

[Laura Schang](#), Yrjänä Hynninen, Alec Morton, Ahti Salo
**Developing robust composite measures of
healthcare quality – ranking intervals and
dominance relations for Scottish Health
Boards**

**Article (Accepted version)
(Refereed)**

Original citation:

Schang, Laura, Hynninen, Yrjänä , Morton, Alec and Salo, Ahti (2016) *Developing robust composite measures of healthcare quality – ranking intervals and dominance relations for Scottish Health Boards*. [Social Science and Medicine](#), 162. pp. 59-67. ISSN 0277-9536

DOI: [10.1016/j.socscimed.2016.06.026](https://doi.org/10.1016/j.socscimed.2016.06.026)

Reuse of this item is permitted through licensing under the Creative Commons:

© 2016 [Elsevier Ltd.](#)
CC BY-NC-ND 4.0

This version available at: <http://eprints.lse.ac.uk/67145/>

Available in LSE Research Online: July 2016

LSE has developed LSE Research Online so that users may access research output of the School. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. You may freely distribute the URL (<http://eprints.lse.ac.uk>) of the LSE Research Online website.

Accepted Manuscript

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

Laura Schang, Yrjänä Hynninen, Alec Morton, Ahti Salo



PII: S0277-9536(16)30310-0

DOI: [10.1016/j.socscimed.2016.06.026](https://doi.org/10.1016/j.socscimed.2016.06.026)

Reference: SSM 10699

To appear in: *Social Science & Medicine*

Received Date: 24 August 2015

Revised Date: 24 May 2016

Accepted Date: 15 June 2016

Please cite this article as: Schang, L., Hynninen, Y., Morton, A., Salo, A., Developing robust composite measures of healthcare quality – Ranking intervals and dominance relations for Scottish Health Boards, *Social Science & Medicine* (2016), doi: 10.1016/j.socscimed.2016.06.026.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

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

Correspondence to:

Laura Schang, PhD
Department of Health Services Management
Ludwig-Maximilians-Universität München
Schackstraße 4
80539 DE-Munich
schang@bwl.lmu.de

1 ABSTRACT

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

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

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

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

23 **1 Introduction**

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

34
35 However, composite indicators also have important disadvantages. In contrast to assessing
36 performance based on a range of separate indicators, rankings based on aggregate
37 measures may disguise the sources of poor performance and thus obscure the best focus
38 for remedial action (Smith, 2002). Composite indicators are also highly sensitive to
39 methodological choices, in particular to the weights attached to constituent indicators (see
40 e.g. Jacobs et al., 2005, Reeves et al., 2007, OECD, 2008). In their analysis of hospital
41 performance based on star ratings in the English NHS, Jacobs et al. (2005) show, for
42 instance, how subtle changes in the weighting system lead some hospitals to jump almost
43 half of the league table. However, the techniques by which weights are determined are not
44 straightforward. In addition, although many comparative quality measures are constructed
45 as ratios, it is not necessarily obvious which indicators should be employed as

46 denominators (Schlaud et al., 1998). In the context of low-birthweight survival rates,
47 Guillen et al. (2011) illustrate how the choice of population denominator results in
48 considerable variation depending on whether survival is reported relative to all births; live
49 births; or neonatal intensive care unit admissions.

50
51 These concerns are critical especially when rankings have serious consequences for the
52 rankees. For example, six of the Chief Executives of the twelve lowest ranked hospitals in
53 England's star rating system (the so-called "dirty dozen") lost their jobs as a result (Bevan
54 and Hamblin, 2009). It has been argued that France and Spain's apparently high ranking in
55 the WHO's 2000 assessment of health systems substantially diminished pressure for
56 reform in these countries (Navarro, 2000). In Medicare's Premier Hospital Quality
57 Incentive Demonstration, a pay-for-performance scheme based on a composite quality
58 score, hospitals below the ninth decile faced a 2% deduction in their Medicare payment
59 (CMS, 2009). With such high stakes, understanding whether ranks are robust to alternative
60 assumptions seems critical.

61
62 This study proposes an alternative approach to handle the lack of information about an
63 appropriate set of weights and about a defensible denominator in composite indicators. We
64 make two main contributions. First, we demonstrate the use of an approach to ranking
65 organisations based on ranking intervals and dominance relations which accounts for the
66 full set of feasible weights. This avoids the need to settle on a single, potentially
67 controversial set of weights as it is required for instance in data envelopment analysis
68 (DEA), in which weights are chosen such that each organisation appears in its best possible

69 light (Cherchye et al., 2007). Feasible weights are less restrictive and thus potentially better
70 able to increase transparency and to acknowledge imperfect information about the
71 “correct” set of weights. The ranking intervals obtained with this approach can be said to
72 be robust in the sense that they reflect the full range of rankings that the organizations
73 under comparison may attain when weights are selected from their respective feasible
74 weight sets. Second, we address the problem of choice of denominator in ratio-based
75 measures of performance.

76

77 **2 Challenges in developing composite indicators of healthcare quality**

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

$$80 \quad C_k = \sum_{j=1}^J w_j x_{jk}, \quad (1)$$

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

88 The focus of this study is on problems (iii) and (vi), how to handle incomplete information
89 about weights and about the choice of denominator. Below we review the conceptual

90 background and problems with conventional strategies to address these challenges. In the
91 empirical application, we explain the approaches taken to problems (i), (ii), (iv) and (v).

92 **2.1 Valuation of multiple healthcare quality measures**

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

98
99 In practice, arriving at explicit trade-offs between different healthcare quality measures –
100 and thus exact specifications of weights – is highly contentious. First, it is often unclear
101 *whose* preferences should be elicited. Weights used often reflect a single set of preferences,
102 although the evidence suggests substantial heterogeneity in preferences between and
103 within groups of policy-makers, patients and the public (Smith, 2002, Decancq and Lugo,
104 2012). Making precise judgments about the relative value of sub-indicators to the
105 composite is typically both politically controversial and cognitively demanding, thus
106 triggering reluctance among respondents to agree on a set of weights.

107
108 Second, there is no consensus on a single best method *how* to elicit weights. Different
109 techniques for valuing health(care) outcomes – from simpler trade-off methods including
110 ranking from most to least desired indicator and voting techniques to elaborate multi-

111 attribute approaches such as conjoint analysis and the analytic hierarchy process – tend to
112 produce different results. Each method has distinct advantages and disadvantages in terms
113 of feasibility, consistency and validity (Dolan, 1997, OECD, 2008, Appleby and Mulligan,
114 2000).

115
116 To circumvent perceived difficulties with normative approaches to set weights, data-driven
117 weighting systems are frequently used. For example, in data envelopment analysis (DEA) –
118 a widespread method to compare organisations with multiple outputs and inputs
119 (Hollingsworth and Street, 2006) – weights are derived from the data so as to maximise
120 each organisation’s performance (Cherchye et al., 2007). Each organisation receives a
121 different set of weights which casts it in the best possible light. However, data-driven
122 weights do not necessarily reflect meaningful trade-offs between performance domains
123 (Decancq and Lugo, 2012). There is no logical reason why an organisation values most
124 some performance domain because it performs relatively well on it: data-driven
125 approaches thus purport to solve a deep philosophical problem of how to derive values
126 from facts (Hume, 1739).

127
128 The conventional recommendation to address incomplete information about weights, and
129 about the best method to elicit weights, is to conduct extensive sensitivity analysis on the
130 chosen weights (Jacobs et al., 2005). However, traditional sensitivity analysis is
131 problematic insofar as the choice of ranges of weights depends on the analyst. This form of
132 sensitivity analysis thus corresponds to a “blind search” which is not explicitly oriented

133 towards changes in ranks and the maximum and minimum plausible ranks an organisation
134 can attain.

135 **2.2 Choice of denominators**

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

143
144 To construct ratio-based quality measures, the denominator should represent the best
145 available proxy for the population at risk (PAR) (Romano et al., 2010). However, the PAR of
146 experiencing a specific event is not always obvious. Suppose a national government wants
147 to assess performance on healthcare-associated infections (HAIs) among local health
148 authorities which are responsible for protecting the health of their local populations. To
149 measure health authority performance on HAIs, two measures of the PAR have been
150 proposed: hospital occupied bed days (OBDs) and total population living in the health
151 authority area (Health Protection Scotland, 2007).

152

153 Using OBDs as the denominator implies that each day spent in the hospital puts patients at
154 risk of acquiring an infection there. However, OBDs ignore that some infections are not
155 acquired in hospital but in the community (Health Protection Scotland, 2014). Using OBDs
156 as the denominator might thus underestimate the actual number of exposed individuals.
157 Total population as a measure of the PAR, in contrast, implies the view that every person
158 could acquire an infection, independent of hospital activity (Health Protection Scotland,
159 2007). Nevertheless, total population might overestimate the PAR by including individuals
160 facing no or a negligible risk of experiencing the event (Marlow, 1995).

161
162 Ideally, one would specify a numerator that is unambiguously linked to one single
163 denominator (McKibben et al., 2005); for example, by excluding community-acquired
164 infections that are present on admission to hospital from the numerator. In practice, it is
165 however often difficult to distinguish between infections that were present on admission
166 and those acquired during a hospital stay (Naessens and Huschka, 2004, Zhan et al., 2007).

167
168 If the “correct” PAR is not obvious, then Guillen et al. (2011) recommend to consider
169 different denominators to acquire a more complete perspective on the outcome of interest.
170 To this end, one could produce multiple ratios between all reasonable numerator and
171 denominator combinations. However, manual comparisons of multiple performance ratios
172 quickly become unwieldy. In a situation with, say, four numerators and three
173 denominators, one would obtain 12 performance ratios for each entity under scrutiny.

174

175

176 3 Methods

177 3.1 Ranking intervals and dominance relations for all feasible 178 weights

179 We here examine the use of an alternative approach to handle incomplete information
180 about appropriate weights and a defensible denominator. This approach consists in
181 developing ranking intervals and dominance relations based on the full set of feasible
182 weights. It is also able to handle different choices of denominator variables.

183

184 We use the ratio-based efficiency analysis (REA) technique (Salo and Punkka, 2011).

185 Suppose there are K Decision-Making Units (DMUs – the entities to be evaluated) that have

186 N different measures for the numerator of a ratio and M measures for the denominator of a

187 ratio. The values of the n th numerator and the m th denominator of the k th DMU are

188 $y_{nk} \geq 0$ and $x_{mk} \geq 0$, respectively. Thus, the possible performance ratios of the DMU k are

189 y_{nk}/x_{mk} , where $n = 1, \dots, N$ and $m = 1, \dots, M$.

190

191 REA enables the aggregation of different numerators and denominators in a summary

192 measure of performance. The relative importance of the n th numerator and the m th

193 denominator is captured by nonnegative weights u_n and v_m , respectively. The aggregated

194 performance ratio of DMU k is defined as

$$195 \quad E_k(u, v) = \frac{\sum_n u_n y_{nk}}{\sum_m v_m x_{mk}}. \quad (2)$$

196

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

$$203 \quad D_{k,l}(u, v) = \frac{E_k(u,v)}{E_l(u,v)}. \quad (3)$$

204

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

208

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

216

217 REA-based results are computed using general programming methods such as linear
 218 programming and mixed integer programming (Bertsimas & Tsitsiklis, 1997). The idea

219 behind the use of these optimisation methods is to find, for each DMU, the highest
220 (respectively the lowest) ranking of that DMU for all feasible numerator and denominator
221 weights.

222

223 **3.2 Method strengths and limitations**

224 There are several innovative characteristics, and advantages, to this approach. First, the
225 aggregation of numerators and the denominators is achieved without fixing a single set of
226 weights for each DMU. The key innovation of REA is that one compares the relative
227 magnitude of the performance ratios between DMUs for all feasible weights (rather than
228 applying only the most favourable weighting of variables to each organisation as in DEA
229 (Cherchye et al., 2007)). Although one can obtain ranking intervals with DEA (by applying
230 different sets of weight restrictions), these intervals still represent the highest possible
231 performance for each set of weight restrictions. REA by contrast produces robust
232 information about organizational performance in the sense that the resulting intervals
233 reflect the full range of rankings that DMUs may attain for all feasible (from most to least
234 advantageous) weights.

235

236 Second, REA calculates pairwise comparisons between DMUs rather than comparing each
237 DMU to an efficiency frontier as in DEA or stochastic frontier analysis. This makes REA
238 results more robust than frontier-based results, since the introduction or removal of an
239 outlier DMU can substantially change the location of the efficiency frontier (Banker et al.,

240 1986). In contrast, already established pairwise dominance relations obtained from REA
241 cannot change if a new DMU is added; and the end points of any DMU's ranking interval can
242 shift towards lower performance by at most one ranking.

243
244 Third, because REA is based on pairwise comparisons, it requires a minimum of only two
245 DMUs. In contrast, frontier-based methods require a larger number of DMUs to construct
246 the frontier. For DEA, for instance, Banker et al. (1986) proposed the simple rule of thumb
247 that the number of DMUs should be at least three times the number of variables. This is
248 problematic because the number of indicators typically far outstrips the number of
249 organisations.

250
251 Where the choice of denominator is straightforward, ratio-based analysis is not necessary.
252 One can calculate individual performance rates for the respective indicators and aggregate
253 them as a weighted sum as in equation (1). This is akin to evaluating the numerator of the
254 performance ratio (2).

255
256 We here use ratio-based analysis in order to illustrate robustness to different choices of
257 denominator while, which is an important innovation of REA, simultaneously varying the
258 numerators weights. Ratio-based measures have limitations. In particular, the use of a ratio
259 function does not account for structural differences (such as a higher share of fixed costs)
260 between organisations. This assumption implies that, in evaluating organisational
261 performance, one does for instance not "allow" an organisation a higher number of HAIs (in
262 ratio terms, e.g. per 100,000 population) only because it is relatively small in size. However,

263 this assumption seems justified in contexts where health policy objectives include the
264 principle of ensuring equal quality of care regardless of a person's place of residence and
265 where structural differences have been compensated for (e.g. via the funding system, as
266 outlined below) so as to ensure a level playing field across organisations.

267
268 Ratio measures may be preferred when there is primarily a concern with evaluation
269 (examining which organisations perform better or worse) rather than explanation
270 (examining why organisations achieve particular performance outcomes, as in regression
271 analysis). This paper addresses the problem of comparative evaluation.

272 **3.3 System context and data**

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

284

285 ***Part I:*** To examine robustness to choices of weights, we analyze six indicators from the
286 HEAT target system intended to measure Boards' relative degree of achievement in
287 ensuring appropriate treatment. This analysis uses an additive model akin to analyzing the
288 numerator of the performance ratio (equation 2) subject to uncertainty about weights.

289 ***Part II:*** To examine robustness to alternative choices of denominator alongside uncertainty
290 about numerator weights, we relate the number of two types of HAIs (MRSA/MSSA and
291 C.difficile infections) to OBDs and total population. This analysis relies on the more
292 complex ratio-based model in equation (2). We focus on HAIs because there is a good
293 justification for two alternative denominators (as set out in section 2.2). REA-based
294 analyses with two numerators and two denominators thus show the full strength of the
295 ratio-based approach.

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

299

300 **Environmental adjustment.** The 14 Health Boards differ in terms of demographic,
301 epidemiological and regional factors which are beyond their control but might influence
302 observed performance. However, in Scotland, Health Boards are allocated resources based
303 on a formula that takes account of variations in healthcare needs which arise from
304 differences in age and sex composition, morbidity, life circumstances, and excess costs of
305 delivering services in some (especially rural) regions which are deemed unavoidable (ISD

306 Scotland, 2010). Thus, Boards have already been compensated for structural differences so
307 that they can ensure the same level of quality. We acknowledge that the risk adjustment
308 provided by this formula is not perfect. However, following this argument, it is not
309 unreasonable to assume that Boards are comparable with respect to the performance
310 indicators analysed here.

311 *Tables 1 and 2 about here*

312 **3.4 Weight restrictions on quality measures**

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

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

325

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

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

338
339 In flexible weighting systems, the composite score may be heavily influenced by a sub-
340 indicator that is marginally important in the wider health system context (Goddard and
341 Jacobs, 2009). To address this problem, for Part I we made the (illustrative but reasonable)
342 assumption that avoiding a particular event can at most have half of the overall value
343 attached to avoiding an event of each of the six quality measures. This resulted in the
344 following proportional weight restrictions: avoiding an event of the worst healthcare
345 quality measure cannot be more than ten times as valuable as avoiding an event of the least
346 bad quality measure (since with six indicators, a minimum weight of 1/10 means that one
347 quality measure can have at most half of the weight mass).

348

349 For part II, we made the (illustrative but reasonable) assumption that avoiding one
350 *C.difficile* infection must be at least 1/4 as valuable as avoiding one MRSA/MSSA infection.
351 No weight restrictions for denominator variables were used. In efficiency analysis,
352 denominator weights have a clear interpretation, as they indicate the substitutability of
353 different types of inputs (labor, capital, intermediate inputs). In quality comparisons,
354 denominators represent different populations at risk. However, denominator weights lack
355 a clear interpretation as in efficiency analysis since it is hard to think about trade-offs
356 between different populations at risk.

357

358 **4 Results**

359 **4.1 Robustness to choices of weights: Unrestricted and restricted** 360 **ranking intervals for feasible weight sets**

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

367

368 However, the impact of weight restrictions on reductions in uncertainty differs across
369 Boards. For Boards *L* and *H*, ordinal weight restrictions narrow the ranking interval from

370 11 respectively 12 ranks (Figure 1) to 3 possible ranks (Figure 2), thus clarifying Board
371 performance. In contrast, for Boards *N*, *E*, *M* and *A*, ranking intervals remain wider, because
372 these Boards perform better on some indicators, but worse on others (Table 2). Hence, the
373 remaining flexibility to set weights influences the ranks these Boards may attain. For 7 out
374 of 14 Boards (*K*, *F*, *B*, *E*, *C*, *A*, *J*), the additional use of proportional weight restrictions
375 (Figure 3) further decreases uncertainty about relative ranks.

376

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

383

Figures 1 to 3 about here

384 **4.2 Dominance relations and comparative scope for improvement**

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

391

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

403

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

411

412 Multiple values on the same row mean that a Board is dominated by several Boards and
413 would be situated on lower levels of the dominance graph. Looking horizontally, one can

414 see the improvements needed for the five worst performing Boards J, G, D, M, A to become
415 non-dominated by the better-performing Boards. Looking vertically, one can identify the
416 distance that differentiates each Board from the national leaders, Boards K, L and N.

417 *Figure 4 about here*

418 *Table 3 about here*

419

420 **4.3 Ratio-based analysis: Robustness to choice of denominator**

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

427

428 REA-based ranking intervals, which show composite performance on MRSA/MSSA and
429 C.difficile relative to OBDs and population, reveal seven Boards (marked in bold in Table 4)
430 with a ranking interval spanning seven or more ranks. This uncertainty in ranking reflects,
431 first, sensitivity to choice of denominator (e.g. Borders jumps up four ranks when
432 MRSA/MSSA and C.difficile are measured relative to total population). Second, this may
433 show differences in performance on MRSA/MSSA as opposed to C.difficile (e.g. Forth Valley

434 is ranked 13th on the former but 2nd on the latter relative to OBDs).

435 *Table 4 about here*

436

437

438 **5 Discussion**

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

449

450 In a context where information is inevitably incomplete but policy-makers remain
451 interested in an overall measure of health system performance (OECD, 2008), we have
452 demonstrated how REA offers a third way that openly provides indications of the
453 uncertainty inherent in the valuation of objectives and choices of denominators. The
454 approach is essentially based on agnosticism: When there are multiple reasonable
455 denominators which each highlight aspects of performance – such as that an organisation
456 can deliver high quality in terms of few HAIs relative to hospitalised and/or general

457 populations – then analysts need not restrict themselves to a single denominator. Our
458 results reinforce the insight that healthcare quality may be best thought of as a collection of
459 possible rates depending on how the denominator is specified rather than as a single
460 “right” rate (Guillen et al., 2011). Ranking intervals based on multiple denominators thus
461 may enable a more complete account of performance.

462
463 Similarly, if we know that quality measures are heterogeneous but are ignorant of the best
464 method to weight them, then methods to construct composite indicators need to capture
465 that lack of knowledge. Sensitivity analysis on weights is not a new idea; prior work –
466 especially in the multidimensional well-being literature – includes explicit use of ranges of
467 weights (Zhou et al., 2010); computation of multiple weighting schemes (Osberg and
468 Sharpe, 2002); and global sensitivity analysis (Saltelli et al., 2008).

469
470 The REA approach adds to this work in two ways. First, consideration of incomplete
471 information is built into the structure of the model. Ranking intervals give policy-makers a
472 sense of the uncertainty around ranks, indicating the extent to which action is warranted.
473 Our results show that, when one assumes complete ignorance about the relative weights
474 assigned to different indicators, then it is impossible to differentiate the performance of
475 Scottish Health Boards (Figure 1). Thus, one cannot say which organisations perform better
476 or worse. Regulatory action based on such rankings would clearly be premature.

477
478 However, once some reasonable ordinal and proportional weight restrictions are applied,
479 organizational performance appears more clarified. The choice of weight restrictions may

480 differ between groups of people: different individuals may come up with different
481 orderings or proportionate weights concerning the relative badness (or goodness) of
482 particular events. However, if weight restrictions can be established (e.g. based on existing
483 consensus or medical evidence of disease severity), then they may provide useful insights.
484 When an organisation consistently appears at the bottom (Board G) or at the top (Board L;
485 in Figure 2) whichever set of weights is used, this may strengthen the rationale for policy
486 intervention. It supports the notion that settling on a unique set of weights is not always
487 necessary to inform well-founded judgments (Foster and Sen, 1997).

488

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

497

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

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

503 (ii) there are concerns about rank reversals due to sensitivity to outliers and the
504 introduction or removal of organisations (since pairwise comparisons make REA
505 results relatively robust to these biases); and
506 (iii) there are relatively few organisations (since a large number of organisations is
507 not needed to construct an efficiency frontier). However, there are also no inherent
508 limitations to applying REA to large datasets.

509

510

511

512 **6 Implications for policy and research**

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

523

524 Evidence of uncertainty reinforces the need to use the results as signals for further
525 analysis, rather than for definitive judgments. Since weakly correlated indicators will make

526 rankings more sensitive to different sets of weights (Foster et al., 2012), the careful use of
527 weight restrictions becomes particularly important. Weight restrictions will tend to clarify
528 the results and make explicit the impact of subjective choices about the relative value of
529 different quality indicators on performance rankings.

530
531 Dominance relations that are based on pairwise comparisons between Boards provide
532 comparative performance assessments one can be confident about. Since dominance
533 relations indicate that some DMU k performs at least as well as some other DMU l for all
534 feasible weights and there exist some weights for which it performs strictly better, this
535 information could, for instance, be used for setting performance targets across all
536 indicators included in the analysis. Since improvements on some indicators may require
537 less effort than others, indicator-specific improvements would also be informative.
538 However, this would require a different approach. Gouveia et al. (2015), for instance,
539 employ slack-variables (which define the variable-specific distance to the efficiency
540 frontier) to estimate the improvements required for a DMU to reach the best performing
541 organisation. However, this approach does not indicate the improvements needed to reach
542 some specific, non-efficient DMU as it is possible with our approach. This is particularly
543 relevant for policy and management and a strength of our study, since the top performing
544 organisation may not always be the most meaningful (and practically feasible) benchmark
545 for worse performing organisations. In a collegiate rather than competitive environment,
546 such results could help organisations to learn from better performing (dominating) peers.

547
548 For a large number of organisations (and dominance relations), the clear presentation and

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

554
555 Finally, it is essential to re-emphasize the importance of the other methodological choices
556 (listed in section 2) that must be made when constructing a composite indicator; in
557 particular, the initial selection of indicators and risk adjustment for environmental
558 (uncontrollable) determinants of performance. If important indicators are omitted or
559 irrelevant variables are included, then performance evaluations will be meaningless
560 (Smith, 1997). The choice of performance metrics therefore needs to reflect a country's
561 definition of valued outcomes of the health service (Dowd et al., 2014).

562
563 Concerning risk adjustment, in Scotland the funding formula is designed to enable all NHS
564 Boards to produce equal levels of performance. Since this formula takes account of
565 differences in population and local characteristics (e.g. rurality), in this study we have
566 followed the argument that risk adjustment has been implemented via the funding system
567 (Jacobs et al., 2006). However, the degree to which this argument holds depends on the
568 accuracy and comprehensiveness of the formula. While for our study the direction of any
569 potential bias is difficult to determine, it is possible that inadequate risk adjustment has
570 affected observed Board performance on the constituent indicators.

571

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

577

ACCEPTED MANUSCRIPT

578 REFERENCES

- 579 Allen, R., Athanassopoulos, A., Dyson, R. G. & Thanassoulis, E. (1997). Weights restrictions and value
580 judgements in Data Envelopment Analysis: Evolution, development and future directions. *Annals of*
581 *Operations Research*, 73, 13-34.
- 582 Appleby, J. & Mulligan, J. (2000). *How well is the NHS performing? A composite performance indicator based on*
583 *public consultation*, London, The King's Fund.
- 584 Banker, R. D., Conrad, R. F. & Strauss, R. P. (1986). A Comparative Application of Data Envelopment Analysis
585 and Translog Methods - an Illustrative Study of Hospital Production. *Management Science*, 32, 30-44.
- 586 Bertsimas, D., & Tsitsiklis, J. N. (1997). *Introduction to linear optimization* (Vol. 6). Belmont, MA: Athena
587 Scientific.
- 588 Bevan, G. & Hamblin, R. (2009). Hitting and missing targets by ambulance services for emergency calls:
589 Effects of different systems of performance measurement within the UK. *Journal of the Royal*
590 *Statistical Society. Series A*, 172, 161-190.
- 591 Carinci, F., Van Gool, K., Mainz, J., Veillard, J., Pichora, E. C., Januel, J. M., et al. (2015). Towards actionable
592 international comparisons of health system performance: expert revision of the OECD framework
593 and quality indicators. *International Journal for Quality in Health Care*, 27, 137-146.
- 594 Cherchye, L., Moesen, W., Rogge, N. & Van Puyenbroeck, T. (2007). An introduction to 'benefit of the doubt'
595 composite indicators. *Social Indicators Research*, 82, 111-145.
- 596 CMS (2009). Centers for Medicare & Medicaid Services. Premier Hospital Quality Incentive Demonstration:
597 Fact sheet.
598 <http://www.cms.hhs.gov/HospitalQualityInits/downloads/HospitalPremierFactSheet200907.pdf> [9
599 May 2014].
- 600 Decancq, K. & Lugo, M. A. (2012). Weights in multidimensional indices of wellbeing: An overview.
601 *Econometric Reviews*, 32, 7-34.
- 602 Dolan, P. (1997). Valuing health states: A comparison of methods. *Journal of Health Economics*, 16, 617-617.
- 603 Dowd, B., Swenson, T., Kane, R., Parashuram, S. & Coulam, R. (2014). Can Data Envelopment Analysis Provide
604 A Scalar Index Of 'Value'? *Health Economics*, 23, 1465-1480.
- 605 Foster, J. & Sen, A. (1997). *On Economic Inequality*, Oxford, Oxford University Press.
- 606 Foster, J. E., McGillivray, M. & Seth, S. (2012). Composite Indices: Rank Robustness, Statistical Association, and
607 Redundancy. *Econometric Reviews*, 32, 35-56.
- 608 Goddard, M. & Jacobs, R. (2009). Using composite indicators to measure performance in health care. In: Smith,
609 P., Mossialos, E., Papanicolas, I. & Leatherman, S. (eds.) *Performance measurement for health system*
610 *improvement: experiences, challenges and prospects*. Cambridge: Cambridge University Press, pp. 339-
611 368.
- 612 Gouveia, M., Dias, L., Antunes, C., Mota, M., Duarate, E. & Tenreiro, E. (2015). An application of value-based
613 DEA to identify the best practices in primary health care. *OR Spectrum*. DOI 10.1007/s00291-015-
614 0407-x.
- 615 Guillen, Ú., Demauro, S., Ma, L., Zupancic, J., Wang, E., Gafni, A. & Kirpalani, H. (2011). Survival rates in
616 extremely low birthweight infants depend on the denominator: avoiding potential for bias by
617 specifying denominators. *American Journal of Obstetrics and Gynecology*, 205, 329.e1-329.e7.
- 618 Hauck, K. & Street, A. (2006). Performance assessment in the context of multiple objectives: a multivariate
619 multilevel analysis. *Journal of Health Economics*, 25, 1029-48.
- 620 Healthcare Commission (2005). *2005 Performance Ratings*, London, Healthcare Commission.
- 621 Health Protection Scotland (2007). Annual report on the surveillance of Clostridium difficile associated
622 disease (CDAD) in Scotland, Glasgow, Health Protection Scotland.
- 623 Health Protection Scotland (2014). *Healthcare Associated Infection Annual Report 2013*, Glasgow, Health
624 Protection Scotland.
- 625 Hollingsworth, B. & Street, A. (2006). The market for efficiency analysis of health care organisations. *Health*
626 *Economics*, 15, 1055-1059.
- 627 Hume, D. (1739). *A Treatise of Human Nature*. London: John Noon.
- 628 Hussey, P. S., De Vries, H., Romley, J., Wang, M. C., Chen, S. S., Shekelle, P. G. & Mcglynn, E. A. (2009). A
629 systematic review of health care efficiency measures. *Health Services Research*, 44, 784-805.

- 630 ISD Scotland (2010). Resource allocation. <http://www.isdscotland.org/Health-Topics/Finance/Resource->
631 [Allocation-Formula/information.asp](http://www.isdscotland.org/Health-Topics/Finance/Resource-Allocation-Formula/information.asp) [11 November 2014].
- 632 Jacobs, R., Goddard, M. & Smith, P. (2005). How robust are hospital ranks based on composite performance
633 measures? *Medical Care*, 43, 1177-84.
- 634 Jacobs, R., Smith, P. & Street, A. (2006). *Measuring efficiency in health care: analytic techniques and health*
635 *policy*, Cambridge University Press.
- 636 Marlow, A. (1995). Potential years of life lost: what is the denominator? *Journal of Epidemiology & Community*
637 *Health*, 49, 320-2.
- 638 Mckibben, L., Horan, T., Tokars, J. I., Fowler, G., Cardo, D. M., Pearson, et al. (2005). Guidance on Public
639 Reporting of Healthcare-Associated Infections: Recommendations of the Healthcare Infection Control
640 Practices Advisory Committee. *American Journal of Infection Control*, 33, 217-226.
- 641 Naessens, J. M. & Huschka, T. R. (2004). Distinguishing hospital complications of care from pre-existing
642 conditions. *International Journal for Quality in Health Care*, 16, I27-I35.
- 643 Navarro, V. (2000). Assessment of the world health report 2000. *Lancet*, 356, 1598-1601.
- 644 NHS England (2015). *Revised Never Events Policy and Framework*, London, NHS England.
- 645 OECD (2008). *Handbook on Constructing Composite Indicators*, Paris, OECD.
- 646 Osberg, L. & Sharpe, A. (2002). An index of economic well-being for selected OECD countries. *Review of*
647 *Income and Wealth*, 48, 291-316.
- 648 Pedraja-Chaparro, F., Salinas-Jimenez, J. & Smith, P. (1997). On the Role of Weight Restrictions in Data
649 Envelopment Analysis. *Journal of Productivity Analysis*, 8, 215-230.
- 650 Reeves, D., Campbell, S. M., Adams, J., Shekelle, P. G., Kontopantelis, E. & Roland, M. O. (2007). Combining
651 multiple indicators of clinical quality: an evaluation of different analytic approaches. *Medical Care*,
652 45, 489-96.
- 653 Romano, P., Hussey, P. & Ritley, D. (2010). *Selecting Quality and Resource Use Measures: A Decision Guide for*
654 *Community Quality Collaboratives*, Rockville, Agency For Healthcare Research And Quality.
- 655 Salo, A. & Punkka, A. (2011). Ranking intervals and dominance relations for ratio-based efficiency analysis.
656 *Management Science*, 57, 200-214.
- 657 Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., et al. (2008). *Global Sensitivity Analysis:*
658 *The Primer*, Wiley E-book.
- 659 Schlaud, M., Brenner, M. H., Hoopmann, M. & Schwartz, F. W. (1998). Approaches to the denominator in
660 practice-based epidemiology: a critical overview. *Journal of Epidemiology & Community Health*, 52,
661 13S-19S.
- 662 Scottish Government (2014). NHS Boards. <http://www.scotland.gov.uk/Topics/Health/NHS->
663 [Workforce/NHS-Boards](http://www.scotland.gov.uk/Topics/Health/NHS-Workforce/NHS-Boards) [14 May 2014].
- 664 Smith, P. (1997). Model misspecification in data envelopment analysis. *Annals of Operations Research*, 73,
665 233-252.
- 666 Smith, P. (2002). Developing composite indicators for assessing health system efficiency. In: SMITH, P. (ed.)
667 *Measuring up: improving health system performance in OECD countries*. Paris: OECD, pp. 295-318.
- 668 Smith, P. C. (2003). Formula funding of public services: An economic analysis. *Oxford Review of Economic*
669 *Policy*, 19, 301-322.
- 670 WHO (2000). *The world health report 2000 - Health systems: improving performance*, Geneva, World Health
671 Organization.
- 672 Zhan, C. L., Elixhauser, A., Friedman, B., Houchens, R. & Chiang, Y. P. (2007). Modifying DRG-PPS to include
673 only diagnoses present on admission - Financial implications and challenges. *Medical Care*, 45, 288-
674 291.
- 675 Zhou, P., Ang, B. W. & Zhou, D. Q. (2010). Weighting and aggregation in composite indicator construction: A
676 multiplicative optimization approach. *Social Indicators Research*, 96, 169-181.

677 **TABLES AND FIGURES**678 **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

679 Sources: ^aHEAT target system; ^bNational Records of Scotland. All data are for 2012/13.

680

681 **Table 2 Comparative performance of Boards on the constituent six quality**
 682 **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

683

684

685 **Table 3 Comparative scope for improvement needed to reach another target or**
 686 **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													

687

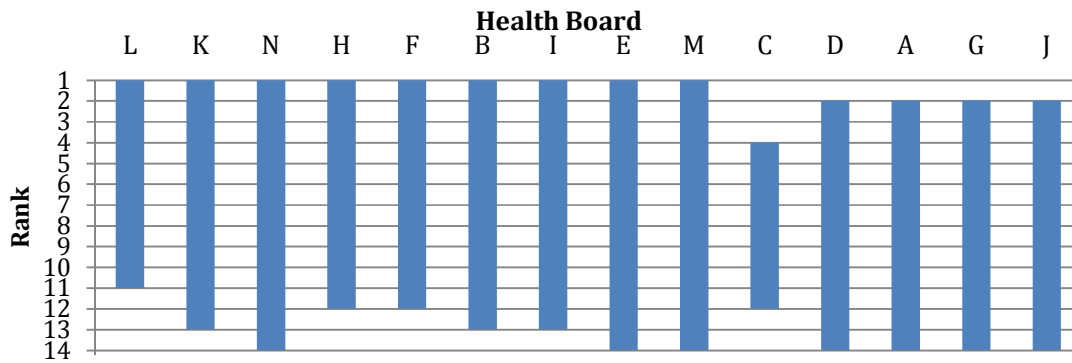
688

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

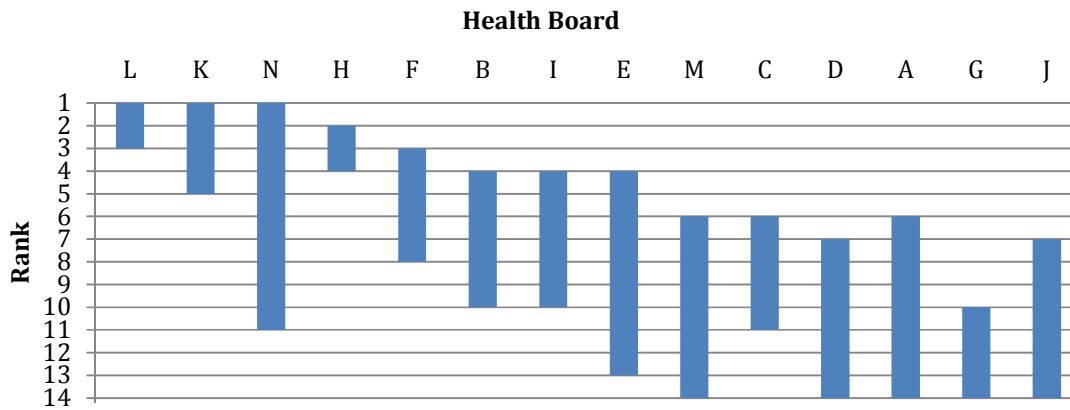
Board	Per 100,000 OBDs		Per 100,000 population		Per 100,000 OBDs		Per 100,000 population		Ranking interval for composite performance on MRSA/MSSA and C.difficile relative to OBDs and population
	Number of MRSA/MSSA	Rank	Number of MRSA/MSSA	Rank difference compared to OBDs	Number of C.difficile	Rank	Number 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

690

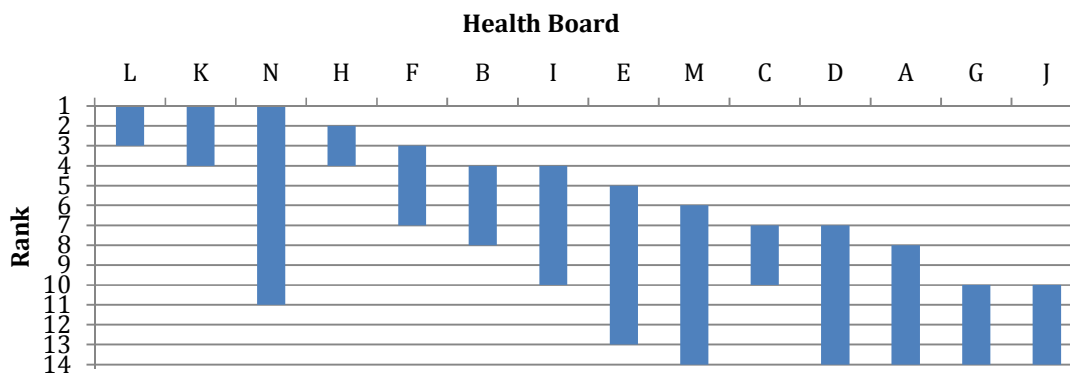
691

692 **Figure 1 Performance rankings for all feasible weights**

693

694 **Figure 2 Performance rankings with ordinal weight restrictions**

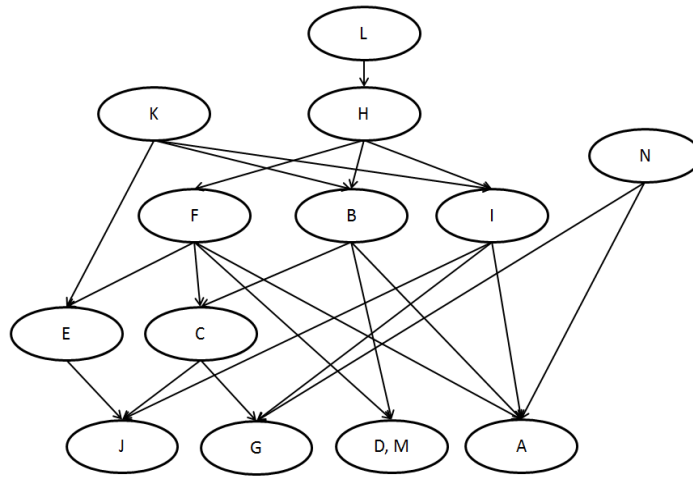
695

696 **Figure 3 Performance rankings with ordinal and proportional weight restrictions**

697

698

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



701

Role of the funding source : The research was supported by the New Professors Fund at the University of Strathclyde. The sponsor had no role in study design; in the collection, analysis and interpretation of data; in the writing of the article; and in the decision to submit it for publication.

Acknowledgements: We thank Andrew Street, Irene Papanicolas, Yan Feng, four anonymous reviewers, and the participants of the HESG Glasgow 2014 and of the EURO conference Glasgow 2015 for their helpful comments. The authors are responsible for any errors.

Research highlights

- Proposes a method to handle lack of information on weights and denominators in composite metrics
- Ranking intervals and dominance relations show performance rankings one can have confidence in
- Quality comparisons of Scottish Health Boards illustrate the impact of incomplete information