WP 23/14

Multi-directional Robust Benefit of the Doubt model: a comprehensive measure for the quality of health care in OECD countries

Francesco Vidoli; Elisa Fusco; Giacomo Pignataro and Calogero Guccio

August 2023

Multi-directional Robust Benefit of the Doubt model: a comprehensive measure for the quality of health care in OECD countries

Vidoli F.[a], Fusco E.[b], Pignataro G.[c,d], Guccio C.[c]

[a] *University of Urbino Carlo Bo, Department of Economics, Society and Politics, Via Aurelio Saffi, 42, Urbino, 61029, PU, Italy*
[b] *SOGEI S.p.A, Department of Economic Modelling and Statistical Analysis for Policy Making, Via M. Carucci n. 99, Rome, 00143, RM, Italy*
[c] *University of Catania, Department of Economics and Business, Corso Italia, 55, Catania, 95129, CT, Italy*
[d] *Politecnico di Milano, Department of Management, Economics and Industrial Engineering, Via Raffaele Lambruschini, 4/B, Milano, 20156, MI, Italy*

Abstract

While individual indicators in health care quality assessment provide detailed insights into specific aspects, they often fail to capture the full and relevant information. Consequently, there is a growing need to develop composite measures that comprehensively assess the overall quality or performance of health care systems, especially those not covered by official OECD measures. A novel multi-directional robust Benefit of the Doubt approach is proposed to measure overall health care quality, through a composite indicator, while, at the same time, highlighting the potential improvement directions for each single component indicator. The method is developed within a robust framework. To show its advantages, the approach is applied, first, to simulated data, and then to country-level OECD data, drawn from the Healthcare Quality and Outcomes program, relative to acute care services.

Keywords: Robust composite indicators, Non-compensatory, Multi-Directional Benefit of the Doubt, Health care quality
JEL: C14, C43, C44, I18

[*] Corresponding author

URL: francesco.vidoli@uniurb.it (Vidoli F.), fusco_elisa@libero.it (Fusco E.), giacomo.pignataro@unict.it (Pignataro G.), guccio@unict.it (Guccio C.)

1. Introduction

The need to ensure that the vast resources devoted to health care in most developed countries produce substantial results for the health of the populations of those countries has driven the search for ways to measure these results. To quickly and conveniently identify the research context to which the application of the methodology developed in this paper is intended to contribute, we refer to the field of quality measurement in health care.

The main problem in measuring quality is its inevitable multidimensionality, due to its very general meaning. This is also due to the multiple dimensions of the concept of health (or good health) and the multiplicity of different health needs addressed by different treatments. A well-known and broad classification of the different dimensions of quality of health care provision is provided by Donabedian (1966), who identifies three areas for its measurement: outcome, "*in terms of recovery, restoration of function and of survival*" (Donabedian, 1966, p. 167); the process of care, since "*one is interested not in the power of medical technology to achieve results, but in whether what is now known to be "good" medical care has been applied*" (Donabedian, 1966, p. 169); structure, that is the "*settings in which it [the process of care] takes place and the instrumentalities of which it is the product*" (Donabedian, 1966 p. 170). Within these three headings, it is possible to record an extremely large effort of development of specific measures, related to the different aspects of each heading, the different treatments, the different areas of care as well as the different health needs. As a practical result, we can now read statistics of quality of care, at the national as well as at the international level, which are made of a very large number of indicators. Beaussier et al. (2020), in a study on how statutory hospital regulators in four countries (France, England, Germany and the Netherlands) measure quality of health care provision, surveyed 1,100 different indicators of quality of care[1]. At the international level, the WHO (World Health Organization) and the OECD (Organisation for Economic Co-operation and Development)

[1] They also report that "*In England's National Health Service (NHS) for example, the number of performance indicators has skyrocketed from 70 in 1982 to more than 2000 today. Likewise in the US, the number of healthcare quality indicators endorsed by the National Quality Forum has more than doubled over the last decade to 1078.*" (Beaussier et al., 2020, p. 501).

have developed sets of indicators for measuring the performance of health-care systems[2]., which are not as large as the ones at the national level, but still, they include numerous measures.

While the multiplicity of indicators can provide granular information on very specific aspects of the quality of care, there is other relevant information for health care policy (as well as for managerial actions) that cannot be conveyed by single indicators or by a set of indicators and, therefore, there has been now a longstanding claim for building up composite measures of quality of care. Smith (2002) noted that "*the broad arguments for developing a composite indicator of performance are that it offers a more rounded assessment of system performance than piecemeal inspection of individual performance indicators, and that it facilitates judgments on overall system efficiency*" (Smith, 2002, p. 298). Jacobs and Goddard (2007) add that composite indicators allow "*focusing attention on important policy issues, offering a more rounded assessment of performance and presenting the 'big picture' in a way which the public can understand. In contrast to piecemeal indicators based on individual performance measures, they can offer policy-makers a summary of complex multi-dimensional issues... They provide an attractive option for accountability purposes, as it is easier to track the progress of a single indicator over time rather than a whole package of indicators*" (Jacobs and Goddard, 2007, p. 103).

Since the need for composite measures is real, it is easy to find several attempts to develop and use such indicators for measuring the overall quality (or more in general, performance) of health care provision. Kara et al. (2022) provide an updated and large survey of the use of different approaches to construct composite measures of quality of care. Among the most well-known efforts, which have also been the subject of field applications, it is possible to mention the World Health Report 2000 by the WHO[3], the English NHS star rating system for hospitals (and, afterward, also for primary care trusts)[4], the Hospital Compare Overall Hospital Quality Ratings by the US Centers for

[2] For WHO, please see https://www.who.int/data/data-collection-tools/harmonized-health-facility-assessment/introduction; for OECD https://www.oecd.org/health/health-systems/health-care-quality-outcomes-indicators.htm

[3] https://www.who.int/publications/i/item/924156198X

[4] Since the NHS documents are no longer publicly accessible at the original URLs, see a brief representation of the star rating system in Jacobs et al. (2006).

Medicare and Medicaid Services[5]. There are several flaws in these as in other composite indicators, which have been widely discussed (*e.g.*, Smith, 2002; Jacobs et al., 2005; Jacobs and Goddard, 2007; Shwartz et al., 2015; Barclay et al., 2019; Friebel and Steventon, 2019; Hofstede et al., 2019). They are not a mere list of potential disadvantages but, at least some of them, also affect the reliability of the information they can convey to decision-makers, above all when they are presented in the form of rankings of providers (NHS and CMS star rating) or systems (WHO report).

In this paper, we aim to contribute to the literature on composite indicators and their use in measuring the quality of health care by using a methodology that addresses some of the main drawbacks of composite indicators and their specific implications for the area of our interest. Firstly, the main problem faced in any attempt to use a composite measure is how to weight the component indicators. This is a particularly sensitive issue in health care, as the different indicators to be included in a single index may refer to objectives related to the health of different individuals, characterized, for example, by different health needs. We believe that the main problem here, rather than the heterogeneity of the weights of the individual indicators, is the potential heterogeneity of the set of weights across the different units under consideration (health systems, providers, etc.). Since the prioritization of different health needs falls either in the realm of social value judgments at the system level (which can be heterogeneous across different countries) or depends on the composition of health needs in a given geographical area, for providers that have a service obligation to cover the needs of their population, assessing the performance of different decision makers, in terms of quality of their provision of care, on the basis of a uniform set of weights, may produce misleading information. This is the reason why we depart from the prevailing methodologies of measuring composite indicators, in the health care field, and our exercise is run within the group of methodologies known as Benefit of the Doubt (`BoD` - Nardo et al., 2008). The `BoD` approach is basically derived from the non-parametric frontier analysis, and it is characterized by the endogenous derivation of the weights for each single unit under exam. To the best of our knowledge, there have not been attempts of applying `BoD` to provide a composite measure of the quality of

[5] https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/hospitalqualityinits/hospitalcompare

health care provision, except for two recent papers by Matos et al. (2021) and Pereira et al. (2021b) that use a directional BoD measure to study the convergence of the Member States of the World Health Organization in terms of the United Nations' Sustainable Development Goal (SDG) 'Good health and well-being'.

However, our study differs from theirs, since we focus on the quality of care, on the basis of a set of indicators belonging to a consolidated framework of analysis based on the twenty-year-long experience of the OECD. Matos et al. (2021) consider, instead, four different areas of evaluation: quality, efficiency, access, and finance, with a choice of indicators, in some cases, left to the discretion of the authors. Moreover, their technique is based on `BoD` models that do not take into account some recent developments, which increase their robustness and their informational content, as we shall explain in presenting our methodology.

We also differ from the current, and prevailing, methods of calculating composite indicators in that we do not assume perfect compensability of the services measured by the different indicators, since a social assessment of the quality of health care may well take into account that potential health losses due to low quality in some treatments or areas of care are not necessarily compensated by other potential health gains due to higher quality in other treatments or areas of care.

Finally, the methodology used in our work allows us to relate overall performance, as measured by the composite indicator, to the individual indicators' performances, by measuring the potential improvement of each of them. We exploit the latest advances in the development of `BoD` approach, using the Multi-directional Benefit of the Doubt (`MDir_BoD`) proposed by Fusco (2023). Moreover, we develop this method so as to enhance its robustness in dealing with outliers, thus providing a contribution also to the `BoD` literature.

The paper proceeds as follows. In section 2, we will briefly survey the `BoD` literature, to stress how the approach, in general, and the specific models we are going to use can address the specific questions arising from the objective of measuring composite indicators for the quality of health care provision. In the following section 3, we provide technical aspects of our original methodology called Multi-directional Robust Benefit of the Doubt (`MDir_RBoD`) composite indicator. In section 4, we present a simulation for testing our methodology, while in section 5 we apply the `MDir_RBoD` to assess the quality of health care using OECD data. Finally, the last section provides some concluding remarks.

2. Literature

One of the most debated issues in the literature on composite indicators, and also in the literature on their application to health care, is the way in which the individual indicators are weighted. In their survey of the works on the use of composite measures in health care, Kara et al. (2022) show the different methods used, within the surveyed works, for assigning weights to single indicators. They range from the "simplest" solution of assigning equal weights to all indicators (or, equivalently, assigning no weights) to a choice based on experts' opinions or on statistical techniques, like Principal Component Analysis. While the choice of a particular set of weights may severely affect the outcome of the composite indicator, there is no theoretical construct that can consistently address this choice, since it essentially reflects "value judgments" about the relative importance of the component indicators. In the field of health care quality indicators, weighting implies the attribution of a "priority", in the measurement of the overall quality of a provider or of the entire healthcare system, for instance, to the outcomes of given treatments (and, consequently, of given patients) relative to others, or to the results in some areas of care (*e.g.*, hospital care) over others (*e.g.*, primary care), or to given aspects of the provision of care (*e.g.*, effectiveness) relative to others (*e.g.*, responsiveness). In this as in other fields, of course, it is feasible and meaningful to prioritize across different objectives. However, when using composite indicators to compare different health care systems, or different providers (like hospitals, for instance) enjoying relevant degrees of autonomy in their allocation choices, "imposing" a unique set of weights, that is a unique set of priorities over the different objectives as represented by the different indicators, may be misleading. In general, following Jacobs et al. (2004), "*The weights used reflect a single set of preferences, whilst the evidence suggests there exist a great diversity in preferences across policy makers, individual unit actors and the broader public. There is likely to be considerable variation in the preferences of respondents*" (Jacobs et al., 2004, p. 49). What Jacobs et al. (2004) call preferences may well arise from the prevalence of different diseases, which may not be uniform across different countries/providers' areas, and the relevant decision-makers make a differential effort in treating the different diseases, consistently with their prevalence. It may also be that, at the system level, for instance, the relevant decision-makers favor some areas of care (*e.g.*, hospital care) with respect to others (*e.g.*, primary care) because of the geographical dispersion of the population. What is relevant, in terms of the choice of weights for the composite measure, is that a unique set of weights for all the benchmarked units may not be meaningful for assessing their overall quality performance in the provision of health care.

Our choice of a model within the BoD group (Nardo et al., 2008) allows for the heterogeneity of weighting across the different units, in a way that assumes that each unit, in its own context, makes allocation decisions, in pursuing the different objectives measured by the single indicators, to maximize their performance. This method (or better this family of methods) provides a comprehensive framework for evaluating the relative performance of health care systems by integrating multiple dimensions of health care provision into a single metric. BoD frontier composite indicators are derived from the concept of frontier analysis, which originates from the field of production economics.

Starting with the seminal papers of Cherchye et al. (2005, 2007) and Zhou et al. (2010), the baseline BoD method served as a basis for theoretical improvements regarding the aggregation criterion (Karagiannis, 2017; Rogge, 2018b,a; Verbunt and Rogge, 2018; Van Puyenbroeck and Rogge, 2017). The basic BoD model, apart from the undoubted advantages of the endogenous derivation of the weights, suffers from certain disadvantages due to the linear optimization procedure itself. One of the main problems is related to the perfect compensability of simple indicators. Perfect compensability assumes that all dimensions of performance are equally important and that deficiencies in one dimension can be perfectly compensated for by strengths in other dimensions. From a managerial point of view this means, as pointed out in Fusco (2023), that "*enhancements in the composite indicator continue to follow the same proportion of the actual mix of them, i.e., a larger weight is assigned to a simple indicator on which the unit performs well, instead of encouraging an increase in indicators on which the unit performs worse*" and also *"from an economic point of view, this means stick to the past not really changing for the better the unit condition"*.

While this can simplify indicator construction and facilitate comparisons, it can also result in a loss of valuable information, since important variations and nuances in individual dimensions of performance may be masked. This problem has a special relevance in the health care field. Even if, as discussed before, it is possible that different decision makers may behave differently in terms of the priorities assigned to the different objectives, if the composite indicator has to reflect a "social" evaluation of the quality of health care, above all when the measure is referred to health outcomes for different treatments (and patients), it may not be "acceptable" that underperformance in one area is entirely compensated by overperformance in some other areas of care, since underperformance in the pursuit of some objectives may entail, directly or indirectly, a health loss (or a loss of health gain), which cannot be compensated by any other health gain. In addition, maintaining the same mix as in the past may not be an optimal management strategy for achieving a balanced health care system.

For these reasons, starting with Fusco (2015) Directional BoD (`DBoD`) proposal, many authors have revised the basic `BoD` approach to a directional perspective including weight restrictions (D'Inverno and De Witte, 2020; Pereira et al., 2021a), undesirable indicators (Fare et al., 2019), considering the subindexes as non-compensatory (Mahdiloo et al., 2023) and extending the directional approach to panel data (Oliveira et al., 2020). In the meantime, some empirical directional papers have been proposed for public health (Pereira and Marques, 2022), in the field of education (Sahoo et al., 2017) and for evaluating the environmental performance of municipal utilities (Mergoni et al., 2022). However, the choice of the optimal or preferred direction has always remained an inherent limitation within a procedure that aspired to be not tied to subjective choices. Fusco (2023) has proposed the *Multi-directional Benefit of the Doubt model* (`MDir_BoD`), an extension of the `BoD` that introduces the non-compensability among simple indicators by finding a unit-specific preference structure (the direction) directly from the data. This can be achieved by separating the benchmark selection from the efficiency measurement. Instead of an implicit selection based on improvements proportional to the actual mix of dimensions, as in `BoD`, or in favor of a preferred mix, as in `DBoD`, the benchmark selection is based on adjustments in simple indicators proportional to the potential improvements given by the input/output specific excesses. From a technical point of view, `MDir_BoD` finds an *ideal* vector of simple indicators for each unit and moves in the direction needed to reach this *potential* value. In particular, following Fusco (2023), *"with MDBoD they [the units whose performance is assessed] reach the frontier by enhancing more the simple indicator where they perform worse and, given the unbalanced mix, a low specific efficiency in a simple indicator is not compensated by a high specific efficiency in the other one".* This means that the method finds, other than the global composite score, specific scores for each simple indicator, and therefore, they can be examined separately and used to make practical recommendations for improving the performance of the unit. This is not possible in BoD and DBoD because information about input-specific or output-specific (simple indicator-specific) inefficiencies is masked. This feature is of particular relevance for the use of composite indicators in the health care field. Jacobs et al. (2004) include in their list of potential drawbacks of the use of composite indicators the risk that "*as measures become aggregated it becomes more difficult to determine the source of poor performance and where to focus remedial action*" (Jacobs et al., 2004, p. 2). This risk is especially serious when composite indicators are used as a ranking device of the units under exam (countries, providers), as it happens with the most well-known applications of composite measures (WHO, English star rating, CMS hospital compare). In this case, even the information provided by the score of the composite measure is flattened by the instrumentality of its use for the positioning of the units in the ranking, which is the sort of dominant "message" arising from rankings. As noted by Oliver (2012), with respect to international rankings of health care systems, "*Some countries seem to perform very well on specific aspects of health care, and those from other countries should attempt to learn how they do this, and deduce whether policies can be transferred to and within the institutional structure of their own system without undermining other important health policy goals*" (Oliver, 2012, p. 17)[6]. This outcome, however, can be achieved only if the information from the composite measure is complemented with other information on how the single components of the composite measure contribute to the overall performance as it happens with MDir_BoD.

However, MDir_BoD, like BoD and DBoD, is not robust to outliers, causing the frontier to shift toward the outlier and underestimate the performance scores of all other observations. Robustness to outlier data, in fact, is crucial in composite indicators as it enhances the reliability, credibility, comparability, and policy relevance of the indicator's results given that it ensures the reliability and stability of the composite indicator results over time and across different data sources or variations in the methodology, enhances the credibility and trustworthiness of composite indicators among policymakers and the stakeholders, allow for meaningful comparisons, enabling policymakers to identify best practices, areas requiring improvement, and potential inter-

[6] On this issue, see also Street and Smith (2021).

ventions when conducting comparative analysis between different countries. From a methodological point of view, Vidoli and Mazziotta (2013) proposed a robust approach to the BoD method (called RBoD) using a resampling procedure aimed at decreasing the effect of outliers on the scores of the other units under analysis. The Robust Directional BoD (RDBoD, Vidoli et al., 2015) and the later spatial extension (Fusco et al., 2020) are intended to combine the advantage of robust estimation within a directional model. What we do, in this paper is to extend the MDir_BoD so as to achieve robustness.

3. Robust Multidirectional BoD

Against this background, in order to enhance the lack of robustness control in MDir_BoD approach, by following RBoD and RDBoD methods, a resampling procedure of the MDir_BoD is here proposed.
The underlying idea is to repeatedly compare each unit to subsets of observations of size $m < N$ instead of the entire dataset, thus obtaining a maximal expected frontier of order m and reducing the effect of outliers[7].

In formal terms, as usual, let's consider a matrix of q simple indicators treated as outputs ($Y_q \in \mathbb{R}_+, \forall q = 1, ..., Q$) and an input vector equal to one for all the N observations i. The probabilistic formulation of the production set is defined as:

$$H(\mathbb{1}, \mathbf{y}) = Prob(X \equiv \mathbb{1}, \mathbf{Y} \geq \mathbf{y}) \tag{1}$$

where Ψ is the support of $H(\mathbb{1}, \mathbf{y})$:

$$\Psi = \left\{ (\mathbb{1}, \mathbf{y}) \in \mathbb{R}_+^{1+Q} | H(\mathbb{1}, \mathbf{y}) > 0 \right\} \tag{2}$$

In accordance with this formulation, the maximum possible increment of the q-th indicator for a specific unit, *i.e.* $\widehat{\mathbf{y}}_q$, is obtained by solving Q linear programming problems, that maximize each indicator q, keeping the remaining simple indicators $\mathbf{y}_{-q}$ fixed, as described in detail in Fusco (2023):

$$\widehat{\mathbf{y}}_q = \sup \{\mathbf{y}_q | (\mathbb{1}, \mathbf{y}_q, \mathbf{y}_{-q}) \in \Psi\} \tag{3}$$

where $\widehat{\mathbf{y}}_q$ is a vector of size N.

[7] The relative function will be available on R Compind package (https://cran.r-project.org/web/packages/Compind/index.html).

Then, considering a sample of m (with $m < N$) random variables with replacement $S_m = \{\mathbf{Y_i}\}_{i=1}^m$ drawn from the density of $\mathbf{Y}$, the random set $\widetilde{\Psi}_m$ is defined as:

$$\widetilde{\Psi}_m = \bigcup_{j=1}^{m} \{(\mathbb{1}, \mathbf{y}) \in \mathbb{R}_+^{1+Q} | X \equiv \mathbb{1}, \mathbf{Y_j} \geq \mathbf{y}\}. \tag{4}$$

As said before, since the individual unit is not compared to all others, but to a sample subset of size m, the effect of an abnormal or outlying unit is attenuated.

This generalization enables the iterative computation of the sample subset of size m (for $b = 1, \ldots, B$ times) and, for each iteration b, the maximum possible increment for the single indicator, following equation 3, is given by:

$$\widetilde{\mathbf{y}}_{m;q}^{b} = \sup\left\{\mathbf{y}_q | \left(\mathbb{1}, \mathbf{y}_q, \mathbf{y}_{-q}\right) \in \widetilde{\Psi}_m\right\}, \ \forall b = 1, \ldots, B; \ q = 1, \ldots, Q \tag{5}$$

where $\widetilde{\mathbf{y}}_{m;q}^{b}$ is a vector of size N.

The *potential improvements* of each unit, *i.e.*, the directional vector at iteration b can be calculated as follows:

$$\widetilde{\mathbf{g}}_m^{PI_b} = (\widetilde{\mathbf{y}}_m^b - \mathbf{y}_m) = (\widetilde{\mathbf{y}}_{m;1}^b - \mathbf{y}_{m;1}, \ldots, \widetilde{\mathbf{y}}_{m;q}^b - \mathbf{y}_{m;q}, \ldots, \widetilde{\mathbf{y}}_{m;Q}^b - \mathbf{y}_{m;Q}), \forall b = 1, \ldots, B \tag{6}$$

where $\widetilde{\mathbf{g}}_m^{PI_b}$ is a matrix of size $N \times Q$ and $\widetilde{\mathbf{g}}_{m;q}^{PI_b} = (\widetilde{\mathbf{y}}_{m;q}^b - \mathbf{y}_{m;q})$ is the vector of the specific direction for the simple indicator q at iteration b.

Once the b vectors of the directions have been found, the b benchmark selections, relative to the specific potential improvements of the simple indicators, are obtained with a further maximization problem:

$$\widetilde{D}_m^b(\mathbb{1}, \mathbf{y}; \widetilde{\mathbf{g}}_m^{PI_b}) = \sup\left\{\beta | \left(\mathbb{1}, \mathbf{y} + \beta \widetilde{\mathbf{g}}_m^{PI_b}\right) \in \widetilde{\Psi}_m\right\}, \forall b = 1, \ldots, B \tag{7}$$

where $\widetilde{D}_m^b(\mathbb{1}, \mathbf{y}; \widetilde{\mathbf{g}}_m^{PI_b})$ is a vector of size N and $\beta \in [0, 1]$ measures the proportion by which each of the simple indicators must be increased in order to reach the frontier.

Then, robust directions and selected benchmarks are given by the expected value of the related bootstrap distributions, *i.e.*:

$$\mathbf{g}^{PI} = E\left[\widetilde{\mathbf{g}}_m^{PI_1}, \ldots, \widetilde{\mathbf{g}}_m^{PI_b}, \ldots, \widetilde{\mathbf{g}}_m^{PI_B}\right] \tag{8}$$

where $\mathbf{g}^{PI} = (\mathbf{g}_1^{PI}, \ldots, \mathbf{g}_q^{PI}, \ldots, \mathbf{g}_Q^{PI})$ is a matrix of size $N \times Q$.

$$D(\mathbb{1},\mathbf{y};\widetilde{\mathbf{g}}_m^{PI}) = E\left[\widetilde{D}_m^b\left(\mathbb{1},\mathbf{y};\widetilde{\mathbf{g}}_m^{PI_1}\right),\ldots,\widetilde{D}_m^b\left(\mathbb{1},\mathbf{y};\widetilde{\mathbf{g}}_m^{PI_b}\right),\ldots,\widetilde{D}_m^b\left(\mathbb{1},\mathbf{y};\widetilde{\mathbf{g}}_m^{PI_B}\right)\right] \tag{9}$$

where $D(\mathbb{1},\mathbf{y};\widetilde{\mathbf{g}}_m^{PI})$ is a matrix of size $N \times Q$.

Expected values are approximated, as usual, with empirical means over B. Note that in this case, unlike RBoD or RDBoD, which handle vectors, the computation involves the mean over B of the columns of matrix blocks of size $N \times Q$.

The relative robust multi-directional scores vector for the simple indicator q is then calculated as the following:

$$e_q = \frac{\mathbf{y}_q}{\mathbf{y}_q + \beta^* \mathbf{g}_q^{PI}} \tag{10}$$

where $\mathbf{g}_q^{PI}$ is the q-th element of $\mathbf{g}^{PI}$, *i.e.*, the vector of units specific directions of the simple indicator q.

The overall CI_{RMDir_BoD} scores, to be consistent with Fusco (2023), are determined as the difference to 1 of the *normalized potential improvements inefficiency index*, proposed by Bogetoft and Hougaard (1999), related to the benchmark:

$$CI_{MDir\_RBoD} = 1 - \frac{\beta^* \sum_{q=1}^{Q} \mathbf{g}_q^{PI}}{\sum_{q=1}^{Q} \mathbf{y}_q + \beta^* \mathbf{g}_q^{PI}} \tag{11}$$

B-order resampling, finally, allows us to reconstruct the confidence interval of the estimated CI values. According to the t-distribution, given that the population standard deviation is unknown, the CI confidence interval is equal to:

$$\bar{x} \pm t \cdot (s/\sqrt{B}) \tag{12}$$

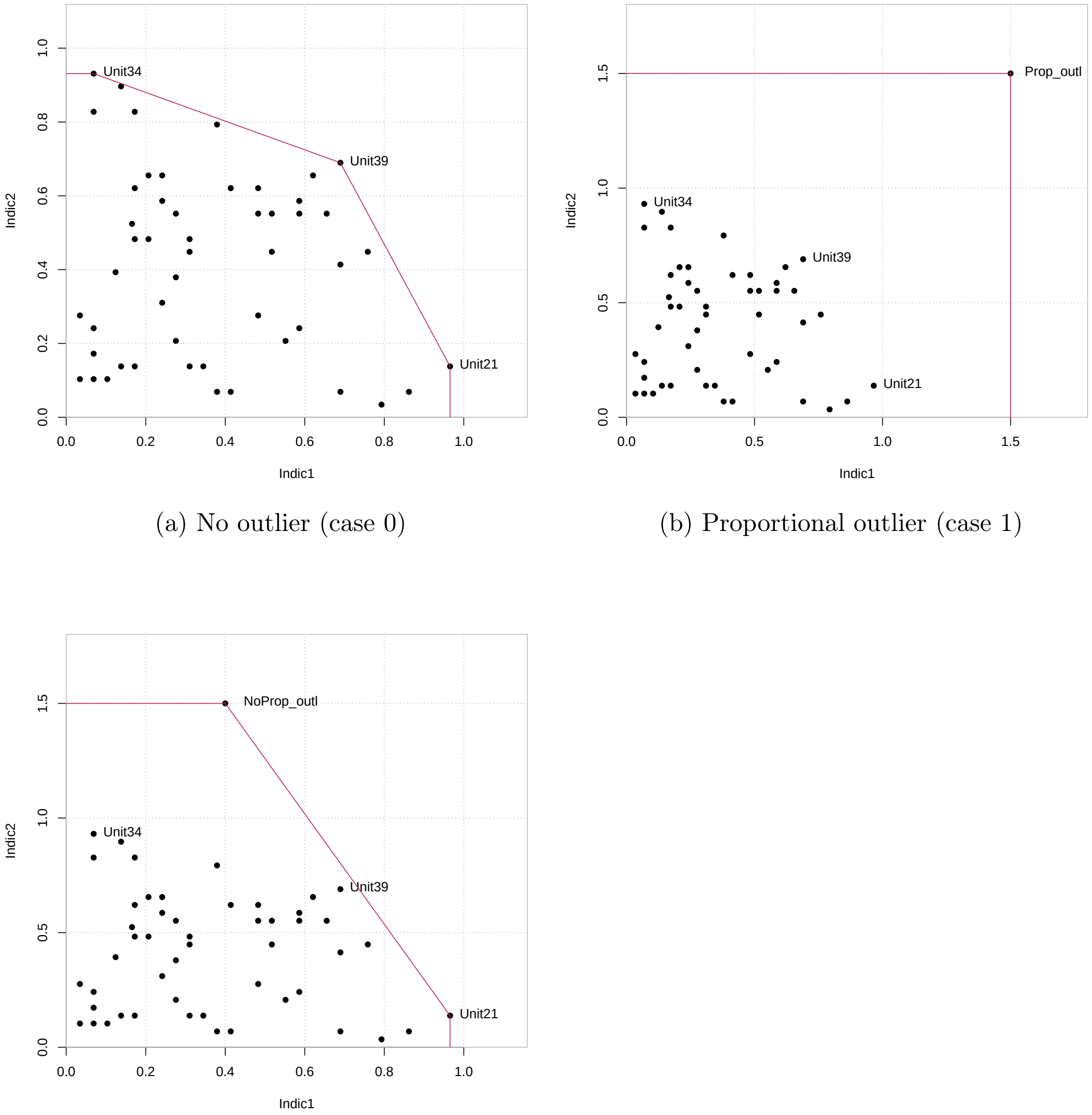
where $\bar{x}$ is the mean of the sample data, t is the critical t-value from the t-distribution depending on the desired confidence level and degrees of freedom ($df = n - 1$, where n is the sample size), s is the standard deviation of the sample data and B is the sample size equal to the number of iterations. It is important to emphasize, again, that `RMDir_BoD` as `MDir_BoD` provides a weighting scheme that *"is not only the most favourable for each unit, but is the «most favorable in the desirable direction» looking for potential improvements instead of following past production (like in BoD) or a specific direction (like in DBoD)."* (Fusco, 2023) and that it adds a further enrichment, *i.e.*, the comparison with a frontier robust to outliers and abnormal units.

4. Simulations

In order to better highlight the additional property of the proposed method, i.e., the robustness to outliers, a descriptive example has been implemented on simulated data.

In the baseline setting (Figure 1.a), two simple indicators (`Indic1` and `Indic2`) for 50 units have been extracted from a uniform distribution, ranging from 0 to 1, with concavity constraint; three efficient units (Unit34, Unit39 and Unit21) define the frontier, which is the benchmark for the other units.

To show the effect of outliers, both on the composite indicator and on the improvement directions, three cases have been proposed: the first one is the baseline, where there are no outliers or out-of-scale units (Figure 1.a, case 0); in the second one - maintaining simulated non-abnormal data - an outlier with both out-of-scale indicators (called *proportional outlier*, Figure 1.b, case 1) has been added causing a proportional frontier shift; in the third simulation, the outlier unit presenting an anomalous value on a single indicator (called *non-proportional outlier*, Figure 1.c, case 2) has been added causing only a partial shift of the frontier (please note that only Unit21 remains on the frontier).



(a) No outlier (case 0)

(b) Proportional outlier (case 1)

(c) Non-proportional outlier (case 2)

Figure 1: Simulated cases

Under these assumptions, the proposed method (`MDir_RBoD`) has been compared with the non-robust one (`MDir_BoD`), to test its ability to be unaffected by the presence of outlier units. Operationally, we will then add the outlier data, calculate CI scores and directions for improvement for all units, and then analyze the results on the 50 non-outlier units only. We, therefore, expect that the scores obtained on the 50 non-outlier units with the `MDir_BoD` method in the baseline setting (case 0, namely the comparison term in terms of both CI and estimated directions) will be very different in the proportional and non-proportional cases, while both CI scores and directions will not be different when using the `MDir_RBoD` method.

Table 1 shows the average differences (in absolute terms) between the CI scores calculated with the `MDir_BoD` method and its robust version `MDir_RBoD`, while Table 2 reports the average of the absolute value differences between the directions for both the first (`Indic1`) and the second (`Indic2`) indicator. Some findings emerge:

- In case 0 (first column of Table 1 and 2), the mean differences are minimal with the `MDir_RBoD` method, *i.e.* the robust version of the multidirectional BoD method has no impact on both the composite indicators (0.066) and especially the directions (0.134 and 0.103), which remain substantially similar. This result is true for the directions both on average and for each unit (Figure A.1) showing substantial stability in the directions of improvement for each unit.

- In case 1 (second column of Table 1 and 2), the differences between the methods are evident both in terms of CI scores and in terms of directions: if the `MDir_BoD method`, as it is obvious, is severely affected in terms of CI score compared to the robust version (0.038 vs. 0.050), even in terms of mean changes in directions (0.709 and 0.704 vs. 0.061 and 0.172), the improvement induced by the `MDir_RBoD method` stands out. Figure 2.a shows this result even more clearly by highlighting, for each unit and indicator, the biasing effect of the single outlier in the directions calculated by the `MDir_BoD model` as opposed to the robust version (Figure 2.b).

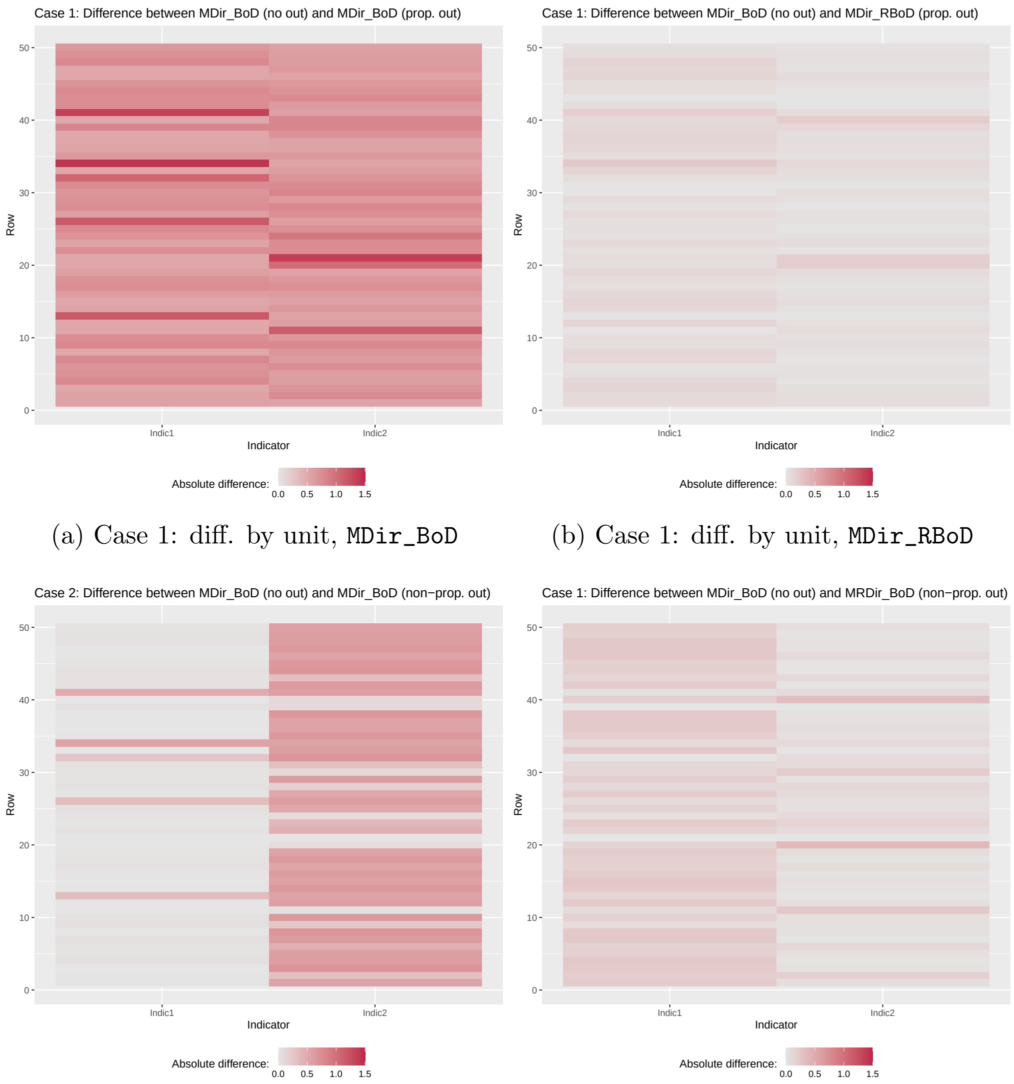
- The non-proportional outlier case (case 2, the third column of Table 1 and 2) highlights slightly different results for the `MDir_BoD method`: on the one hand a limited distortion in terms of the CI score, but regarding the directions an interesting aspect, namely that the directions are distorted in terms of the outlier indicator while safeguarding the other dimension (see also Figure 2.c). The robust method again allows the biasing effect of the outlier to be controlled.

	No outlier (case 0)	Proport. outlier (case 1)	Non-Proport. outlier (case 2)
MDir_BoD		0.348	0.099
MDir_RBoD	0.066	0.050	0.074

Table 1: Average differences (in absolute term) between the MDir_BoD CIs calculated in no outliers case and the other settings and method

	No outlier (case 0)		Proport. out (case 1)		Non-Proport. out (case 2)	
	Indic1	Indic2	Indic1	Indic2	Indic1	Indic2
MDir_BoD			0.709	0.704	0.058	0.494
MDir_RBoD	0.134	0.103	0.093	0.061	0.172	0.076

Table 2: Average differences (in absolute term) between the MDir_BoD directions calculated in no outliers case and the other settings and method



Case 1: Difference between MDir_BoD (no out) and MDir_BoD (prop. out)

Case 1: Difference between MDir_BoD (no out) and MDir_RBoD (prop. out)

(a) Case 1: diff. by unit, `MDir_BoD`

(b) Case 1: diff. by unit, `MDir_RBoD`

Case 2: Difference between MDir_BoD (no out) and MDir_BoD (non−prop. out)

Case 1: Difference between MDir_BoD (no out) and MRDir_BoD (non−prop. out)

(c) Case 2: diff. by unit, `MDir_BoD`

(d) Case 2: diff. by unit, `MDir_RBoD`

Figure 2: Direction differences by unit between MDir_BoD (case 0) and other case/method

The generalization of this particular illustrative setting is not straightforward because clearly linked with our specific simulation/hypothesis on the input data. But still, within our illustrative setting, we can ask what would be the effect of a higher outlier (*e.g.* a non-proportional outlier) in terms of mean differences from the baseline case; in other terms, if the outlier had been a multiple (in the example 1 to 4) of the outlier shown in Figure 1.c, what would be the results in terms of the MDir_BoD and MDir_RBoD computations?

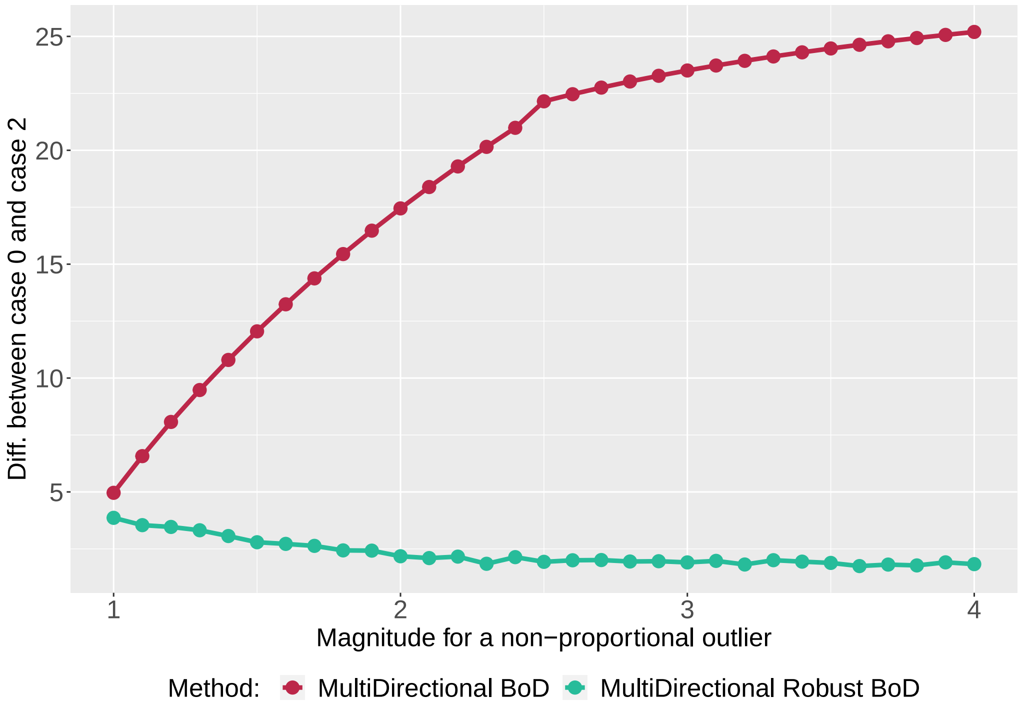


Figure 3: Average differences (in absolute term) between the MDir_BoD CIs calculated in case 0 and case 2 by the magnitude of the outlier unit and by method

Figure 3 shows that as the magnitude of the outlier unit increases, the differences from the baseline case in terms of CI score increase markedly in the non-robust case, whereas they are optimally controlled for by our robust method.

5. Health care quality in OECD countries

In this section, we apply the methodology developed in section 3 to the quality assessment of health care carried out by OECD[8]. In 2001, the OECD

[8] https://www.oecd.org/health/health-care-quality-outcomes-indicators.htm

started the Healthcare Quality Indicators project with the aim of collecting data for the development and reporting of health care quality indicators and for international comparisons. Since then, the initiative has evolved with the incorporation of quality measures into a conceptual framework for health system performance (Arah et al., 2006). The current Healthcare Quality and Outcomes program (HCQO) includes a total of 64 indicators and covers 40 countries[9]. The data are regularly published on a dedicated page on the OECD website and in the organization's well-known publications, such as Health at a Glance. To the best of our knowledge, OECD has not attempted to provide a measure of the overall quality of health care in different countries, and cross-country comparisons can be made only for each individual indicator.

The indicators are grouped into the following areas: acute care; cancer care; mental health care; end of life care; integrated care; mental health - patient reported experience measures; patient experience; patient safety; primary care prescribing; primary care. For our exercise, we decided to focus on the acute care indicators: AMI (acute myocardial infarction) 30-day mortality; hemorrhagic stroke 30-day mortality; ischaemic stroke 30-day mortality; hip fracture surgery started within 2 days of admission to hospital. We are aware that this choice does not exploit the full potential of our methodology to provide information on the overall performance of the health systems studied, as it excludes very important areas of care. However, our choice is motivated by the fact that this simplification makes it easier to present the value-added information resulting from the use of our methodology.

In this paper, we use data related to the four indicators above, for the last year available[10], with the aim of computing a composite indicator of the quality of care at national level[11], by applying the `MDir_RBoD` developed in section 3.

[9] For an analysis of the evolution of the OECD initiative, please see Carinci et al. (2015).

[10] Collection date: 04 May 2023, source: https://stats.oecd.org/Index.aspx?QueryId=51879. Data are not available for a specific year for all countries considered. France, Greece, Poland and the USA had at least one indicator with missing data for all years considered; they were, therefore, removed from the analysis set because they were not fully comparable with the other countries.

[11] National data have been originally standardized by age, sex, co-morbidity. Ratios are based on linked data focused on each patient (a single patient is counted only once) and use them as a denominator (regardless of the number of admissions or readmissions). Ratios consider deaths occurring anywhere, including inside or outside the hospital of admission.

All the simple indicators have been normalized in the range [0,1] and the sign of negative polarities has been reversed, using the min–max method. Figures A.2, A.3, A.4 and A.5 report the normalized values of the individual indicators, with increasing polarity (the higher the indicator, the better the country performs).

The observation of these elementary data does not highlight the existence of any evident outlier, except for one country (Iceland), characterized by a smaller and younger population than the other countries, which may lead to problems of comparability. Our robust approach, however, allows to control for this aspect.

From such data, therefore, BoD and MDir_RBoD have been calculated, and the relative scores are reported in Tables A.1 and A.2, while Table 3 reports the descriptive statistics of the results. Please note that the proposed method also allows the calculation of confidence intervals for the robust composite indicator by resampling, thus providing an indirect estimate of the reliability of the indicator itself.

Statistic	N	Mean	St. Dev.	Min	Max
BOD	24	0.822	0.163	0.348	1.000
MDir_RBoD	24	0.720	0.207	0.243	1.000

Spearman ranking correlation: 0.904

Table 3: BoD and MDir_RBoD scores - Descriptive statistics

The first obvious thing to be noted is the provision of information on the "big picture" (Jacobs and Goddard, 2007) of the quality performance of the different countries in the provision of acute care services, with both BoD and MDir_RBoD measures. These measures, for each country, are expressed relatively to the score of the best performers (for both composite indicators, Iceland). This information is particularly valuable for those countries, which do not score either well or bad for all the four indicators, in which case the information arising from the single indicators would be sufficient for understanding the relative overall performance of a country. The composite scores, therefore, while confirming a good placement for the European Nordic countries, or for countries like Switzerland, New Zealand and Canada, also allows for appreciating the overall performance of other countries, like Turkey and Romania, which behave well on some indicators, and less well on others.

Even if there is a good degree of correlation between the values of both BoD and MDir_RBoD and the ones for each single indicator, there is no single indicator that can replicate the refined information provided by the composite measures.

The second remark is related to the comparison between MDir_RBoD and BoD scores. It allows to stress the difference made by assuming (or not) the perfect compensability of the performances under the single indicators. As it can be easily observed, from the descriptive statistics as well as from the scores for each single country, MDir_RBoD scores are generally lower than for BoD. The reason is that the countries with an unbalanced performance between the four indicators are penalized by MDir_RBoD, and the stronger the unbalance the higher the penalisation. We can notice, therefore, that there is one country, Hungary, that suffers the most from a strongly unbalanced provision of quality in the four acute services considered by the single indicators, but also several other countries (among them, Canada, Singapore, UK, Finland, Spain, Portugal) show an uneven provision of quality within acute care. If it is not accepted, for the reasons discussed in section 2, that the underperformance in one area of care can be compensated by overperformance in another area, our indicator incorporates this "value judgment" in the assessment of the overall quality of care provided by a country.

Figure 4 illustrates the difference between the two methods more clearly, using the case of Hungary as an example and, without loss of generality, considering only two indicators. On the axes, we measure the normalized values of the two indicators, for each country.

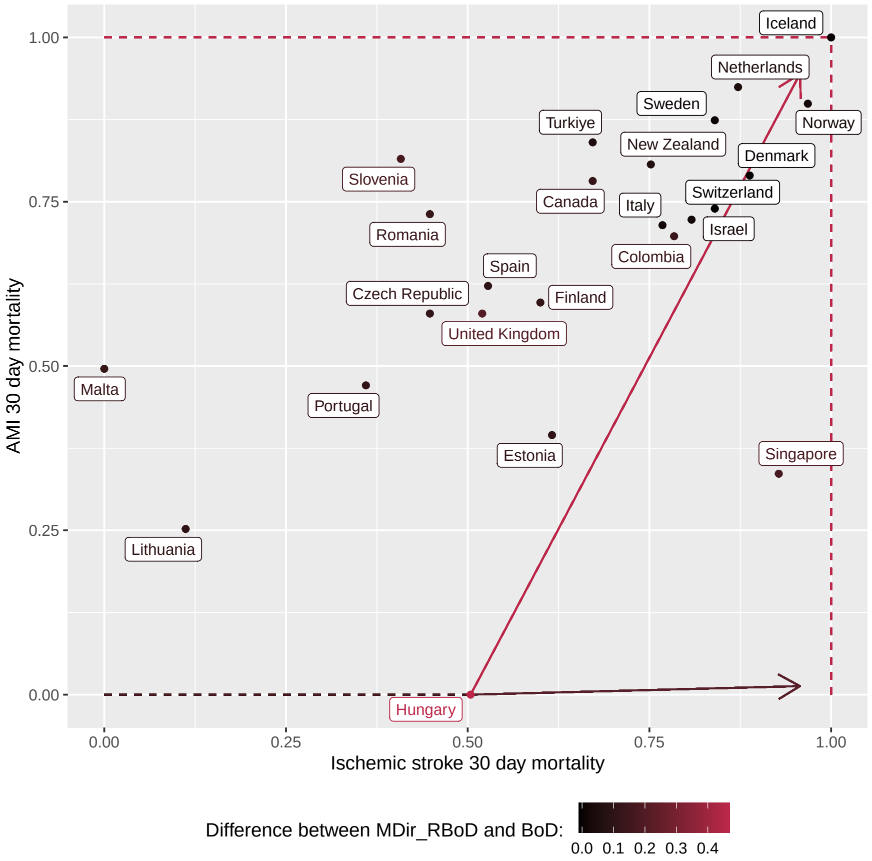


Figure 4: Estimated `MDir_RBoD` (red) vs `BoD` (black) radial direction

Hungary shows the lowest value of the AMI indicator (by construction, its normalized and polarized value is zero), while the value of the ischaemic stroke indicator is significantly higher. In this case, the `BoD` method calculates the composite indicator as a radial distance (black line) from the deterministic frontier (red dashed line), since it can achieve the maximum overall quality by increasing the value of the ischaemic indicator only, thus compensating the very bad underperformance on AMI with a good performance on the other indicator (and more generally following the path of the actual mix of services). The `MDir_RBoD` approach shows that the composite indicator, understood as the distance from a benchmark characterized by an even performance on the two indicators, is greater (red line) and that an improvement of the AMI indicator is also necessary since it cannot be compensated by the better position of Hungary on the ischemic stroke indicator. Finally, note that the red arrow (the distance from the benchmark) does not touch the Iceland point, but is slightly smaller, having dampened the extreme data of the only border benchmark (due to the robustness of our indicator).

The key information provided by the `MDir_RBoD` method is not only the overall quality performance of each country, as measured by the `MDir_RBoD` score, but also the improvement path of each country, in terms of the potential improvement of the different indicators, which is necessary to move to the (balanced) benchmark on the frontier. The "direction" scores, in Table A.1, measure what is the fraction of "optimal" performance that each country lacks, for each indicator. In other words, it measures the relative effort that each country has to make for each indicator, so as to move its overall performance all along the path that takes to the frontier. We can graphically represent this information through a typical radar plot, like the one in Figure 5.

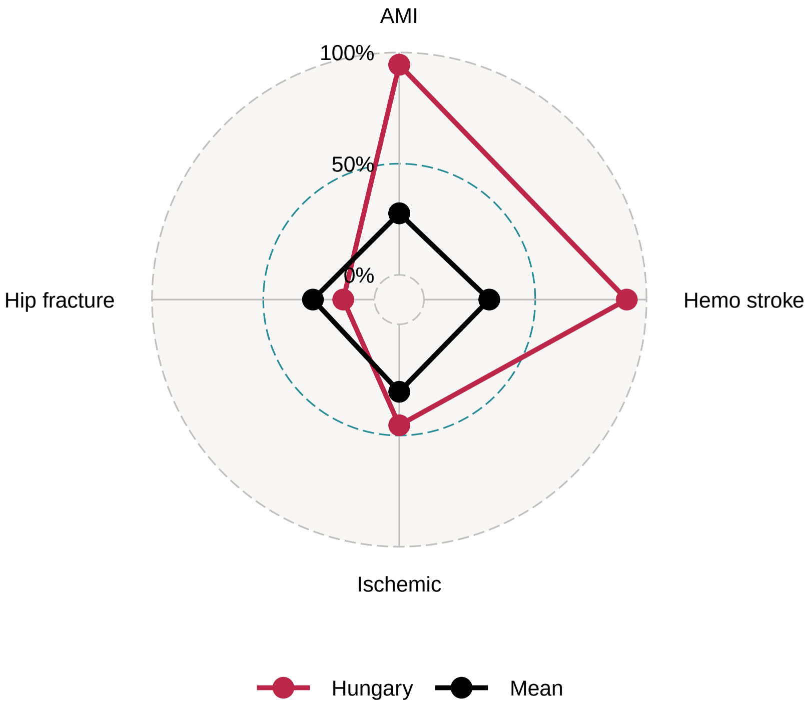


Figure 5: Estimated directions for improvement - Hungary vs directions mean

Using again Hungary as an example, we can see that most of the work that this country has to do to improve its relative low quality of acute care services (measured by a value of `MDir_RBoD` equal to 0.3763) must be made in the areas of AMI and Hemorrhagic stroke mortality. By improving its performance in these two areas, as well as in the other two services, even if to a less extent, Hungary may move to the frontier, improving the overall quality of its acute services in a balanced way.

The area of the radar plot represents the overall effort required to improve the overall quality and is, therefore, proportional to the distance from the robust frontier (Figure 6), that is it is inversely related to the value of the `MDir_RBoD` score.

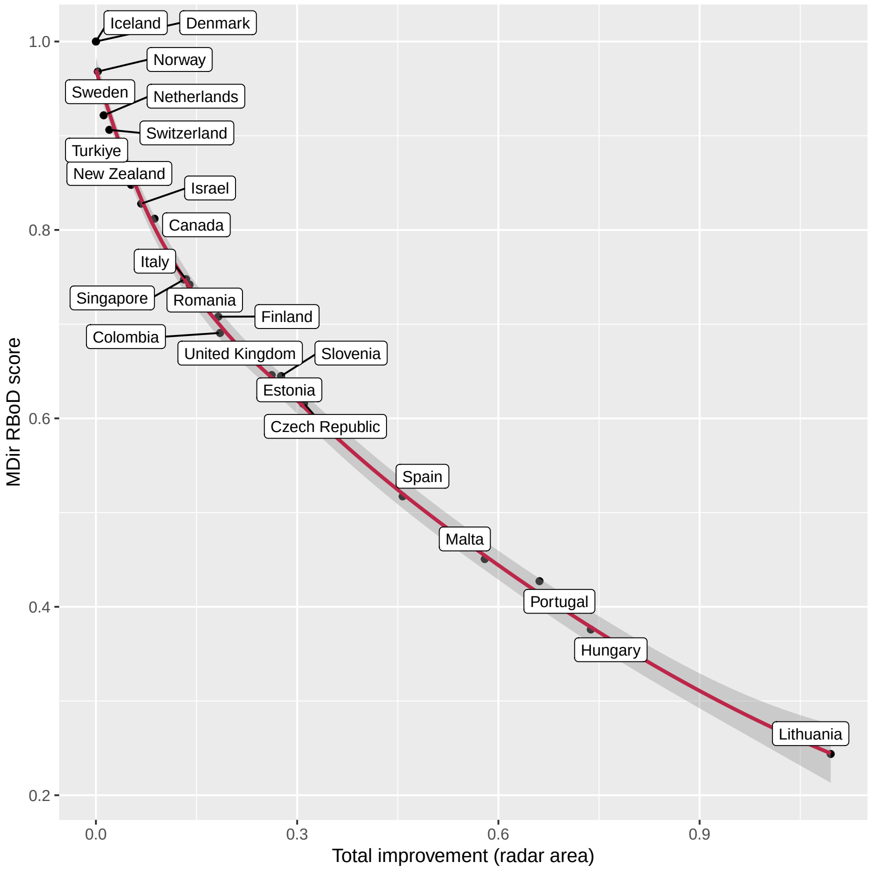


Figure 6: `MDir_RBoD` score and the needed total improvement

6. Concluding remarks

The paper aims to contribute to the literature on composite indicators as well as to their use to measure the quality of health care. While individual indicators in health care quality assessment provide detailed information on specific aspects, in fact, they cannot capture all relevant information. Therefore, there has long been a need to construct composite measures to assess the overall quality or performance of health care, not yet covered by official OECD measurements.

The study departs from the prevailing methodologies by proposing an original multi-directional robust Benefit of the Doubt approach that allows highlighting the improvement directions of individual units within a robust framework. First, an approach based on simulated data has been carried out to better describe the advantages of the proposed approach and then the methodology has been applied to country-level health data highlighting that potential health losses in some areas of care are not necessarily offset by health gains in others.

It is worth noting that the composite score computed in this work, given its features, in particular its robustness to outliers and its non compensability nature, is particularly reliable for its potential use in other analyses. For instance, it could be used for comparing the overall quality performance of each country along time. Surely, it is more significant than single measures in statistical analyses of how policies or other contextual factors (like competition, for instance) impact on the overall quality/performance of health care systems, or for the assessment of the efficient use of their resources. It must also be stressed that the methodology, while applied here to country-level data, can also be employed to evaluate the quality of care of single providers, in such a way to get relevant information for managerial actions.

From a methodological point of view, future enhancements to the proposed robust and multidirectional approach could certainly concern two aspects: *(i)* the integration of conditional approaches (Rogge et al., 2017) that allow composite indicators to be calculated with the same contextual factors, and *(ii)* the development of this approach in a hierarchical framework (Shen et al., 2013) that allow to better evaluate the multidimensional phenomenon in a multi-level setting.

References

Arah, O., Westert, G., Hurst, J., Klazinga, N., 2006. A conceptual framework for the oecd health care quality indicators project. International journal for quality in health care 18 Suppl 1, 5–13.

Barclay, M., Dixon-Woods, M., Lyratzopoulos, G., 2019. The problem with composite indicators. BMJ Quality & Safety 28, 338–344.

Beaussier, A., Demeritt, D., Griffiths, A., Rothstein, H., 2020. Steering by their own lights: Why regulators across europe use different indicators to measure healthcare quality. Health Policy 124, 501 – 510.

Bogetoft, P., Hougaard, J.L., 1999. Efficiency evaluations based on potential (non-proportional) improvements. Journal of Productivity Analysis 12, 233–247.

Carinci, F., Van Gool, K., Mainz, J., Veillard, J., Pichora, E., Januel, J.M., Arispe, I., Kim, S., Klazinga, N., 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.

Cherchye, L., Lovell, K., Moesen, W., Puyenbroeck, T.V., 2005. One market, one number? a composite indicator assessment of eu internal market dynamics. Technical Report. Working Paper Series ces0513, Katholieke Universiteit Leuven, Centrum voor Economische Studien.

Cherchye, L., Moesen, W., Rogge, N., Puyenbroeck, T., 2007. An Introduction to 'Benefit of the Doubt' Composite Indicators. Social Indicators Research 82, 111–145.

D'Inverno, G., De Witte, K., 2020. Service level provision in municipalities: A flexible directional distance composite indicator. European Journal of Operational Research 286, 1129–1141.

Donabedian, A., 1966. Evaluating the quality of medical care. The Milbank Memorial Fund Quarterly 44, 166–206.

Fare, R., Karagiannis, G., Hasannasab, M., Margaritis, D., 2019. A benefit-of-the-doubt model with reverse indicators. European Journal of Operational Research 278, 394–400.

Friebel, R., Steventon, A., 2019. Composite measures of healthcare quality: sensible in theory, problematic in practice. BMJ Quality & Safety 28, 85–88.

Fusco, E., 2015. Enhancing non-compensatory composite indicators: A directional proposal. European Journal of Operational Research 242, 620 – 630.

Fusco, E., 2023. Potential improvements approach in composite indicators construction: The multi-directional benefit of the doubt model. Socio-Economic Planning Sciences 85, 101447.

Fusco, E., Vidoli, F., Rogge, N., 2020. Spatial directional robust benefit of the doubt approach in presence of undesirable output: An application to italian waste sector. Omega 94, 102053.

Hofstede, S.N., Ceyisakar, I.E., Lingsma, H.F., Kringos, D.S., van de Mheen, P.J.M., 2019. Ranking hospitals: do we gain reliability by using composite rather than individual indicators? BMJ Quality & Safety 28, 94–102.

Jacobs, R., Goddard, M., 2007. How do performance indicators add up? an examination of composite indicators in public services. Public Money & Management 27, 103–110.

Jacobs, R., Goddard, M., Smith, P., 2005. How robust are hospital ranks based on composite performance measures? Medical Care 43, 1177–84.

Jacobs, R., Martin, S., Goddard, M., Gravelle, H., Smith, P., 2006. Exploring the determinants of nhs performance ratings: lessons for performance assessment systems. Journal of Health Services Research & Policy 11, 211–217.

Jacobs, R., Smith, P., Goddard, M., 2004. Measuring performance: An examination of composite performance indicators. Center for Health Economics, University of York, CHE Technical Paper Series, n.29.

Kara, P., Valentin, J.B., Mainz, J., Johnsen, S.P., 2022. Composite measures of quality of health care: Evidence mapping of methodology and reporting. PLOS ONE 17, 1–21.

Karagiannis, G., 2017. On aggregate composite indicators. Journal of the Operational Research Society 68, 741–746.

Mahdiloo, M., Andargoli, A.E., Toloo, M., Harvie, C., Duong, T.T., 2023. Measuring the digital divide: A modified benefit-of-the-doubt approach. Knowledge-Based Systems 261.

Matos, R., Ferreira, D., Pedro, M.I., 2021. Economic analysis of portuguese public hospitals through the construction of quality, efficiency, access, and financial related composite indicators. Social Indicators Research 157, 361–392.

Mergoni, A., D'Inverno, G., Carosi, L., 2022. A composite indicator for measuring the environmental performance of water, wastewater, and solid waste utilities. Utilities Policy 74, 101285.

Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., Giovannini, E., 2008. Handbook on constructing composite indicators: Methodology and user guide. OECD Publishing .

Oliveira, R., Zanella, A., Camanho, A.S., 2020. A temporal progressive analysis of the social performance of mining firms based on a malmquist index estimated with a benefit -of -the-doubt directional model. Journal of Cleaner Production 267.

Oliver, A., 2012. The folly of cross-country ranking exercises. Health Economics, Policy, and Law 7, 15–7.

Pereira, M.A., Camanho, A.S., Figueira, J.R., Marques, R.C., 2021a. Incorporating preference information in a range directional composite indicator: The case of portuguese public hospitals. European Journal of Operational Research 294, 633–650.

Pereira, M.A., Camanho, A.S., Marques, R.C., Figueira, J.R., 2021b. The convergence of the world health organization member states regarding the united nations' sustainable development goal 'good health and well-being'. Omega 104, 102495.

Pereira, M.A., Marques, R.C., 2022. The 'sustainable public health index': What if public health and sustainable development are compatible? World Development 149, 105708.

Rogge, N., 2018a. Composite indicators as generalized benefit-of-the-doubt weighted averages. European Journal Of Operational Research 267, 381–392.

Rogge, N., 2018b. On aggregating benefit of the doubt composite indicators. European Journal Of Operational Research 264, 364–369.

Rogge, N., De Jaeger, S., Lavigne, C., 2017. Waste performance of nuts 2-regions in the eu: A conditional directional distance benefit-of-the-doubt model. Ecological Economics 139, 19–32.

Sahoo, B.K., Singh, R., Mishra, B., Sankaran, K., 2017. Research productivity in management schools of india during 1968-2015: A directional benefit-of-doubt model analysis. Omega 66, 118–139.

Shen, Y., Hermans, E., Brijs, T., Wets, G., 2013. Data envelopment analysis for composite indicators: A multiple layer model. Social Indicators Research 114, 739–756.

Shwartz, M., Restuccia, J., Rosen, A., 2015. Composite measures of health care provider performance: A description of approaches. The Milbank Quarterly 93, 788–825.

Smith, P., 2002. Measuring Up: Improving Health System Performance in OECD Countries. OECD Publishing. chapter Developing Composite Indicators for Assessing Health System Efficiency. pp. 295 – 316.

Street, A., Smith, P., 2021. How can we make valid and useful comparisons of different health care systems? Health Services Research 56, 1299–1301.

Van Puyenbroeck, T., Rogge, N., 2017. Geometric mean quantity index numbers with benefit-of-the-doubt weights. European Journal of Operational Research 256, 1004–1014.

Verbunt, P., Rogge, N., 2018. Geometric composite indicators with compromise benefit-of-the-doubt weights. European Journal of Operational Research 264, 388–401.

Vidoli, F., Fusco, E., Mazziotta, C., 2015. Non-compensability in composite indicators: A robust directional frontier method. Social Indicators Research 122, 635–652.

Vidoli, F., Mazziotta, C., 2013. Robust weighted composite indicators by means of frontier methods with an application to european infrastructure endowment. Italian Journal of Applied Statistics 23, 259–282.

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.

Appendix A. Appendix

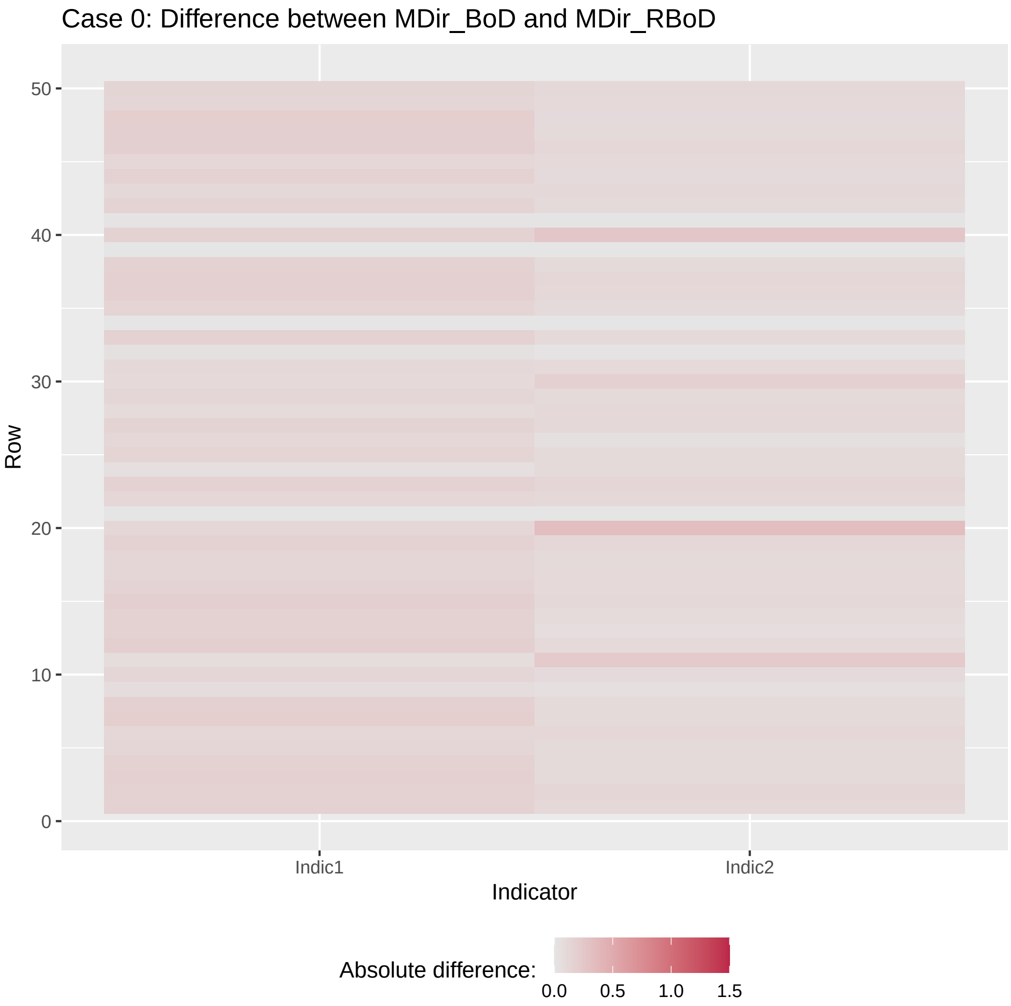


Figure A.1: Direction differences by unit between `MDir_BoD` (case 0) `MDir_RBoD`

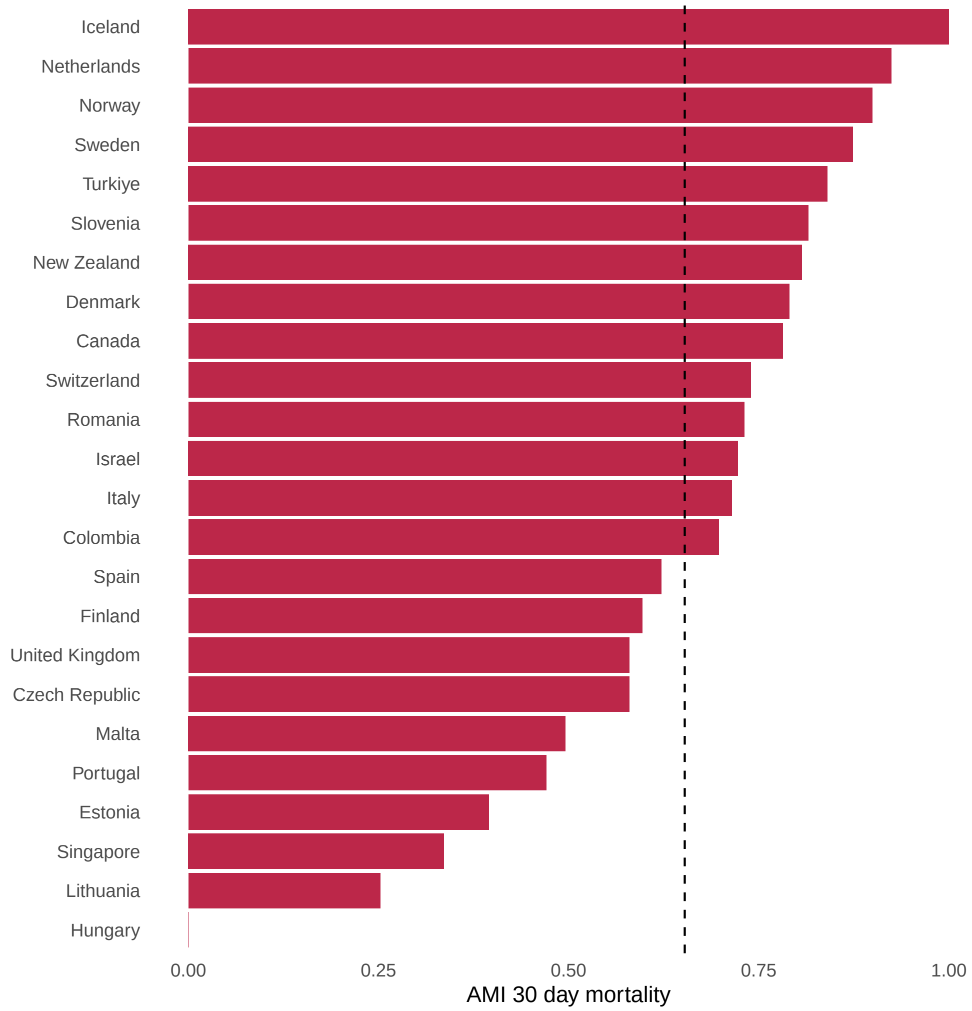


Figure A.2: AMI 30 day mortality, normalized and polarised indicator

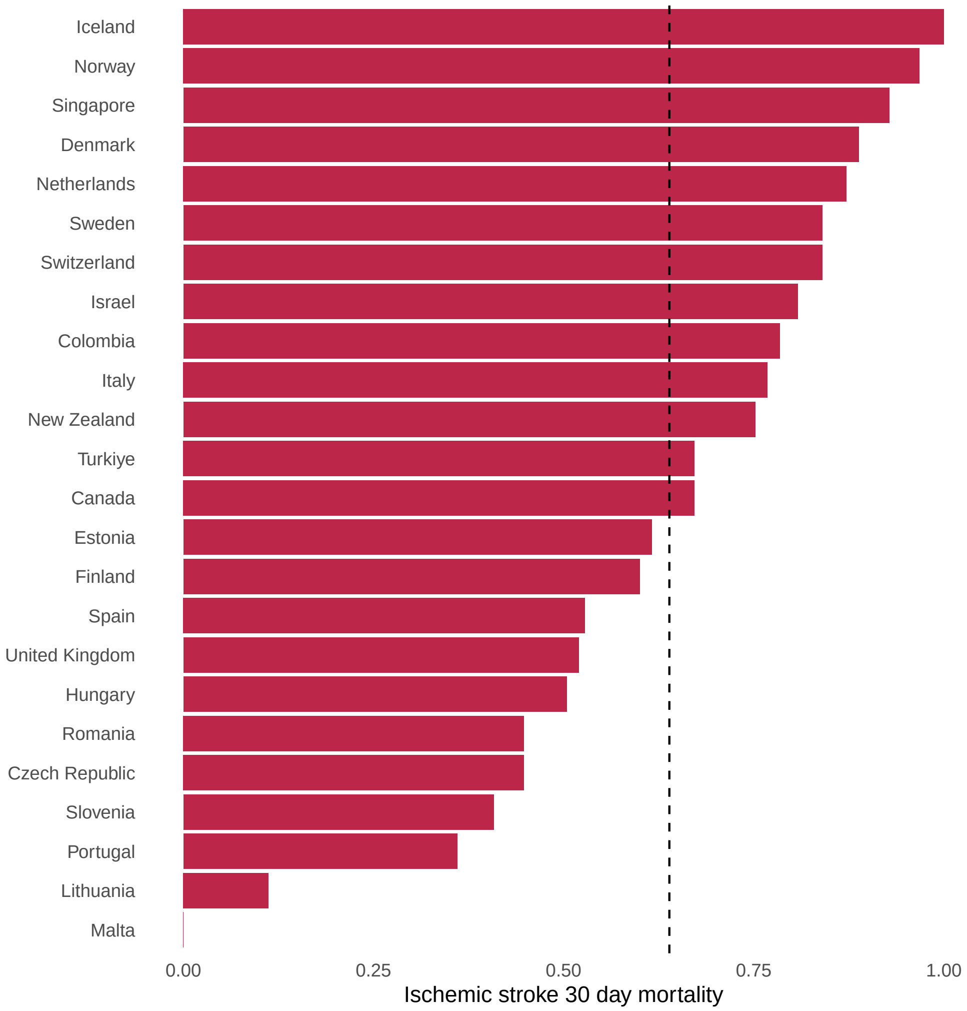


Figure A.3: Ischemic stroke 30 day mortality, normalised and polarised indicator

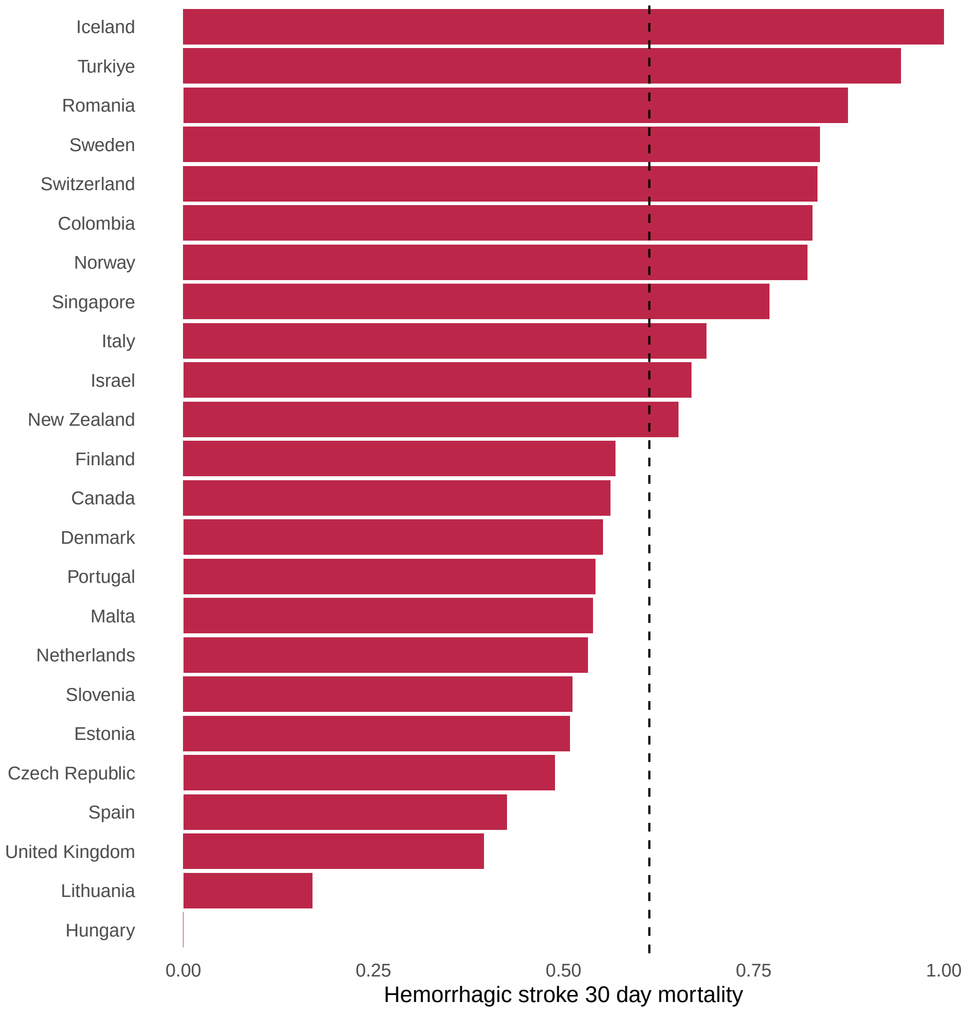


Figure A.4: Hemorrhagic stroke 30 day mortality, normalised and polarised indicator

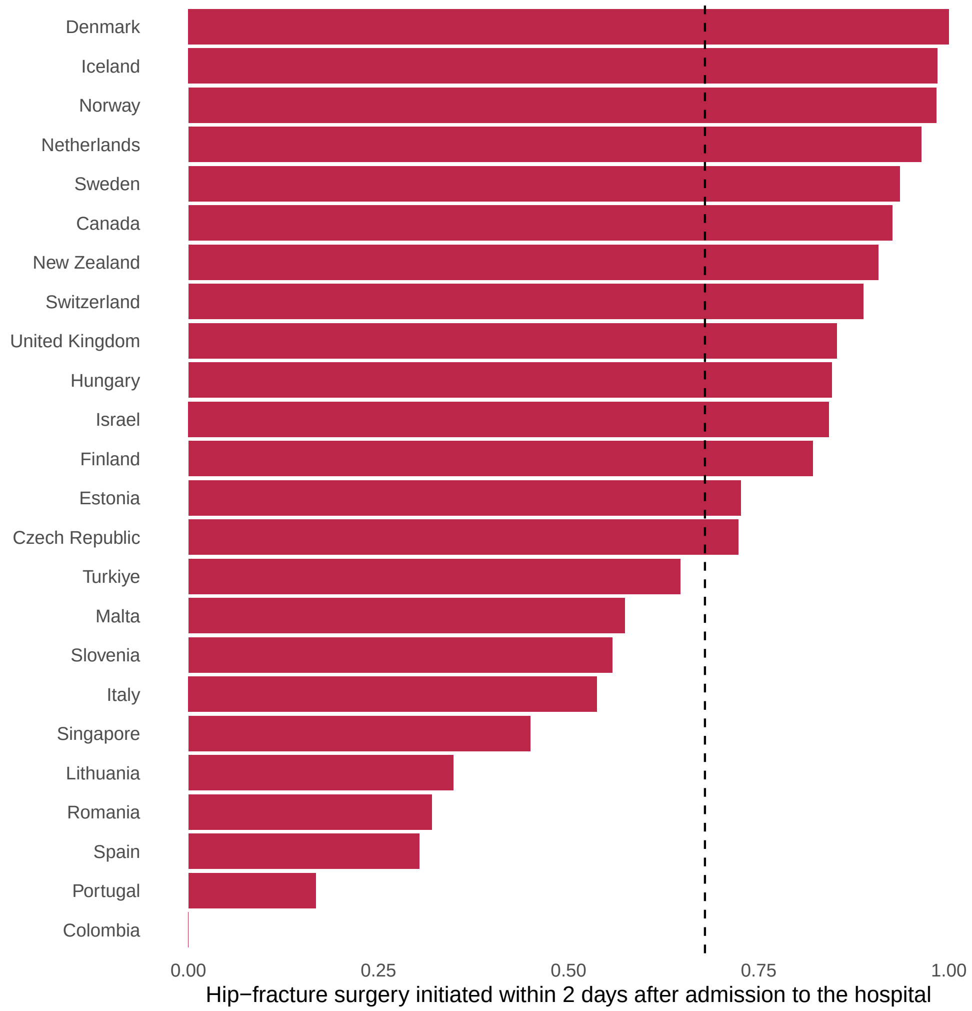


Figure A.5: Hip-fracture surgery initiated within 2 days after admission to the hospital

Country	AMI	Hemo.	Ischemic	Hips	BoD		MDir_RBoD		
					Score	Rank	Score	95% CI	Rank
Denmark	4.50	23.90	4.80	97.60	1.0000	(1)	1.0000	(1.0000, 1.0000)	(1)
Iceland	2.00	10.40	3.40	96.70	1.0000	(1)	1.0000	(1.0000, 1.0000)	(1)
Norway	3.20	15.80	3.80	96.60	0.9947	(3)	0.9686	(0.9673, 0.9699)	(3)
Netherlands	2.90	24.50	5.00	95.40	0.9745	(4)	0.9202	(0.9170, 0.9234)	(5)
Sweden	3.50	15.30	5.40	93.70	0.9458	(5)	0.9405	(0.9381, 0.9428)	(4)
Turkiye	3.90	12.10	7.50	76.30	0.9435	(6)	0.8942	(0.8899, 0.8985)	(7)
Canada	4.60	23.60	7.50	93.10	0.9288	(7)	0.8100	(0.8073, 0.8127)	(10)
Singapore	9.90	17.30	4.30	64.40	0.9280	(8)	0.7433	(0.7372, 0.7495)	(12)
New Zealand	4.30	20.90	6.50	92.00	0.9132	(9)	0.8475	(0.8448, 0.8502)	(8)
Switzerland	5.10	15.40	5.40	90.80	0.8985	(10)	0.9068	(0.9039, 0.9097)	(6)
Romania	5.20	14.20	10.30	56.60	0.8738	(11)	0.7431	(0.7366, 0.7496)	(13)
United Kingdom	7.00	28.60	9.40	88.70	0.8524	(12)	0.6458	(0.6435, 0.6481)	(16)
Israel	5.30	20.40	5.80	88.10	0.8496	(13)	0.8267	(0.8240, 0.8294)	(9)
Hungary	13.90	40.50	9.60	88.30	0.8458	(14)	0.3763	(0.3748, 0.3778)	(23)
Colombia	5.60	15.60	6.10	37.30	0.8272	(15)	0.6927	(0.6879, 0.6976)	(15)
Finland	6.80	23.40	8.40	86.80	0.8247	(16)	0.7057	(0.7033, 0.7080)	(14)
Slovenia	4.20	25.10	10.80	70.90	0.8151	(17)	0.6441	(0.6407, 0.6474)	(17)
Italy	5.40	19.80	6.30	69.70	0.7680	(18)	0.7466	(0.7436, 0.7496)	(11)
Estonia	9.20	25.20	8.20	81.10	0.7299	(19)	0.6173	(0.6151, 0.6195)	(18)
Czech Republic	7.00	25.80	10.30	80.90	0.7260	(20)	0.6147	(0.6125, 0.6169)	(19)
Spain	6.50	27.70	9.30	55.60	0.6218	(21)	0.5178	(0.5158, 0.5197)	(20)
Malta	8.00	24.30	15.90	71.90	0.5810	(22)	0.4515	(0.4495, 0.4535)	(21)
Portugal	8.30	24.20	11.40	47.40	0.5415	(23)	0.4275	(0.4257, 0.4292)	(22)
Lithuania	10.90	35.40	14.50	58.30	0.3483	(24)	0.2434	(0.2426, 0.2443)	(24)

Table A.1: Simple indicators, BoD and MDir_RBoD CI with 95% confidence interval

Country	MDir_RBoD			Directions			
	Score	95% CI	Rank	AMI	Hemo.	Isch.	Hips
Denmark	1.0000	(1.0000, 1.0000)	(1)	0.000	0.000	0.000	0.000
Iceland	1.0000	(1.0000, 1.0000)	(1)	0.000	0.000	0.000	0.000
Norway	0.9686	(0.9673, 0.9699)	(3)	0.050	0.088	0.016	0.002
Sweden	0.9405	(0.9381, 0.9428)	(4)	0.060	0.075	0.085	0.024
Netherlands	0.9202	(0.9170, 0.9234)	(5)	0.037	0.228	0.062	0.012
Switzerland	0.9068	(0.9039, 0.9097)	(6)	0.171	0.084	0.090	0.066
Turkiye	0.8942	(0.8899, 0.8985)	(7)	0.078	0.027	0.159	0.165
New Zealand	0.8475	(0.8448, 0.8502)	(8)	0.134	0.249	0.196	0.070
Israel	0.8267	(0.8240, 0.8294)	(9)	0.215	0.248	0.147	0.132
Canada	0.8100	(0.8073, 0.8127)	(10)	0.163	0.331	0.282	0.061
Italy	0.7466	(0.7436, 0.7496)	(11)	0.224	0.239	0.187	0.432
Singapore	0.7433	(0.7372, 0.7495)	(12)	0.465	0.124	0.045	0.398
Romania	0.7431	(0.7366, 0.7496)	(13)	0.165	0.080	0.362	0.456
Finland	0.7057	(0.7033, 0.7080)	(14)	0.350	0.352	0.360	0.164
Colombia	0.6927	(0.6879, 0.6976)	(15)	0.219	0.097	0.152	0.897
United Kingdom	0.6458	(0.6435, 0.6481)	(16)	0.365	0.516	0.439	0.135
Slovenia	0.6441	(0.6407, 0.6474)	(17)	0.132	0.420	0.535	0.412
Estonia	0.6173	(0.6151, 0.6195)	(18)	0.553	0.423	0.345	0.261
Czech Republic	0.6147	(0.6125, 0.6169)	(19)	0.369	0.445	0.515	0.264
Spain	0.5178	(0.5158, 0.5197)	(20)	0.326	0.520	0.433	0.684
Malta	0.4515	(0.4495, 0.4535)	(21)	0.452	0.405	0.963	0.412
Portugal	0.4275	(0.4257, 0.4292)	(22)	0.476	0.405	0.604	0.819
Hungary	0.3763	(0.3748, 0.3778)	(23)	0.945	0.912	0.454	0.141
Lithuania	0.2434	(0.2426, 0.2443)	(24)	0.694	0.775	0.852	0.639

Table A.2: `MDir_RBoD` CI with 95% confidence interval, Directions for improvement