Table 4 about here

5 DISCUSSION

We have proposed a methodological approach to address two pervasive challenges which make the use of composite measures for robust performance comparisons in healthcare difficult: How should heterogeneous indicators be weighted to obtain an aggregate measure of performance? How to handle a lack of clarity about the “correct“ denominator in ratio-based indicators? As Jacobs et al. (2005) note, two responses to the uncertainty inherent in composite indicators would be to dismiss composite indicators altogether and instead estimate relative performance separately for each objective (an example of this is Hauck and Street’s (2006) multivariate multilevel approach that requires no aggregation and weighting of multiple objectives at all); or to invest considerable resources into more sophisticated modelling, such as by means of elaborate preference elicitation.

In a context where information is inevitably incomplete but policy-makers remain interested in an overall measure of health system performance (OECD, 2008), we have demonstrated how the REA approach offers a third way that openly provides indications of the uncertainty inherent in the valuation of objectives and choices of denominators. The approach is essentially based on agnosticism: When there are multiple reasonable denominators which each highlight aspects of performance – such as that an organisation can deliver high-quality in terms of few HAIs relative to hospitalised and/or general populations – then analysts need not restrict themselves to a single denominator. Our results reinforce the insight that healthcare quality may be best thought of as a collection of possible rates depending on how the denominator is specified rather than as a single "right" rate (Guillen et al., 2011). Ranking intervals based on multiple denominators thus may enable a more complete account of performance.

Similarly, if we know that quality measures are heterogeneous but are ignorant of the best method to weight them, then methods to construct composite indicators need to capture that lack of knowledge. Sensitivity analysis on weights is not a new idea; prior work – especially in the multidimensional well-being literature – includes explicit use of ranges of weights (Zhou et al., 2010); computation of multiple weighting schemes (Osberg and Sharpe, 2002); and global sensitivity analysis (Saltelli et al., 2008).

The REA approach adds to this work in two ways. First, consideration of incomplete information is built into the structure of the model. Ranking intervals give policy-makers a sense of the uncertainty around ranks, indicating the extent to which action is warranted. Our results show that, when one assumes complete ignorance about the relative weights assigned to different indicators, then it is impossible to differentiate the performance of Scottish Health Boards (Figure 1). Thus, one cannot say which organisations perform comparatively better or worse. Regulatory action based on such rankings would clearly be premature.

However, once some reasonable ordinal and proportional weight restrictions are applied, organizational performance appears much clarified. Importantly, the use of REA without any weight restrictions involves no subjectivity (in the sense that weights are derived for all feasible combinations for each pairwise comparison). In contrast, the choice of weight restrictions may differ between groups of people: different individuals may come up with different orderings or proportionate weights concerning the relative badness (or goodness) of particular events. However, if weight restrictions can be established (e.g. based on existing consensus or medical evidence of disease severity), then they may provide useful insights. When an organisation consistently appears at the bottom (Board G) or at the top (Board L; in Figure 2) whichever set of weights is used, this may strengthen the rationale for policy intervention. It supports the notion that settling on a unique set of weights is not always necessary to inform well-founded judgments (Foster and Sen, 1997).

Second, ranking intervals and dominance relations appear to offer relatively intuitive ways to synthesise key messages contained in disparate indicators. This may help to communicate in a unified way the results of comparative assessments to policy-makers, possibly addressing the limitations of frontier-based approaches such as DEA and stochastic frontier analysis whose complexity has tended to limit their practical influence outside academic circles (Hussey et al., 2009, Hollingsworth and Street, 2006). Visualisation of uncertainty also mitigates the loss of transparency due to opaque methodological choices made about the valuation of objectives (Hauck and Street, 2006).

REA-type analyses are likely to be particularly useful under conditions where:

(i) the audience are policy-makers and managers rather than academics (since results such as being “30% below the efficient frontier“ may not be easily accessible to non-technical audiences and REA requires no concept of an efficient frontier);

(ii) there are concerns about rank reversals due to sensitivity to outliers and the introduction or removal of organisations (since pairwise comparisons make REA results relatively robust to these biases); and

(iii) there are relatively few organisations (since a large number of organisations is not needed to construct an efficient frontier). However, there are also no inherent limitations to applying REA to large datasets.

6 IMPLICATIONS FOR POLICY AND RESEARCH

The agnosticism implied in the REA approach may come at a price of incomplete orderings (in the form of wide and overlapping ranking intervals). Ranking intervals will become wider and more overlapping the more performance indicators are used (compared to the number of organisations) and, at the same time, the weaker the correlation between these indicators (i.e. the less information good or poor performance on one indicator provides about relative performance on other indicators). The number of indicators and the appropriate degree of correlation will depend on the purpose of the analysis. Wide and overlapping ranking intervals do not indicate that REA is not applicable. For policy-makers and managers, a key strength of REA is that wide and overlapping intervals visualize in a transparent way the existing uncertainty.

Evidence of uncertainty reinforces the need to use the results as signals for further analysis, rather than for definitive judgments. Since weakly correlated indicators will make rankings more sensitive to different sets of weights (Foster et al., 2012), the careful use of weight restrictions becomes particularly important. Weight restrictions will tend to clarify the results and make explicit the impact of subjective choices about the relative value of different quality indicators on performance rankings.

Dominance relations that are based on pairwise comparisons between Boards provide comparative performance assessments one can be confident about. Since dominance relations indicate that some DMU k performs at least as well as some other DMU l for all feasible weights and there exist some weights for which it performs strictly better, this information could, for instance, be used for setting performance targets across all indicators included in the analysis. Since improvements on some indicators may require less effort than others, indicator-specific improvements would also be informative. However, this would require a different approach. Gouveia et al. (2015), for instance, employ slack-variables (which define the variable-specific distance to the efficient frontier) to estimate the improvements required for a DMU to reach the best performing organisation. However, this approach does not indicate the improvements needed to reach some specific, non-efficient DMU as it is possible with our approach. This is particularly relevant for policy and management and a strength of our study, since the top performing organisation may not always be the most meaningful (and practically feasible) benchmark for worse performing organisations. In a collegiate rather than competitive environment, such results could help organisations to learn from better performing (dominating) peers.

For a large number of organisations (and dominance relations), the clear presentation and communication of results to decision-makers becomes even more important. To simplify the dominance graph, DMUs which perform similarly can be grouped together (as with DMUs D and M in Figure 4). A large number of dominance relations can also be visualized using a matrix (see Table 3) which shows both the dominance relations and the magnitude of dominance.

Finally, it is essential to re-emphasize the importance of the other methodological choices (listed in section 2) that must be made when constructing a composite indicator; in particular, the initial selection of indicators and risk adjustment for environmental (uncontrollable) determinants of performance. If important indicators are omitted or irrelevant variables are included, then performance evaluations will be meaningless (Smith, 1997). The choice of performance metrics therefore needs to reflect a country's definition of valued outcomes of the health service (Dowd et al., 2014).

Concerning risk adjustment, in Scotland the funding formula is designed to enable all NHS Boards to produce equal levels of performance. Since this formula takes account of differences in population and local characteristics (e.g. rurality), in this study we have followed the argument that risk adjustment has been implemented via the funding system (Jacobs et al., 2006). However, the degree to which this argument holds depends on the accuracy and comprehensiveness of the formula. While for our study the direction of any potential bias is difficult to determine, it is possible that inadequate risk adjustment has affected observed Board performance on the constituent indicators.

As Smith (2003) notes, formula funding is fraught with challenges, such as that performance criteria have proved hard to include in the formula. This means that poor quality of care which increases levels of morbidity might be 'rewarded' with higher levels of funding. As a result, the link between resource allocation and performance measurement remains complex and an important avenue for future research.

REFERENCES

ALLEN, R., ATHANASSOPOULOS, A., DYSON, R. G. & THANASSOULIS, E. 1997. Weights restrictions and value judgements in Data Envelopment Analysis: Evolution, development and future directions. *Annals of Operations Research,* 73**,** 13-34.

APPLEBY, J. & MULLIGAN, J. 2000. *How well is the NHS performing? A composite performance indicator based on public consultation,* London, The King's Fund.

BANKER, R. D., CONRAD, R. F. & STRAUSS, R. P. 1986. A Comparative Application of Data Envelopment Analysis and Translog Methods - an Illustrative Study of Hospital Production. *Management Science,* 32**,** 30-44.

BEVAN, G. & HAMBLIN, R. 2009. Hitting and missing targets by ambulance services for emergency calls: Effects of different systems of performance measurement within the UK. *Journal of the Royal Statistical Society. Series A,* 172**,** 161-190.

CARINCI, F., VAN GOOL, K., MAINZ, J., VEILLARD, J., PICHORA, E. C., JANUEL, J. M., ARISPE, I., KIM, S. M. & KLAZINGA, N. S. O. B. O. T. O. H. C. Q. I. E. G. 2015. Towards actionable international comparisons of health system performance: expert revision of the OECD framework and quality indicators. *International Journal for Quality in Health Care,* 27**,** 137-146.

CHERCHYE, L., MOESEN, W., ROGGE, N. & VAN PUYENBROECK, T. 2007. An introduction to 'benefit of the doubt' composite indicators. *Social Indicators Research,* 82**,** 111-145.

CMS 2009. Centers for Medicare & Medicaid Services. Premier Hospital Quality Incentive Demonstration: Fact sheet. http://www.cms.hhs.gov/HospitalQualityInits/downloads/HospitalPremierFactSheet200907.pdf [9 May 2014].

DECANCQ, K. & LUGO, M. A. 2012. Weights in multidimensional indices of wellbeing: An overview. *Econometric Reviews,* 32**,** 7-34.

DOLAN, P. 1997. Valuing health states: A comparison of methods. *Journal of Health Economics,* 16**,** 617-617.

DOWD, B., SWENSON, T., KANE, R., PARASHURAM, S. & COULAM, R. 2014. CAN DATA ENVELOPMENT ANALYSIS PROVIDE A SCALAR INDEX OF 'VALUE'? *Health Economics,* 23**,** 1465-1480.

FOSTER, J. & SEN, A. 1997. *On Economic Inequality,* Oxford, Oxford University Press.

FOSTER, J. E., MCGILLIVRAY, M. & SETH, S. 2012. Composite Indices: Rank Robustness, Statistical Association, and Redundancy. *Econometric Reviews,* 32**,** 35-56.

GODDARD, M. & JACOBS, R. 2009. Using composite indicators to measure performance in health care. *In:* SMITH, P., MOSSIALOS, E., PAPANICOLAS, I. & LEATHERMAN, S. (eds.) *Performance measurement for health system improvement: experiences, challenges and prospects.* Cambridge: Cambridge University Press, pp. 339-368.

GOUVEIA, M., DIAS, L., ANTUNES, C., MOTA, M., DUARATE, E. & TENREIRO, E. 2015. An application of value-based DEA to identify the best practices in primary health care. *OR Spectrum. DOI 10.1007/s00291-015-0407-x*.

GUILLEN, Ú., DEMAURO, S., MA, L., ZUPANCIC, J., WANG, E., GAFNI, A. & KIRPALANI, H. 2011. Survival rates in extremely low birthweight infants depend on the denominator: avoiding potential for bias by specifying denominators. *American Journal of Obstetrics and Gynecology,* 205**,** 329.e1-329.e7.

HAUCK, K. & STREET, A. 2006. Performance assessment in the context of multiple objectives: a multivariate multilevel analysis. *Journal of Health Economics,* 25**,** 1029-48.

HEALTHCARE COMMISSION 2005. *2005 performance ratings,* London, Healthcare Commission.

HEALTH PROTECTION SCOTLAND 2007. Annual report on the surveillance of Clostridium difficile associated disease (CDAD) in Scotland*,* Glasgow, Health Protection Scotland.

HEALTH PROTECTION SCOTLAND 2014. *Healthcare Associated Infection Annual Report 2013,* Glasgow, Health Protection Scotland.

HOLLINGSWORTH, B. & STREET, A. 2006. The market for efficiency analysis of health care organisations. *Health Economics,* 15**,** 1055-1059.

HUME, D. (1739). *A Treatise of Human Nature.* London: John Noon.

HUSSEY, P. S., DE VRIES, H., ROMLEY, J., WANG, M. C., CHEN, S. S., SHEKELLE, P. G. & MCGLYNN, E. A. 2009. A systematic review of health care efficiency measures. *Health Services Research,* 44**,** 784-805.

ISD SCOTLAND 2010. Resource allocation. http://www.isdscotland.org/Health-Topics/Finance/Resource-Allocation-Formula/information.asp [11 November 2014].

JACOBS, R., GODDARD, M. & SMITH, P. 2005. How robust are hospital ranks based on composite performance measures? *Medical Care,* 43**,** 1177-84.

JACOBS, R., SMITH, P. & STREET, A. 2006. *Measuring efficiency in health care: analytic techniques and health policy*, Cambridge University Press.

MARLOW, A. 1995. Potential years of life lost: what is the denominator? *Journal of Epidemiology & Community Health,* 49**,** 320-2.

MCKIBBEN, L., HORAN, T., TOKARS, J. I., FOWLER, G., CARDO, D. M., PEARSON, M. L., BRENNAN, P. J. & THE HEALTHCARE INFECTION CONTROL PRACTICES ADVISORY, C. 2005. Guidance on Public Reporting of Healthcare-Associated Infections: Recommendations of the Healthcare Infection Control Practices Advisory Committee. *American Journal of Infection Control,* 33**,** 217-226.

NAESSENS, J. M. & HUSCHKA, T. R. 2004. Distinguishing hospital complications of care from pre-existing conditions. *International Journal for Quality in Health Care,* 16**,** I27-I35.

NAVARRO, V. 2000. Assessment of the world health report 2000. *Lancet,* 356**,** 1598-1601.

NHS ENGLAND 2015. *Revised Never Events Policy and Framework,* London, NHS England.

OECD 2008. *Handbook on Constructing Composite Indicators,* Paris, OECD.

OSBERG, L. & SHARPE, A. 2002. An index of economic well–being for selected OECD countries. *Review of Income and Wealth,* 48**,** 291-316.

PEDRAJA-CHAPARRO, F., SALINAS-JIMENEZ, J. & SMITH, P. 1997. On the Role of Weight Restrictions in Data Envelopment Analysis. *Journal of Productivity Analysis,* 8**,** 215-230.

REEVES, D., CAMPBELL, S. M., ADAMS, J., SHEKELLE, P. G., KONTOPANTELIS, E. & ROLAND, M. O. 2007. Combining multiple indicators of clinical quality: an evaluation of different analytic approaches. *Medical Care,* 45**,** 489-96.

ROMANO, P., HUSSEY, P. & RITLEY, D. 2010. *Selecting Quality and Resource Use Measures: A Decision Guide for Community Quality Collaboratives,* Rockville, Agency for Healthcare Research and Quality.

SALO, A. & PUNKKA, A. 2011. Ranking intervals and dominance relations for ratio-based efficiency analysis. *Management Science,* 57**,** 200-214.

SALTELLI, A., RATTO, M., ANDRES, T., CAMPOLONGO, F., CARIBONI, J., GATELLI, D., SAISANA, M. & TARANTOLA 2008. *Global Sensitivity Analysis: The Primer*, Wiley E-book.

SCHLAUD, M., BRENNER, M. H., HOOPMANN, M. & SCHWARTZ, F. W. 1998. Approaches to the denominator in practice-based epidemiology: a critical overview. *Journal of Epidemiology & Community Health,* 52**,** 13S-19S.

SCOTTISH GOVERNMENT 2014. NHS Boards. http://www.scotland.gov.uk/Topics/Health/NHS-Workforce/NHS-Boards [14 May 2014].

SMITH, P. 1997. Model misspecification in data envelopment analysis. *Annals of Operations Research,* 73**,** 233-252.

SMITH, P. 2002. Developing composite indicators for assessing health system efficiency. *In:* SMITH, P. (ed.) *Measuring up: improving health system performance in OECD countries.* Paris: OECD, pp. 295-318.

SMITH, P. C. 2003. Formula funding of public services: An economic analysis. *Oxford Review of Economic Policy,* 19**,** 301-322.

WHO 2000. *The world health report 2000 - Health systems: improving performance,* Geneva, World Health Organization.

ZHAN, C. L., ELIXHAUSER, A., FRIEDMAN, B., HOUCHENS, R. & CHIANG, Y. P. 2007. Modifying DRG-PPS to include only diagnoses present on admission - Financial implications and challenges. *Medical Care,* 45**,** 288-291.

ZHOU, P., ANG, B. W. & ZHOU, D. Q. 2010. Weighting and aggregation in composite indicator construction: A multiplicative optimization approach. *Social Indicators Research,* 96**,** 169-181.

TABLES AND FIGURES

Table 1 Variables and descriptive statistics

	Definition	Mean	SD	Min	Max
Data for part I: robustness to choices of weights and dominance relations					
18WRTT[a]	Number of patient journeys from referral to treatment over 18 weeks (among patients seen) per 100,000 RTT patient journeys from referral to treatment (among patients seen)	7,361	3,475	2,209	15,123
4-hour A&E waiting[a]	Number of recorded A&E waits lasting over 4 hours per 100,000 A&E attendances	4,739	3,090	730	9,172
Emergency admissions[a]	Number of emergency admissions among +75 years per 100,000 population	2,887	424	2,239	3,646
MRSA/MSSA[a]	Number of MRSA/MSSA infections per 100,000 population	23	10	4	36
C.difficile[a]	Number of Clostridium difficile infections per 100,000 population	44	28	14	123
Delayed discharges[a]	Number of bed days lost due to delayed discharges per 100,000 occupied bed days	29	18	6	69
Data for part II: robustness to choices of denominator					
Quality indicators (numerator variables)					
C.difficile[a]	Number of Clostridium difficile infections	133	123	8	399
MRSA/MSSA[a]	Number of MRSA/MSSA infections	108	114	1	413
Population indicators (denominator variables)					
Total population[b]	Resident population (mid-year estimates)	475,232	318,214	113,880	1,214,587
OBD[a]	Number of occupied bed days	113,244	98,182	20,723	365,951

Sources: [a]HEAT target system; [b]National Records of Scotland. All data are for 2012/13.

Table 2 Comparative performance of Boards on the constituent six quality indicators, based on rates as shown in Table 1, part I

		18WRTT	4-hour A&E waiting	Emergency admissions	MRSA/MSSA	C.difficile	Delayed discharges
A	**Ayrshire & Arran**	8,691	8,312	3,646	23	49	14
B	**Borders**	6,204	3,267	3,612	21	44	10
C	**Dumfries & Galloway**	6,170	5,987	3,130	27	36	29
D	**Fife**	6,899	4,559	2,725	35	26	69
E	**Forth Valley**	15,123	8,238	2,513	26	14	50
F	**Grampian**	9,343	3,812	2,239	25	24	43
G	**Greater Glasgow & Clyde**	8,523	6,956	3,061	34	33	17
H	**Highland**	5,817	2,199	2,825	17	24	45
I	**Lanarkshire**	5,551	8,667	2,671	24	35	24
J	**Lothian**	12,293	9,172	2,495	30	42	43
K	**Orkney**	2,649	1,663	2,661	9	84	6
L	**Shetland**	2,209	730	2,555	13	34	14
M	**Tayside**	8,701	1,119	2,964	36	50	21
N	**Western Isles**	4,876	1,666	3,320	4	123	21

Table 3 Comparative scope for improvement needed to reach another target or reference Board in Scotland

Dominated Board		Target or Reference Board													
		A	B	C	D	E	F	G	H	I	J	K	L	M	N
Ayrshire & Arran	**A**		8 %				2 %		25 %	2 %		22 %	36 %		2 %
Borders	**B**								9 %			14 %	27 %		
Dumfries & Galloway	**C**		<1 %				7 %		21 %			15 %	31 %		
Fife	**D**		3 %				11 %		24 %			17 %	32 %		
Forth Valley	**E**						7 %		12 %			3 %	21 %		
Grampian	**F**								6 %				15 %		
Greater Glasgow & Clyde	**G**		9 %	8 %			16 %		29 %	11 %		22 %	36 %		2 %
Highland	**H**												10 %		
Lanarkshire	**I**								12 %			6 %	23 %		
Lothian	**J**		4 %	2 %		6 %	18 %		23 %	11 %		18 %	33 %		
Orkney	**K**														
Shetland	**L**														
Tayside	**M**		8 %				4 %		20 %			25 %	36 %		
Western Isles	**N**														

Table 4 Performance on healthcare-associated infections relative to different choices of denominator

Board	Per 100,000 OBDs		Per 100,000 population		Per 100,000 OBDs		Per 100,000 population		Ranking interval for composite performance on MRSA/MSSA and C.difficile relative to OBDs and population
	Number of MRSA/MSSA	Rank	Number of MRSA/MSSA	Rank difference compared to OBDs	Number of C.difficile	Rank	Number of C.difficile	Rank difference compared to OBDs	
Shetland	21	3	13	0	55	1	34	*-5*	1-3
Highland	87	4	17	0	124	6	24	+3	1-4
Forth Valley	**148**	**13**	**26**	***+4***	**78**	**2**	**14**	**+1**	**1-10**
Orkney	**13**	**2**	**9**	**0**	**114**	**5**	**84**	***-8***	**2-13**
Western Isles	**4**	**1**	**4**	**0**	**140**	**7**	**123**	***-7***	**2-14**
Grampian	108	6	25	-2	105	3	24	+1	4-6
Lanarkshire	113	8	24	+1	162	10	35	+3	5-8
Borders	**116**	**9**	**21**	***+4***	**241**	**14**	**44**	**+4**	**5-14**
Dumfries & Galloway	117	10	27	0	161	9	36	+1	6-10
Greater Glasgow & Clyde	**113**	**7**	**34**	***-5***	**109**	**4**	**33**	**-1**	**6-13**
Fife	**211**	**14**	**35**	**+1**	**155**	**8**	**26**	**+4**	**6-14**
Ayrshire & Arran	**99**	**5**	**23**	**-1**	**211**	**13**	**49**	**+2**	**7-13**
Lothian	127	11	30	0	177	11	42	+2	10-13
Tayside	141	12	36	-2	195	12	50	0	12-14

Figure 1 Performance rankings for all feasible weights

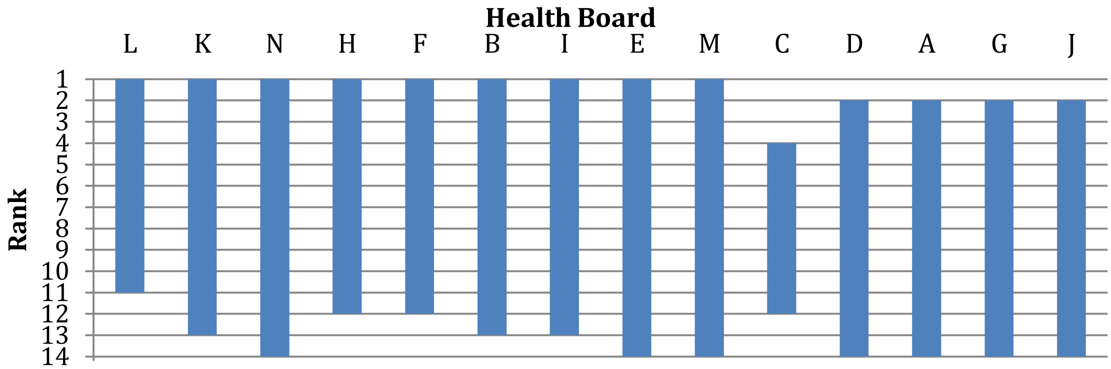

Figure 2 Performance rankings with ordinal weight restrictions

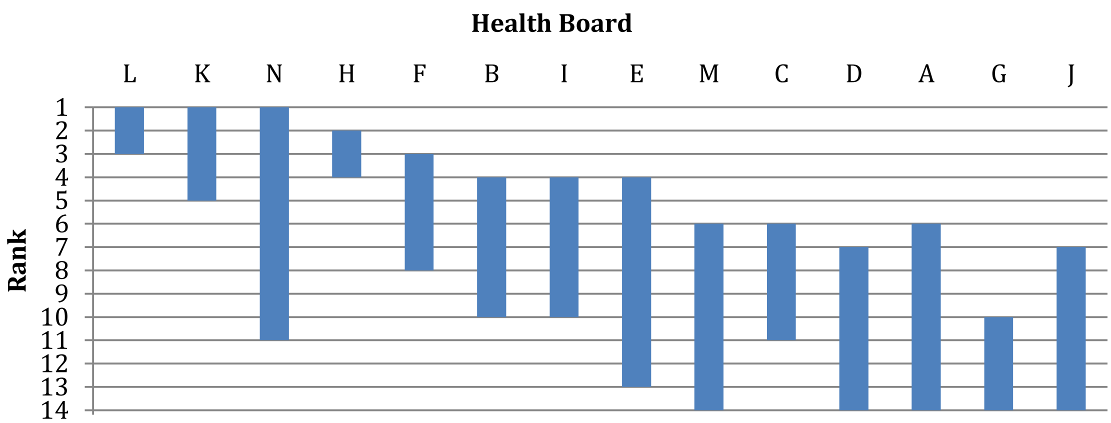

Figure 3 Performance rankings with ordinal and proportional weight restrictions

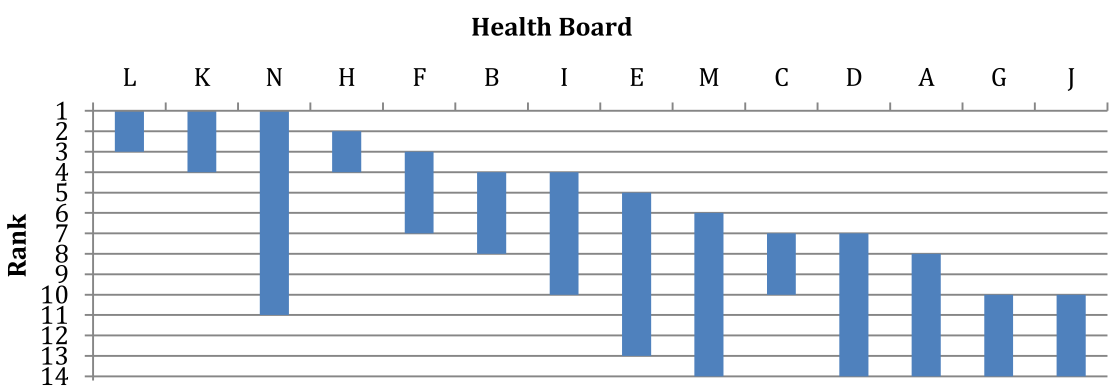

Figure 4 Dominance graph for Scottish Health Boards, based on ordinal and proportional weight restrictions