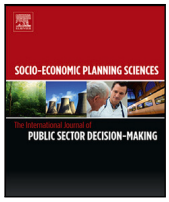
Socio-Economic Planning Sciences 93 (2024) 101877

Contents lists available at ScienceDirect

Socio-Economic Planning Sciences

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



Multi-directional Robust Benefit of the Doubt model: An application to the measurement of the quality of acute care services in OECD countries

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

[a] University of Urbino Carlo Bo, Department of Economics, Society and Politics, Via Aurelio Saffi, 42, Urbino, 61029, PU, Italy
[b] University of Florence, Department of Statistics, Computer Science, Applications "G. Parenti", Viale Morgagni, 59, Firenze, 50134, FI, 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

ARTICLE INFO

JEL classification:
C14
C43
C44
I18

Keywords:
Robust composite indicators
Non-compensatory
Multi-directional Benefit of the Doubt
Acute care quality

ABSTRACT

While individual metrics in evaluating healthcare quality offer in-depth insights into particular areas, they frequently fail to encompass all pertinent information. Consequently, there is a growing need to develop composite measures that comprehensively assess the overall quality or performance of specific care services, especially those not covered by official OECD measures. A novel multi-directional robust Benefit-of-the-doubt approach is proposed to measure overall acute care services quality through a composite indicator while, at the same time, highlighting the potential improvement directions for each single component indicator. 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 OECD data drawn from the Healthcare Quality and Outcomes programme.

1. Introduction

The concern of ensuring that substantial investments in healthcare systems in most developed nations are matched by improvements in efficiency and quality of care delivery has spurred efforts to measure these advancements. Regarding quality, specifically, even within the Donabedian [4] basic categorisation of quality into outcome, process, and structure of care, there has been a substantial endeavour in creating a vast array of indicators.[1] Parallel to this effort of providing granular information on very specific aspects of quality of care, there is also a significant body of work dedicated to the pursuit of composite measures of quality.[2] Recently, Kara et al. [5] presented an extensive and up-to-date review of various methodologies for the construction and use of composite quality of care.[3] The significant drawbacks of most composite indicators used on a national and international scale to assess healthcare quality (*e.g.*, [2,6–11]), coupled with the aim of extracting pertinent policy and "political" implications, have raised considerable doubts about their informational value. A key concern

* Corresponding author.
E-mail addresses: francesco.vidoli@uniurb.it (F. Vidoli), elisa.fusco@unifi.it (E. Fusco), giacomo.pignataro@unict.it (G. Pignataro), guccio@unict.it (C. Guccio).

[1] Beaussier et al. [1], in a study on how statutory hospital regulators in four countries (France, England, Germany, and the Netherlands) measure quality of healthcare provision, surveyed 1,100 different indicators of quality of care. At the international level, the WHO (World Health Organisation) and the OECD (Organisation for Economic Co-operation and Development) have developed sets of indicators for measuring the performance of healthcare systems in a very broad sense, which are not as large as the ones at the national level, but still include numerous measures. For WHO, see https://www.who.int/data/data-collection-tools/harmonized-health-facility-assessment/introduction; for OECD, see https://www.oecd.org/health/health-systems/health-care-quality-outcomes-indicators.htm.

[2] As Smith [2] noted, within the analysis of healthcare system performance, "*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*" ([2], p. 298).

[3] 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 (https://www.who.int/publications/i/item/924156198X); the English NHS star rating system for hospitals (since NHS documents are no longer publicly accessible at the original URLs, see a brief representation of the star rating system in [3]); the Hospital Compare Overall Hospital Quality Ratings by the US Centers for Medicare and Medicaid Services (https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/hospitalqualityinits/hospitalcompare).

https://doi.org/10.1016/j.seps.2024.101877
Received 25 November 2023; Received in revised form 13 March 2024; Accepted 20 March 2024
Available online 22 March 2024

revolves around the reliability of these indicators, specifically how accurately they reflect true differences in quality through the variations in performance they measure. Although composite indicators have the potential to enhance reliability compared to single indicators due to the inclusion of a larger number of data (on this issue, see [10]), they are susceptible to biases stemming from discretionary decisions on weighting and aggregating component indicators, as well as from heterogeneity among units evaluated. Additionally, the actionable value of composite indicators,[4] that is their ability to provide information on actionable steps for performance improvement, is limited, as such improvements are contingent on the measures of individual indicators. These limitations pose the risk of rendering composite indicators entirely ineffective in providing valuable information about healthcare quality.

The primary objective of this paper is to introduce a methodology to assess healthcare quality through composite measures, in addition to quantifying the possible improvements that can be achieved within individual domains of care. This methodology aims to provide a comprehensive framework to assess health care quality and identify areas for improvement in different aspects of care. To achieve this objective, we extend a recent methodology called Multi-directional Benefit of the Doubt (MDir_BoD), introduced by Fusco [13]. MDir_BoD is part of the broader group of methodologies known as Benefit of the Doubt (BoD - [14]), which construct composite measures from individual indicators by comparing performance across a set of units (healthcare systems, providers, etc.) using a non-parametric frontiers approach. MDir_BoD, specifically, enables the estimation of a composite measure while identifying potential improvements for each component indicator, resulting in the desired enhancement of the overall performance. We extend the MDir_BoD methodology to address potential biases in composite measure estimates resulting from significant heterogeneity within the set of units and the presence of outliers, thus strengthening its robustness. We then use this new methodology, which we call Multi-directional Robust Benefit-of-the-Doubt (MDir_RBoD), to measure the quality of acute healthcare services in 29 OECD countries. We use the most recent data provided by the OECD for the four indicators related to acute care within the Healthcare Quality and Outcomes programme (HCQO): AMI (acute myocardial infarction) 30-day mortality; hemorrhagic stroke 30-day mortality; ischaemic stroke 30-day mortality; and hip fracture surgery started within 2 days of admission to hospital. Composite scores are calculated for each country and potential improvements for each of the four indicators are quantified for each country as well.

We believe that this paper offers different contributions to the existing literature on the assessment of quality of healthcare through composite measures. First, we depart from the predominant methodologies used to measure composite indicators within the healthcare domain, instead opting to work within the BoD framework. To our knowledge, there have been no prior attempts to utilise BoD for creating a composite measure of healthcare quality, with the exception of two recent studies by Matos et al. [15] and Pereira et al. [16]. We think that employing a BoD composite measure can effectively tackle one of the main challenges faced in this literature, that is how to determine weights for the component indicators, which is crucial for maintaining the measure's reliability. The main problem here is not solely determining the varying weights for individual indicators, but rather the potential inconsistency of weights across different units. In fact, prioritising different health needs is influenced by social value judgments at the system level, which can vary between countries, or by the specific composition of health needs within a particular geographic area. Employing a uniform set of weights to assess the performance of decision-makers responsible for meeting the needs of their population could lead to misleading conclusions. The BoD approach is characterised by the endogenous derivation of the indicators' weights, separately for each unit under evaluation. In contrast to prevalent methodologies that allocate weights discretionally and uniformly across the units, the BoD approach avoids subjective judgment, thereby reducing the potential for arbitrary distortions in measuring quality variations among units. Furthermore, it enables the consideration that various units may assign different priorities to the distinct domains of care, evaluated by individual indicators.

Second, our choice of the specific BoD methodology represents an additional enhancement in the reliability of the composite measure of the quality of healthcare. It departs from the assumption of perfect compensability between the performances assessed by the various individual indicators. This departure is vital for the healthcare sector, as it acknowledges the possibility that potential health losses resulting from low-quality treatments or care settings may not be counterbalanced by the corresponding health gains from higher quality treatments or care settings. Furthermore, our extension of the MDir_BoD introduces a novel methodological enhancement in terms of the robustness of composite measures, mitigating their susceptibility to outliers — units or groups of units that exhibit significant heterogeneity in nature or size compared to others. Integrating the solutions to these challenges, within a framework where we concurrently assess multidimensional enhancements across various individual performances, an aspect often addressed separately in recent advancements of the BoD literature (as detailed in Section 2), distinguishes our work from previous studies [15,16], which also utilises the BoD methodology.

Finally, another contribution of our work pertains to the issue of "actionable" information derived from the evaluation of healthcare quality. Our methodology prevents the evaluation of composite performance from overshadowing the insights gleaned from basic indicators. Through our approach, we can consistently link overall performance, as measured by the composite indicator, to the performance of individual indicators by gauging the potential improvement of each.

The paper proceeds as follows. In Section 2, we will briefly review the BoD literature to highlight how the approach, in general, and the particular models we intend to use, can address the specific issues arising from the objective of measuring composite indicators for the quality provision of healthcare. Following this, in Section 3, we outline the technical details of our original methodology, MDir_RBoD. Section 4 presents a simulation conducted to test our methodology, while Section 5 presents the application of MDir_RBoD in assessing the quality of acute care using OECD data. Finally, the concluding section offers some closing remarks.

2. Literature review

As highlighted in the Introduction, it is widely recognised that the use of composite indicators in the assessment of quality in healthcare care, as well as in other domains, involves several crucial technical decisions that can impact their reliability. Specifically, these decisions affect their ability to accurately portray the genuine variability of quality among the units under examination. Two key technical concerns involve determining the weights of the indicators of the individual components and establishing the aggregation criteria in the composite measure.

Regarding the weighting issue, Kara et al. [5] provide insights into various approaches used to assign weights to individual indicators within composite measures of health care. These methods range from assigning equal weights to all indicators, considered the simplest approach, to relying on expert judgment or statistical techniques like principal component analysis. Regardless of the specific weight set used in each application of composite indicators, it essentially reflects value judgments concerning the relative importance of the component indicators and the particular performance areas they assess. Even when opting for equal weights, it does not imply a lack of weighting; rather,

[4] For further exploration of the concept of indicator actionability, especially within an international context, refer to the work of Carinci et al. [12].

it suggests that all indicators are of equal importance in the overall assessment of performance, which is not necessarily the case. Additionally, particularly in linear aggregation of individual indicators, equal weighting means perfect compensability of the performances measured by each indicator. Setting aside concerns regarding the arbitrariness of expert-based weight assignment or the stringent assumptions of certain statistical techniques, our aim is to highlight how, regardless of the method used to determine weights, employing a uniform set of weights across units under examination – essentially establishing a uniform set of priorities over the various objectives represented by different indicators – may inaccurately represent differences in quality performance among these units. Following Jacobs et al. [17], weights can be conceived as preferences over different objectives, and preferences are generally diverse between different decision-makers and within the general public. What Jacobs et al. [17] 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 favour some areas of care (*e.g.*, hospital care) relative to others (*e.g.*, primary care) due to 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 benchmarked units may not be meaningful in assessing their overall quality performance in the provision of healthcare. It may end up in penalising certain units that have performed excellently in areas with lesser weights, solely because they assigned higher priority to those areas, in contrast to other units. Opting for a model within the BoD framework [14] enables the variability of weighting between different units, under the assumption that each unit, within its own context, makes allocation decisions to optimise performance across the various objectives measured by the individual indicators.

BoD frontier composite indicators are derived from the concept of frontier analysis, which originates in the field of production economics. Starting with the seminal papers of Cherchye et al. [18,19] and Zhou et al. [20], the baseline BoD method served as a basis for theoretical improvements regarding the aggregation criterion [21–27].

The basic BoD model, despite its undeniable benefits in determining endogenous weights, is subject to certain drawbacks mainly related to the linear optimisation procedure itself. A comprehensive review of the literature that addresses these limitations and the methodological advances that stem from such discussions is beyond the scope of this paper. However, recent work by Ferreira et al. [28] provides an up-to-date and in-depth understanding of these developments. Here, we will only briefly address specific issues that directly impact the main objective of our paper: the influence of BoD method limitations on the reliability and actionability of the use of composite indicators in healthcare. Through this discussion, we aim to offer a better perspective on how our contribution may address and overcome these limitations. One of the main problems that can affect the ability of composite indicators to measure the variability of performance in the different units 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. This problem is of special relevance in the field of healthcare. Since a composite indicator aims to represent a "social" evaluation of healthcare quality, particularly in terms of health outcomes between different treatments and patients, it should not be "acceptable" for shortcomings in one area to be completely offset by exceptional performance in others. This is crucial because underperformance in certain objectives can lead to direct or indirect health losses or a diminished potential for health gain, and such losses cannot be adequately compensated for gains in other health areas. Furthermore, adhering to the same historical performance mix may not constitute an optimal management strategy to achieve a well-balanced healthcare system. An important methodological advancement within the BoD framework that addresses the issue of compensability is the Directional BoD (DBoD) method, introduced by Fusco [29]. DBoD imposes non-compensability within BoD through the use of a "direction" vector, which represents an explicit preference structure over the individual indicators. This vector indicates the desired direction of improvement across the various indicators to improve overall performance. Unlike BoD, where the direction of improvement is implicitly determined by the actual trade-off between individual indicators, DBoD utilises a subjective preference structure to define this direction instead. Several authors have revised the basic BoD approach from a directional standpoint, incorporating weight restrictions [30,31], addressing undesirable indicators [32], treating subindexes as non-compensatory [33], and extending the directional approach to panel data [34]. Meanwhile, several empirical studies adopting a directional approach have appeared in various fields such as public health [35], education [36], and in evaluating the environmental performance of municipal utilities [37].

However, selecting the optimal or preferred direction has consistently posed an inherent constraint within a procedure designed to avoid subjective decisions. Fusco [13] introduced the Multi-directional Benefit of the Doubt model (MDir_BoD), as an extension of the Directional BoD. This model addresses non-compensability among individual indicators by deriving a unit-specific preference structure (the direction) directly from the data. This can be achieved by separating benchmark selection from efficiency measurement. The benchmark selection is based on adjustments in simple indicators proportional to the potential improvements given by the specific excesses of input/output. 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 [13], *"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"* ([13], p. 5). This implies that, in addition to the overall composite score, specific scores are determined for each individual indicator. Consequently, these scores can be analysed independently and used to offer practical recommendations for enhancing unit performance, thereby improving what is referred to as the actionability of the composite indicator. This is not possible in BoD and DBoD because information on input-specific or output-specific inefficiencies (*i.e.*, indicators specific) is obscured. This feature is of significant importance in the healthcare sector, where the aggregation of individual indicators may obscure the root causes of poor performance and hinder the identification of effective remedies to enhance overall performance [17]. This concern is particularly pronounced when composite indicators are used to rank units under examination (such as countries or providers), as is the case with prominent applications such as WHO, English star rating, and CMS hospital comparison. In such cases, the information provided by the composite measure's score is overshadowed by its use in unit rankings, which becomes the primary message conveyed by rankings.[5] However, this limitation can be overcome by supplementing the information from the composite measure with information on how each component contributes to overall performance, as happens with MDir_BoD.

MDir_BoD, like BoD and DBoD, is not robust to outliers, causing the frontier to shift towards the outlier and resulting in underestimated performance scores for all other observations. Ensuring robustness to

[5] As noted by Oliver [38], with respect to international rankings of healthcare systems, "*Some countries seem to perform very well on specific aspects of healthcare, 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*" ([38], p. 17). On this issue, see also Street and Smith [39].

outlier data is crucial as it guarantees the reliability and stability of composite indicator results across different data sources or variations in methodology over time. Robustness enhances the credibility and trustworthiness of composite indicators among policymakers and stakeholders, enabling them to identify best practices, areas needing improvement, and potential interventions during comparative analyses between countries. From a methodological perspective, Vidoli and Mazziotta [40] proposed a robust approach to the BoD method (called RBoD), which employs a resampling procedure to reduce the impact of outliers on the scores of the other units under analysis. Robust Directional BoD (RDBoD, [41]) and its subsequent spatial extension [42] aim to combine the benefits of robust estimation within a directional model. In this paper, we extend MDir_BoD to achieve robustness.

3. Robust multidirectional BoD

In order to enhance the lack of robustness control in MDir_BoD approach a resampling procedure of the MDir_BoD, analogous to the RBoD and RDBoD methods [40,41], is proposed here.

The underlying idea is to isolate and thereby reduce the effect of outliers and abnormal values by repeatedly comparing each unit to subsets of observations of size $m < N$ instead of the entire dataset, thus obtaining a maximal expected frontier of order m (maximum expected achievable level of the composite indicator (CI) among m units drawn in the total observations).[6]

In formal terms, as usual in BoD literature, let us consider a matrix of q simple indicators treated as outputs ($Y_q \in \mathbb{R}_+, \forall q = 1, \dots, Q$) and an input vector equal to one for all N observations i. To formalise the resampling process, a probabilistic formulation of the production set (the set of all possible simple indicators values) Ψ is required, and, can be defined as:

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

where $H(\mathbb{1}, \mathbf{y}) = Prob(X \equiv \mathbb{1}, \mathbf{Y} \geq \mathbf{y})$ is the probability of finding a unit with a combination of simple indicators (the CI) greater than that of the unit $(\mathbb{1}, \mathbf{y})$.

As described in detail in Fusco [13] the idea of MDir_BoD is to examine the maximum possible increase in a single simple indicator to reach the frontier, assuming that the other indicators do not increase, and then to find the corresponding directions. Therefore, in accordance with previous formulation, the maximum possible increment of the qth indicator for a specific unit, *i.e.* $\hat{\mathbf{y}}_q$, can be obtained by solving Q linear programming problems, that maximise each indicator q, keeping the remaining simple indicators $\mathbf{y}_{-q}$ fixed:

$$\hat{\mathbf{y}}_q = \sup \left\{ \mathbf{y}_q | (\mathbb{1}, \mathbf{y}_q, \mathbf{y}_{-q}) \in \Psi \right\}, \forall q = 1, \dots, Q \tag{2}$$

where $\hat{\mathbf{y}}_q$ is a vector of size N and so containing the values of all observations.

In this paper, to introduce robustness and attenuate the effect of abnormal or outlying units, each unit is compared to subsets of m observations, as mentioned above. Therefore, in a probabilistic framework, 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 $\mathbf{Y}$, a random set $\widetilde{\Psi}_m$ is considered and defined as:

$$\widetilde{\Psi}_m = \bigcup_{j=1}^{m} \{ (\mathbb{1}, \mathbf{y}) \in \mathbb{R}_+^{1+Q} | X \equiv \mathbb{1}, \mathbf{Y_j} \geq \mathbf{y} \}. \tag{3}$$

This generalisation enables the iterative computation of the sample subset of size m (for $b = 1, \dots, B$ times) and, for each iteration b, the maximum possible increment for the single indicator, following Eq. (2), is given by:

$$\widetilde{\mathbf{y}}_{m;q}^b = \sup \left\{ \mathbf{y}_q | \left(\mathbb{1}, \mathbf{y}_q, \mathbf{y}_{-q} \right) \in \widetilde{\Psi}_m \right\}, \ \forall b = 1, \dots, B; \ q = 1, \dots, Q \tag{4}$$

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

Given the maximum possible increment for each single indicator at iteration b, *i.e.* $\widetilde{\mathbf{y}}_{m;q}^b$, the *potential improvements* of each unit, *i.e.*, the directional vector (at iteration b) can be calculated as follows:

$$\widetilde{\mathbf{g}}_m^{PI_b} = (\widetilde{\mathbf{y}}_m^b - \mathbf{y}_m) = (\widetilde{\mathbf{y}}_{m;1}^b - \mathbf{y}_{m;1}, \dots, \widetilde{\mathbf{y}}_{m;q}^b - \mathbf{y}_{m;q}, \dots, \widetilde{\mathbf{y}}_{m;Q}^b - \mathbf{y}_{m;Q}), \forall b = 1, \dots, B \tag{5}$$

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 (given by the distance between an ideal reference point and the observed point in the dataset). It is important to emphasise that since the directions are different for each simple indicator and observation, the method achieves non-compensability by penalising an unbalanced mix of simple indicators in a customised and objective way for each unit. Once the b vectors of the directions have been found, the b benchmark selections (on the frontier), relative to the specific potential improvements of the simple indicators, are obtained with a further maximisation problem:

$$\widetilde{D}_m^b(\mathbb{1}, \mathbf{y}; \widetilde{\mathbf{g}}_m^{PI_b}) = \sup \left\{ \beta | \left(\mathbb{1}, \mathbf{y} + \beta \widetilde{\mathbf{g}}_m^{PI_b} \right) \in \widetilde{\Psi}_m \right\}, \forall b = 1, \dots, B \tag{6}$$

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 B distributions obtained in the previous iterative steps, *i.e.*:

$$\mathbf{g}^{PI} = E \left[\widetilde{\mathbf{g}}_m^{PI_1}, \dots, \widetilde{\mathbf{g}}_m^{PI_b}, \dots, \widetilde{\mathbf{g}}_m^{PI_B} \right] \tag{7}$$

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

$$D(\mathbb{1}, \mathbf{y}; \widetilde{\mathbf{g}}_m^{PI}) = E \left[\widetilde{D}_m^b \left(\mathbb{1}, \mathbf{y}; \widetilde{\mathbf{g}}_m^{PI_1} \right), \dots, \widetilde{D}_m^b \left(\mathbb{1}, \mathbf{y}; \widetilde{\mathbf{g}}_m^{PI_b} \right), \dots, \widetilde{D}_m^b \left(\mathbb{1}, \mathbf{y}; \widetilde{\mathbf{g}}_m^{PI_B} \right) \right] \tag{8}$$

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

The 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 $N \times Q$.

Having determined both the proportion by which each of the simple indicators must be increased in order to reach the frontier and the direction, the relative robust multi-directional scores vector for the simple indicator q, can be calculated as the following:

$$e_q = \frac{\mathbf{y}_q}{\mathbf{y}_q + \beta^* \mathbf{g}_q^{PI}} \tag{9}$$

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

The overall CI_{MDir_RBoD} scores, to be consistent with Fusco [13], are determined as the difference to 1 of the *normalised potential improvements inefficiency index*, proposed by Bogetoft and Hougaard [43], related to the benchmark:

$$CI_{MDir\_RBoD} = 1 - \frac{\beta^* \sum_{q=1}^{Q} \mathbf{g}_q^{PI}}{\sum_{q=1}^{Q} \mathbf{y}_q + \beta^* \mathbf{g}_q^{PI}} \tag{10}$$

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:

$$\bar{x} \pm t \cdot (s/\sqrt{B}) \tag{11}$$

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.

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

It is important to emphasise, again, that MDir_RBoD as MDir_BoD provides a weighting scheme that *"is not only the most favourable for each unit, but is the <<most favourable in the desirable direction>> looking for potential improvements instead of following past production (like in BoD) or a specific direction (like in DBoD)."* [13] and that it adds a further enrichment, *i.e.*, the comparison with a frontier robust to outliers and abnormal units.

4. Simulations

In this Section, a descriptive example has been implemented using simulated data. To better show the enhancement achieved through MDir_RBoD, we compare the computed CIs using this methodology with those obtained from its non-robust counterpart, MDir_BoD. This comparison highlights the incremental value of robustness over a methodology that is directly comparable. Furthermore, we will also include a comparison with the results of the CI calculation using a basic BoD, thus providing information on the improvements achieved compared to a substantially different methodology that, in contrast with MDir_RBoD, assumes perfect compensability and does not offer information on potential enhancements to overall performance arising from the single indicators (as explained in Section 2).

In the baseline setting (Fig. 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 (Fig. 1.a, case 0); in the second one, maintaining simulated non-abnormal data, an outlier with both out-of-scale indicators (called *proportional outlier*, Fig. 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*, Fig. 1.c, case 2) has been added causing only a partial shift of the frontier (please note that only Unit21 remains on the frontier).

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 analyse the results on the 50 non-outlier units only. Therefore, we 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 presents the average absolute differences in the CI scores computed using the MDir_BoD method and its robust version, MDir_RBoD, in comparison to the standard BoD method, as well as between MDir_BoD and MDir_RBoD themselves, under base case 0 (no outliers or out-of-scale units). Table 2 displays the average absolute differences in CI scores between the three methods under base case 0 and those calculated under case 1 (proportional outlier) and case 2 (non-proportional outlier). Table 3 shows the average absolute value differences between the directions derived from MDir_BoD and MDir_RBoD for both the first (Indic1) and the second (Indic2) indicator.

Some findings emerge:

- In case 0 (third column of Table 1 and first block of columns of Table 3), 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 both the directions on average and for each unit (Fig. A.1), showing substantial stability in the directions of improvement for each unit. The *p*-values for the *t*-tests (in brackets) confirm that the results are equal on average.
- In case 1 (shown in the first column of Table 2 and in the second block of columns of Table 3), notable differences among the methods are apparent. While both the BoD and MDir_BoD methods show considerable impact on CI scores compared to the robust version, the improvement brought about by the MDir_RBoD method is particularly evident (0.349 and 0.348 with statistically significant *t*-test *p*-values compared to 0.05, which is not statistically significant). This observation also holds for the mean changes in direction (0.709 and 0.704 vs. 0.093 and 0.061). Fig. 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 (Fig. 2.b).
- In the case of non-proportional outliers (case 2, represented in the second column of Table 2 and in the third block of columns in Table 3) the MDir_BoD method exhibits slightly different results (0.279 and 0.09 with a borderline *t*-test significance, compared to 0.074 which is not statistically significant). On the one hand, there is a minor distortion in terms of the CI score; however, an interesting aspect emerges with respect to directions. Specifically, the directions are distorted with respect to the outlier indicator, while preserving the other dimension (refer also to Fig. 2.c for MDir_BoD). The robust method (Fig. 2.d) again allows the biasing effect of the outlier to be controlled.

The generalisation of this particular illustrative setting is not straightforward because it is clearly linked with our specific simulation/hypothesis on the input data. But still, within our illustrative

Table 1
Average distances (in absolute term) among methods. CIs calculated in no outliers case 0 (t-test p-values in brackets).

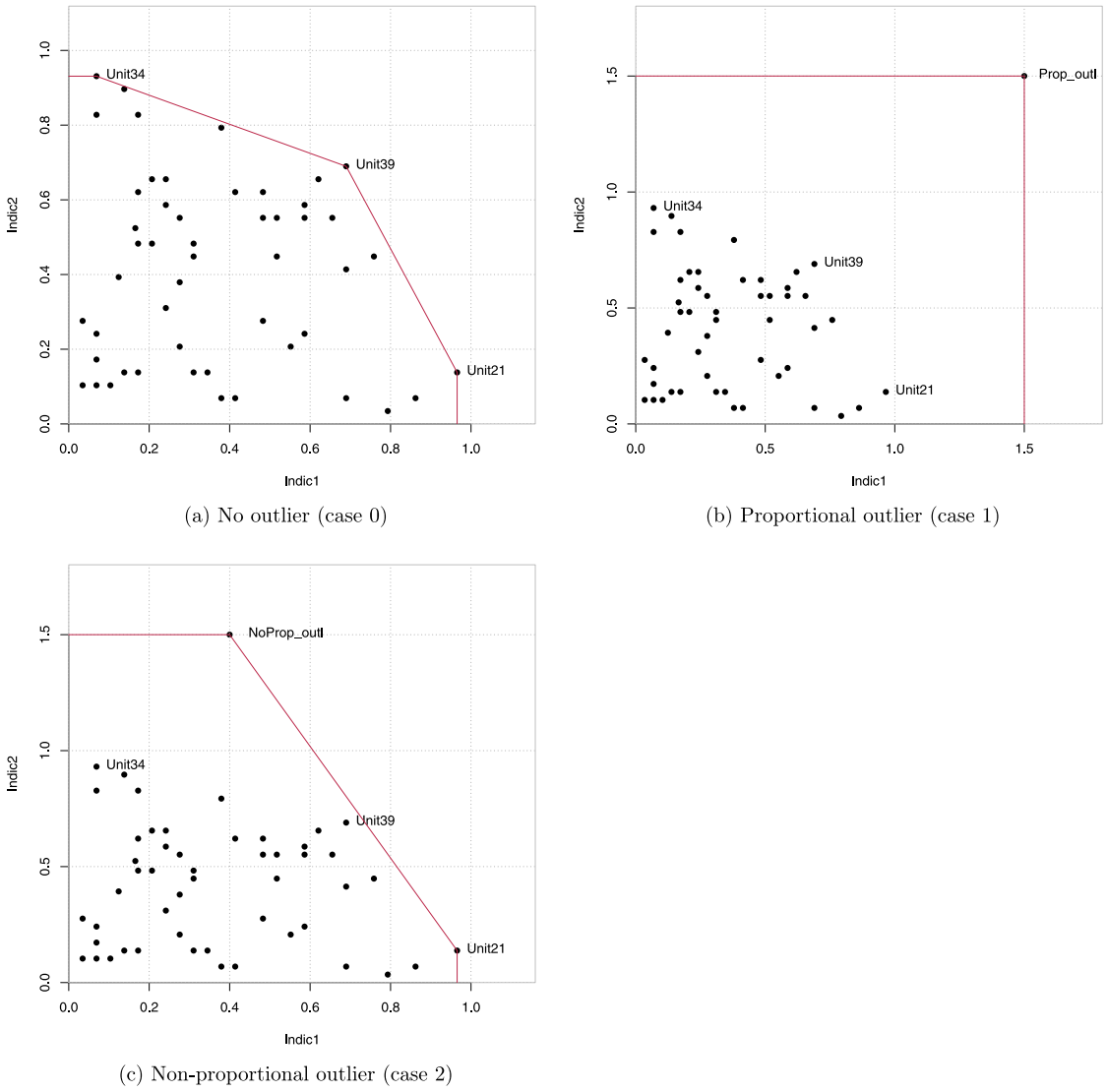
	BoD vs MDirBoD	BoD vs MDirRBoD	MDirBoD vs MDirRBoD
No Outlier	0.036 *(0.495)*	0.036 *(0.592)*	0.066 *(0.229)*

Table 2
Average differences (in absolute term) between the CIs calculated in no outliers case and the other settings and method (t-test p-values in brackets)

	Proport. out (case 1)	Non-Proport. out (case 2)
BoD	0.349 *(0.000)*	0.279 *(0.072)*
MDir_BoD	0.348 *(0.000)*	0.099 *(0.050)*
MDir_RBoD	0.050 *(0.547)*	0.074 *(0.207)*

Table 3
Average differences (in absolute term) between the MDir_BoD directions calculated in no outliers case and the other settings and method (t-test p-values in brackets).

	No outlier (case 0)		Proport. out (case 1)		Non-Proport. out (case 2)	
	Indic1	Indic2	Indic1	Indic2	Indic1	Indic2
MDir_BoD			0.709 (*0.000*)	0.704 (*0.000*)	0.058 (*0.278*)	0.494 (*0.000*)
MDir_RBoD	0.134 (*0.012*)	0.103 (*0.045*)	0.093 (*0.271*)	0.061 (*0.547*)	0.172 (*0.001*)	0.076 (*0.228*)



(a) No outlier (case 0)

(b) Proportional outlier (case 1)

(c) Non-proportional outlier (case 2)

Fig. 1. Simulated cases.

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 examples 1 to 4) of the outlier shown in Fig. 1.c, what would the results be in terms of the `MDir_BoD` and `MDir_RBoD` computations?

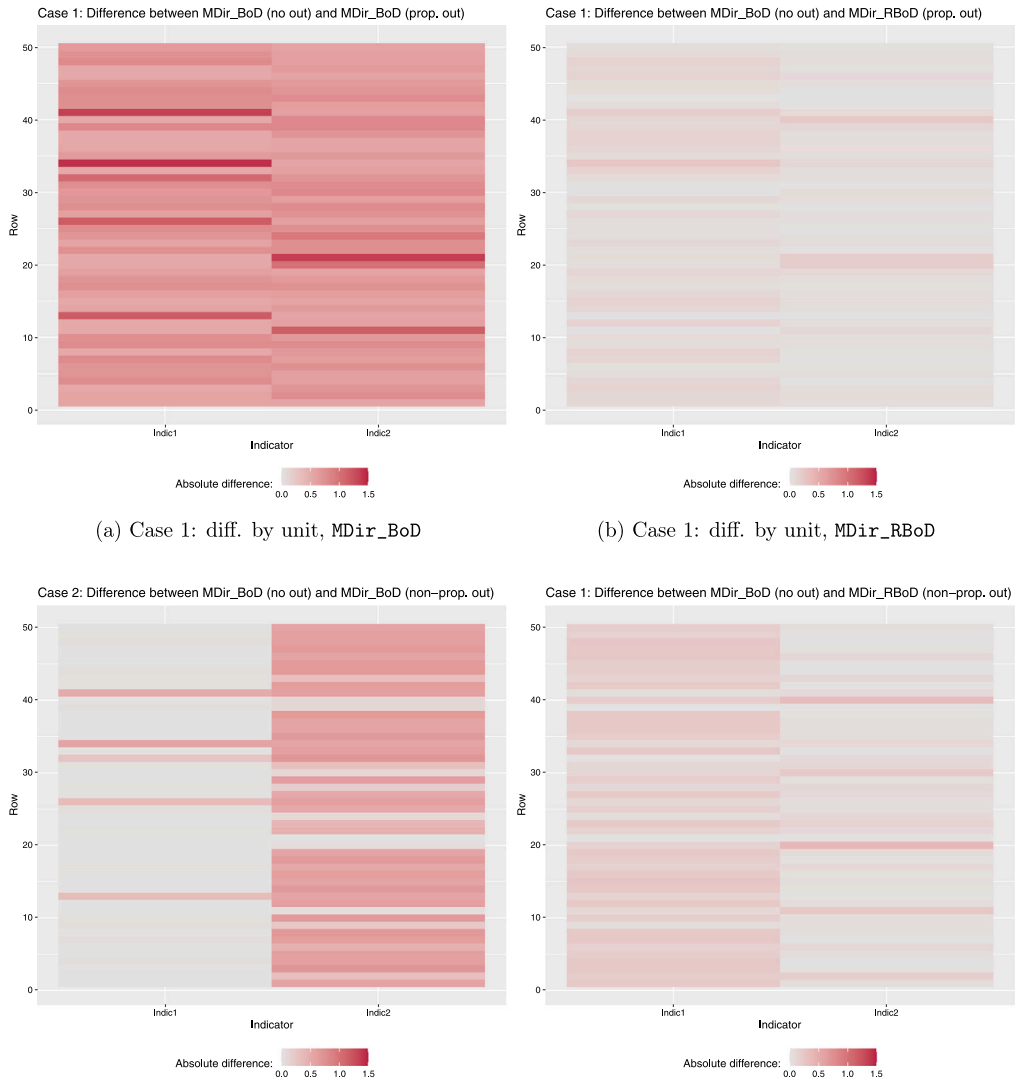
Fig. 3 shows that as the magnitude of the random disturbance outlier increases, the differences from the base case without outlier (case 0) and with non-proportional outlier (case 2) in terms of CI score increase significantly in the non-robust case, while conversely, they are optimally controlled by our robust method.

5. Quality of acute healthcare in OECD countries

In this section, we apply the methodology developed in Section 3 to the assessment of acute care quality in some OECD countries. In 2001, the OECD started the Healthcare Quality Indicators project with the aim of collecting data for the development and reporting of healthcare quality indicators and for international comparisons. Since then, the initiative has evolved with the incorporation of quality measures into a conceptual framework for the performance of the health system [44]. The current Healthcare Quality and Outcomes Programme (HCQO) includes a total of 64 indicators and covers 40 countries.[7] The data are regularly published on a dedicated page on the OECD website and in the organisation's well-known publications, such as Health at a Glance. To our knowledge, the OECD has not attempted to provide a measure of the overall quality of healthcare in different countries, and cross-country comparisons can only be made for each individual indicator.

Indicators are grouped into the following areas: acute care; cancer care; mental healthcare; 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 acute care indicators: AMI (acute myocardial infarction) 30-day mortality; hemorrhagic stroke 30-day mortality; ischaemic stroke 30-day mortality; and hip fracture surgery started within 2 days of admission to the hospital. We are aware that this

[7] Details of the programme can be found on the dedicated page of the organisation's website https://www.oecd.org/health/health-care-quality-outcomes-indicators.htm. For an analysis of the evolution of the OECD initiative, please see Carinci et al. [12].



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

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

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

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

Fig. 2. Direction differences by unit between MDir_BoD (case 0) and other case/method.

choice does not take full advantage of our methodology to provide information on the overall performance of the healthcare systems under examination, 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 latest available year, for 29 countries,[8] with the aim of computing a composite indicator of the quality of acute care at national level,[9] applying the `MDir_RBoD` developed in Section 3. As we did for the simulation exercise, in Section 4, we compare the results of the CI calculation with the ones obtained using `BoD` and `MDir_BoD`.

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

An examination of these elementary data reveals no obvious outliers, with the exception of Iceland, which stands out due to its smaller and younger population compared to the other countries. This could potentially pose challenges in terms of comparability. However, our robust approach enables us to address and control this aspect effectively.

Based on these data, `BoD`, `MDir_BoD` and `MDir_RBoD` CIs have been computed, with the corresponding scores presented in Tables A.1.

[8] 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. For this reason, only the last available year was considered, which is therefore not the same for all countries. Australia, Chile, France, Japan, Korea, Luxembourg, Mexico, Poland, the Slovak Republic and the United States had at least one indicator with missing data for all years considered, and Costa Rica presented unreliable data. Therefore, these countries were removed from the analysis set because they were not fully comparable with the other countries and the subsequent analysis was carried out only in 29 countries.

[9] National data have been originally standardised by age, sex, co-morbidity. The ratios are based on linked data focused on each patient (a single patient is counted only once) and use them as a denominator (irrespective of the number of admissions or readmissions). The ratios consider deaths that occur anywhere, including within or outside the hospital of admission.

Table 4
Results descriptive statistics.

Statistic	N	Mean	St. Dev.	Min	Max
BoD	29	0.814	0.190	0.158	1.000
MDir_BoD	29	0.687	0.209	0.040	1.000
MDir_RBoD	29	0.745	0.214	0.044	1.000

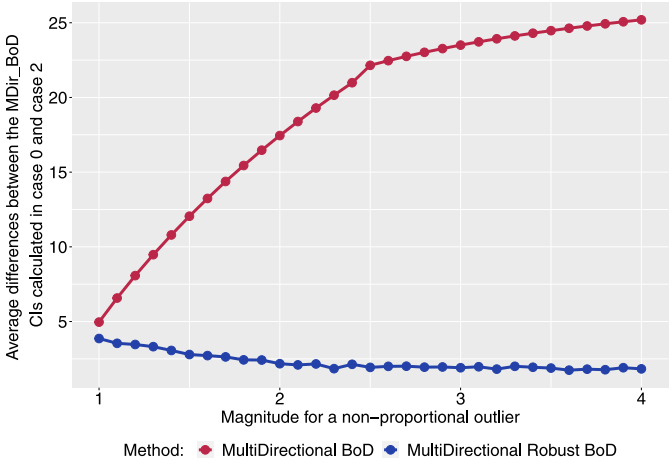


Fig. 3. Average differences (in absolute term) between the MDir_BoD CIs calculated in case 0 and case 2 by the magnitude of the outlier unit and by method.

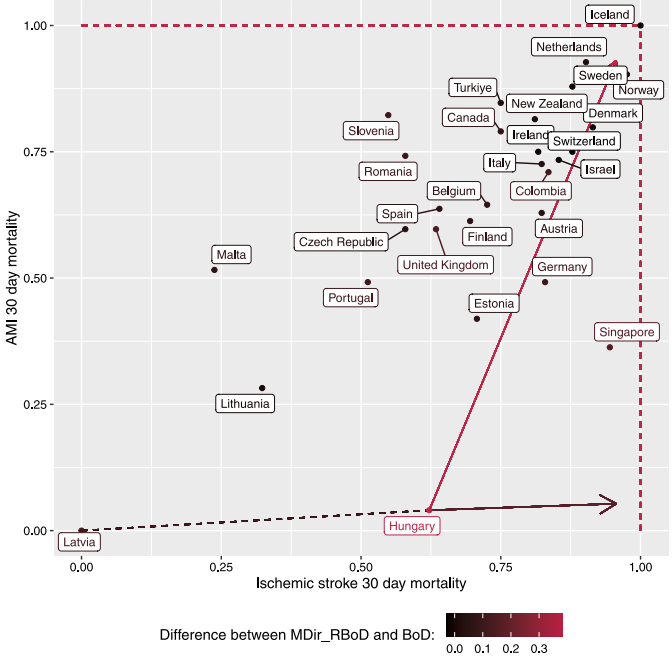


Fig. 4. Estimated MDir_RBoD (red) vs BoD (black) radial direction.

The measures are standardised relative to the scores of the best performers (Iceland for all three methods). Additionally, Table 4 reports the descriptive statistics of the results. It should be noted that the method also allows the computation of confidence intervals for the robust composite indicator by resampling. This approach offers an indirect estimate of the reliability of the indicator.

An initial and very general observation is that the use of CIs allows to have an overview, a "big picture" [7], regarding the quality performance of various countries in delivering acute care services. This information proves particularly valuable for countries that exhibit mixed performance in all four indicators. In such cases, insights from individual indicators are not sufficient to gauge a country's relative overall performance. The results of Table A.1 not only reaffirm the strong performance of Nordic European countries and countries such as Switzerland, New Zealand and Canada, but also allow an appreciation of the overall performance of other countries such as Turkey and Romania, which excel in certain indicators but lag behind others.

Regarding the distinctive feature of MDir_RBoD, namely its robustness, a comparison of the two directional methods alone reveals that isolating outliers with MDir_RBoD prevents the performance of the other units from being significantly underestimated, as observed with MDir_BoD.

Significant findings are associated to the difference between assuming the perfect compensability of performances under the single indicators and not assuming it. From both descriptive statistics and individual country scores, it is noticeable that MDir_RBoD scores generally fall between BoD and MDir_BoD scores. This pattern arises because countries exhibiting imbalanced performance across the four indicators receive penalties from both MDir_BoD and MDir_RBoD, with the severity of the penalty increasing with the degree of imbalance. The penalisation for uneven quality provision is notable for several countries, including Canada, Singapore, UK, Finland, Spain, and Portugal. However, it is especially pronounced for Hungary, which experiences a highly imbalanced quality provision across the four areas. Fig. 4 provides a clearer illustration of the case of Hungary, focussing, without loss of generality, solely on two indicators and comparing BoD with MDir_RBoD. The axes measure the normalised values of the two indicators for each country.

Hungary shows the lowest value of the AMI indicator (by construction, its normalised and polarised value is zero), while the value of the ischaemic stroke indicator is notably higher. With the BoD method, the composite indicator is computed as a radial distance (depicted by the black line) from the deterministic frontier (illustrated by the red dashed line). This method allows maximum overall quality to be achieved by only increasing the value of the ischaemic indicator, thus compensating for significant underperformance in AMI with strong performance in the other indicator. Conversely, the MDir_RBoD approach shows that the composite indicator, interpreted as the distance (indicated by the red line) from a benchmark[10] characterised by a balanced performance in the two indicators, is greater. It also suggests that an improvement in the AMI indicator is necessary, since Hungary's better position in the ischaemic stroke indicator cannot compensate for this deficiency.

Regarding the measure of the penalty relative to a BoD model, it hinges on the ratio of the distance from the multi-directional benchmark to that identified by the BoD (even in a multi-dimensional setting), as well as the presence and magnitude of outliers in the data. Thus, the penalty for each unit is influenced by numerous comparisons between distances, alongside the existence of anomalous data, which may not be present in all individual indicators. In essence, even if it is not possible to generalise, it is evident that a country performing exceptionally well on nearly all indicators but significantly poorly on one will not have a lower overall performance than a country performing marginally acceptably on all indicators. This trend becomes more pronounced as the number of indicators increases.

Finally, the MDir_RBoD method not only provides the overall quality performance of each country, as indicated by the MDir_RBoD

[10] It is worth noting that the red arrow (representing the distance from the benchmark) does not intersect with the point representing Iceland. The robustness of our indicator "mitigates" the impact of an outlier like Iceland, which is the only country on the frontier.

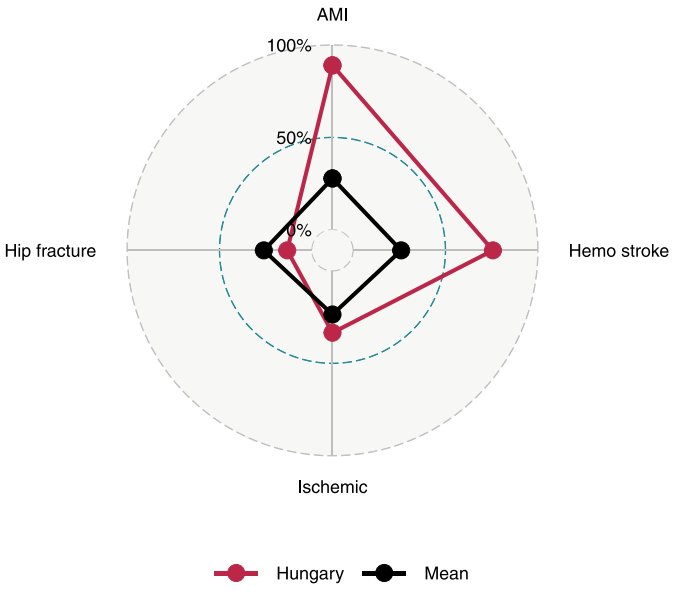


Fig. 5. Estimated directions for improvement — Hungary vs directions mean.

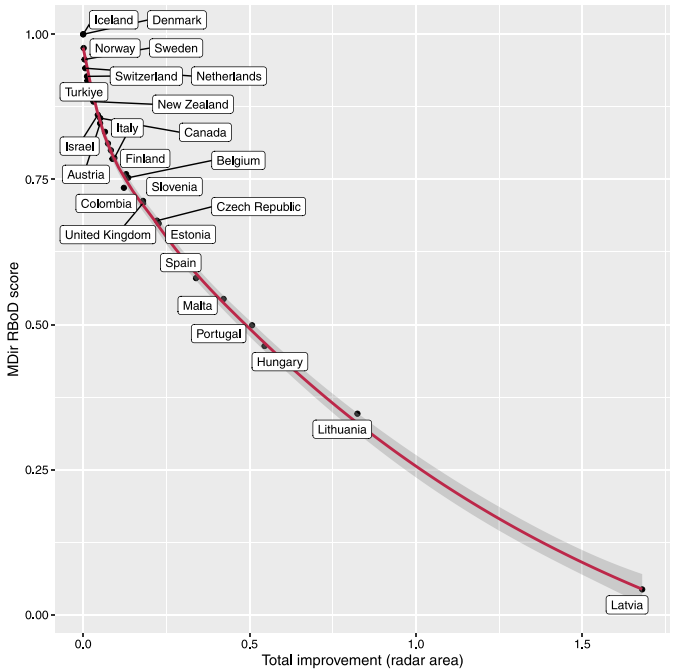


Fig. 6. `MDir_RBoD` score and the needed total improvement.

score, but also outlines the improvement trajectory for each country. This trajectory identifies the potential enhancement of each indicator required to transition towards the balanced benchmark on the frontier. The "direction" scores in Table A.2 quantify the fraction of "optimal" performance that each country lacks for each indicator. Essentially, they measure the relative effort necessary for each country to improve its overall performance along the path leading to the frontier. This information can be visually represented through a typical radar plot, as illustrated in Fig. 5.

Using Hungary again as an example, we can see that most of the work that this country must do to improve its relative low quality of acute care services (measured by a value of `MDir_RBoD` equal to 0.4611) must be done 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 lesser 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 (Fig. 6), that is, it is inversely related to the value of the `MDir_RBoD` score.

6. Concluding remarks

The purpose of this paper is to contribute to the literature on composite indicators and to their use to measure the quality of healthcare. Although individual indicators in the quality assessment of healthcare provide detailed information on specific aspects, in fact, they cannot capture all relevant information. Therefore, there has long been the need to construct composite measures to assess the overall quality or performance of healthcare, 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 how potential health losses in some areas of care are not necessarily offset by health gains in others.

Comparing the results of computing CIs using the method developed in this paper with other methodologies within the BoD framework has revealed how the robustness of `MDir_RBoD` enhances the reliability of the computation. The use of `MDir_RBoD` prevents scores from being biased by outliers, particularly common in international comparisons, and ensures that the measure accurately reflects relevant variations in quality, such as those due to imbalanced performances in the different indicators. Furthermore, applying this methodology to assess acute care quality in OECD countries and discussing the results has also shown how its features can significantly contribute to the actionability of the information derived from its use. It suggests specific directions for improvement in each performance area, thus facilitating the improvement of overall quality of care. It is also 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 example, it could be used to compare the overall quality performance of each country over time. Surely, it is more significant than single measures in statistical analyses of how policies or other contextual factors (like competition, for instance) impact the overall quality/performance of healthcare 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 individual providers, in such a way as to get relevant information for managerial actions.

From a methodological perspective, potential enhancements to the proposed robust and multidirectional approach could focus on three aspects: *(i)* integrating conditional approaches [45] to allow the computation of composite indicators with the incorporation of contextual factors; *(ii)* developing this approach within a hierarchical framework [46] to better assess multidimensional phenomena in a multi-level setting; *(iii)* exploring the possibility of extending the proposed approach within the context of multiplicative models [28] to address limitations associated to issues such as zero multiplier problems [47] and cases where mutual preferential independence is not admissible.

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

Country	AMI	Hemo.	Ischaemic	Hips	BoD		MDir_BoD		MDir_RBoD		
					Score	Rank	Score	Rank	Score	95% CI	Rank
Denmark	4.50	23.90	4.80	97.60	1.0000	(1)	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)	1.0000	(1.0000, 1.0000)	(1)
Norway	3.20	15.80	3.80	96.60	0.9949	(3)	0.9580	(3)	0.9750	(0.9763, 0.9776)	(3)
Netherlands	2.90	24.50	5.00	95.40	0.9746	(4)	0.8724	(5)	0.9386	(0.9416, 0.9446)	(5)
Turkiye	3.90	12.10	7.50	76.30	0.9520	(5)	0.8024	(9)	0.9158	(0.9200, 0.9242)	(7)
Sweden	3.50	15.30	5.40	93.70	0.9461	(6)	0.8994	(4)	0.9544	(0.9566, 0.9588)	(4)
Singapore	9.90	17.30	4.30	64.40	0.9451	(7)	0.6450	(18)	0.7785	(0.7854, 0.7923)	(16)
Canada	4.60	23.60	7.50	93.10	0.9290	(8)	0.8032	(8)	0.8525	(0.8556, 0.8587)	(10)
New Zealand	4.30	20.90	6.50	92.00	0.9134	(9)	0.8278	(7)	0.8812	(0.8841, 0.8870)	(8)
Germany	8.30	22.00	6.20	92.10	0.9130	(10)	0.7519	(13)	0.8083	(0.8119, 0.8156)	(13)
Switzerland	5.10	15.40	5.40	90.80	0.8991	(11)	0.8522	(6)	0.9246	(0.9275, 0.9304)	(6)
Austria	6.60	20.40	6.30	91.00	0.8969	(12)	0.7832	(11)	0.8432	(0.8465, 0.8499)	(11)
Romania	5.20	14.20	10.30	56.60	0.8927	(13)	0.6366	(19)	0.7933	(0.8002, 0.8071)	(14)
Israel	5.30	20.40	5.80	88.10	0.8550	(14)	0.7990	(10)	0.8582	(0.8612, 0.8642)	(9)
Colombia	5.60	15.60	6.10	37.30	0.8531	(15)	0.6023	(23)	0.7297	(0.7356, 0.7414)	(19)
United Kingdom	7.00	28.60	9.40	88.70	0.8524	(16)	0.6634	(17)	0.7065	(0.7091, 0.7117)	(21)
Hungary	13.90	40.50	9.60	88.30	0.8458	(17)	0.4347	(27)	0.4611	(0.4630, 0.4648)	(27)
Finland	6.80	23.40	8.40	86.80	0.8257	(18)	0.7068	(15)	0.7564	(0.7591, 0.7618)	(17)
Belgium	6.40	26.40	7.90	87.00	0.8257	(19)	0.7028	(16)	0.7504	(0.7529, 0.7554)	(18)
Italy	5.40	19.80	6.30	69.70	0.8232	(20)	0.7106	(14)	0.7834	(0.7870, 0.7906)	(15)
Slovenia	4.20	25.10	10.80	70.90	0.8226	(21)	0.6354	(20)	0.7089	(0.7130, 0.7171)	(20)
Ireland	5.10	22.30	6.40	85.40	0.8171	(22)	0.7690	(12)	0.8290	(0.8320, 0.8349)	(12)
Estonia	9.20	25.20	8.20	81.10	0.7328	(23)	0.6213	(22)	0.6715	(0.6740, 0.6766)	(23)
Czech Republic	7.00	25.80	10.30	80.90	0.7275	(24)	0.6283	(21)	0.6764	(0.6789, 0.6813)	(22)
Spain	6.50	27.70	9.30	55.60	0.6402	(25)	0.5283	(24)	0.5777	(0.5799, 0.5822)	(24)
Portugal	8.30	24.20	11.40	47.40	0.6102	(26)	0.4498	(26)	0.4965	(0.4987, 0.5008)	(26)
Malta	8.00	24.30	15.90	71.90	0.6073	(27)	0.4924	(25)	0.5415	(0.5440, 0.5465)	(25)
Lithuania	10.90	35.40	14.50	58.30	0.3513	(28)	0.3167	(28)	0.3450	(0.3464, 0.3478)	(28)
Latvia	14.40	45.80	19.80	46.80	0.1575	(29)	0.0400	(29)	0.0435	(0.0437, 0.0438)	(29)

CRediT authorship contribution statement

F. Vidoli: Writing – original draft, Validation, Software, Methodology, Data curation, Conceptualization. **E. Fusco:** Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **G. Pignataro:** Writing – original draft, Validation, Supervision, Formal analysis, Conceptualization. **C. Guccio:** Writing – original draft, Validation, Supervision, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A

See Figs. A.1–A.5

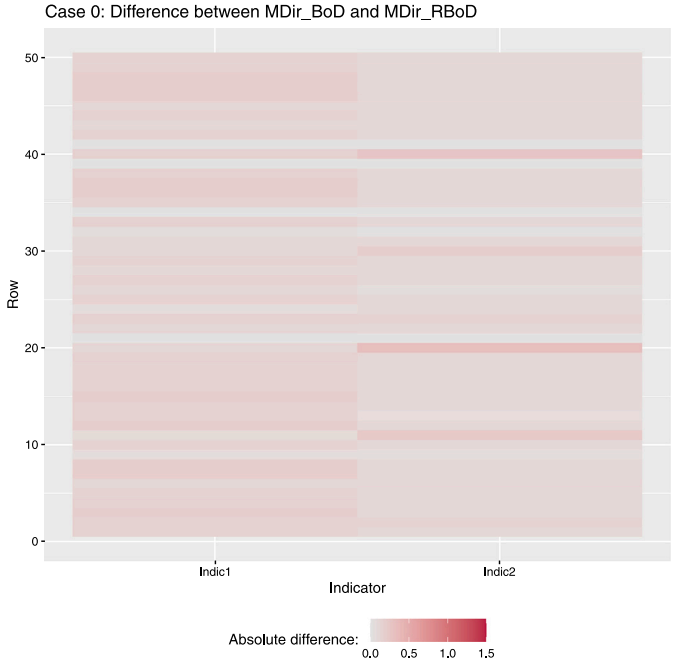


Fig. A.1. Direction differences by unit between MDir_BoD (case 0) MDir_RBoD.

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

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	0.9750	(0.9763, 0.9776)	(3)	0.039	0.062	0.010	0.001
Sweden	0.9544	(0.9566, 0.9588)	(4)	0.051	0.056	0.056	0.021
Netherlands	0.9386	(0.9416, 0.9446)	(5)	0.030	0.166	0.041	0.010
Switzerland	0.9246	(0.9275, 0.9304)	(6)	0.142	0.059	0.057	0.057
Turkiye	0.9158	(0.9200, 0.9242)	(7)	0.062	0.019	0.101	0.137
New Zealand	0.8812	(0.8841, 0.8870)	(8)	0.113	0.189	0.133	0.062
Israel	0.8582	(0.8612, 0.8642)	(9)	0.185	0.189	0.100	0.121
Canada	0.8525	(0.8556, 0.8587)	(10)	0.139	0.249	0.198	0.055
Austria	0.8432	(0.8465, 0.8499)	(11)	0.273	0.181	0.124	0.076
Ireland	0.8290	(0.8320, 0.8349)	(12)	0.176	0.246	0.137	0.173
Germany	0.8083	(0.8119, 0.8156)	(13)	0.408	0.217	0.121	0.064
Romania	0.7933	(0.8002, 0.8071)	(14)	0.136	0.058	0.236	0.389
Italy	0.7834	(0.7870, 0.7906)	(15)	0.192	0.181	0.128	0.406
Singapore	0.7785	(0.7854, 0.7923)	(16)	0.381	0.089	0.029	0.340
Finland	0.7564	(0.7591, 0.7618)	(17)	0.322	0.281	0.265	0.159
Belgium	0.7504	(0.7529, 0.7554)	(18)	0.288	0.363	0.233	0.157
Colombia	0.7297	(0.7356, 0.7414)	(19)	0.182	0.069	0.098	0.805
Slovenia	0.7089	(0.7130, 0.7171)	(20)	0.116	0.331	0.384	0.390
United Kingdom	0.7065	(0.7091, 0.7117)	(21)	0.336	0.418	0.323	0.129
Czech Republic	0.6764	(0.6789, 0.6813)	(22)	0.338	0.358	0.381	0.258
Estonia	0.6715	(0.6740, 0.6766)	(23)	0.514	0.338	0.252	0.255
Spain	0.5777	(0.5799, 0.5822)	(24)	0.297	0.427	0.318	0.677
Malta	0.5415	(0.5440, 0.5465)	(25)	0.422	0.329	0.725	0.407
Portugal	0.4965	(0.4987, 0.5008)	(26)	0.442	0.328	0.448	0.812
Hungary	0.4611	(0.4630, 0.4648)	(27)	0.890	0.757	0.334	0.134
Lithuania	0.3450	(0.3464, 0.3478)	(28)	0.652	0.644	0.639	0.632
Latvia	0.0435	(0.0437, 0.0438)	(29)	0.936	0.942	0.966	0.825

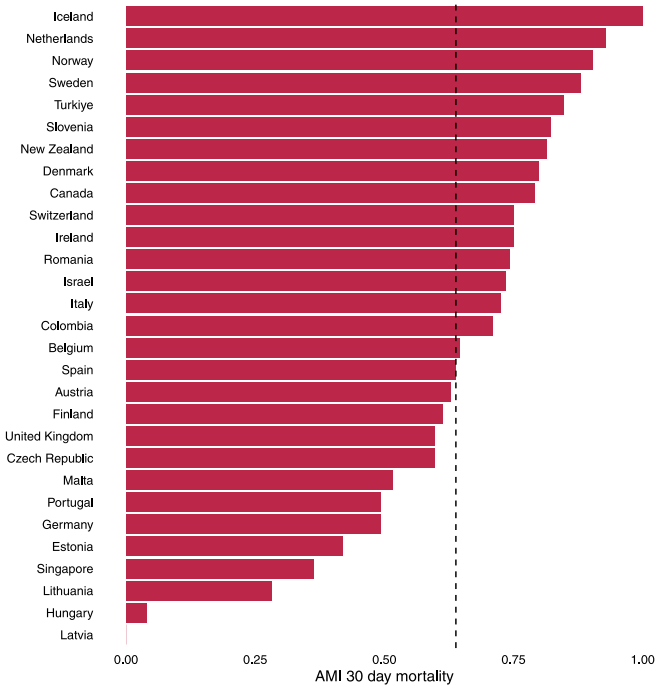


Fig. A.2. AMI 30 day mortality, normalised and polarised indicator.

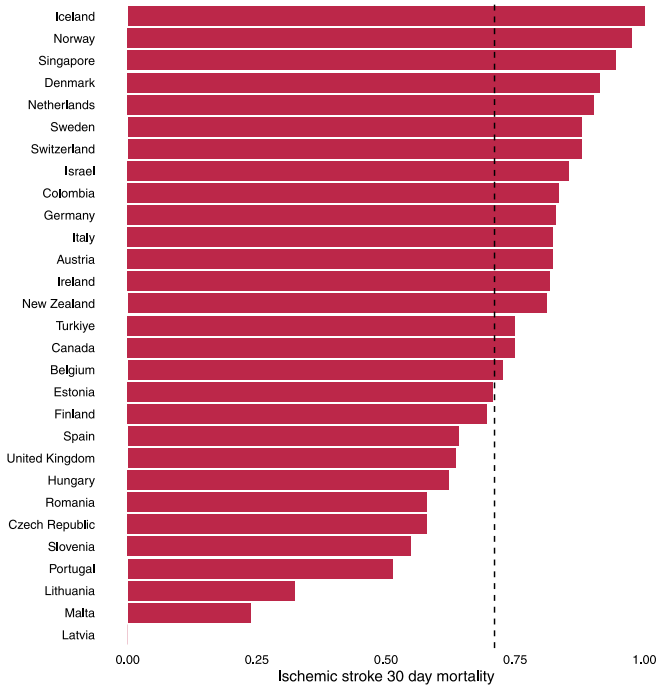


Fig. A.3. Ischaemic stroke 30 day mortality, normalised and polarised indicator.

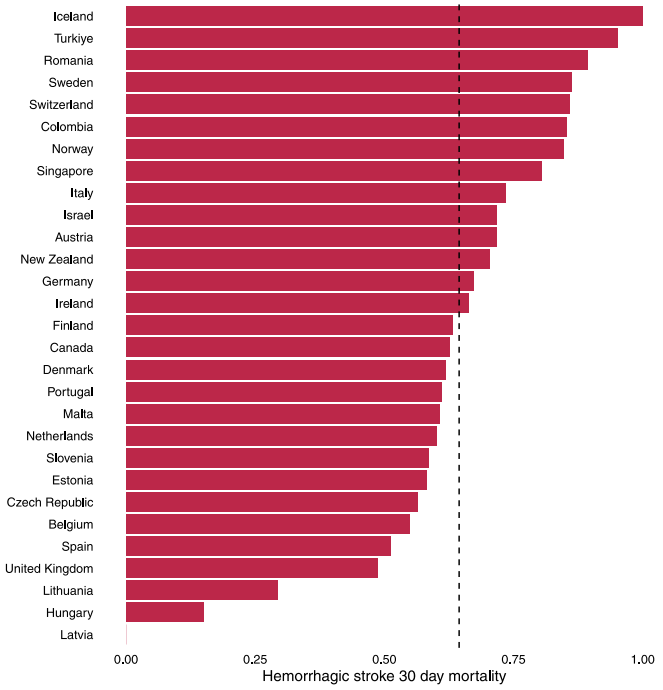


Fig. A.4. Hemorrhagic stroke 30 day mortality, normalised and polarised indicator.

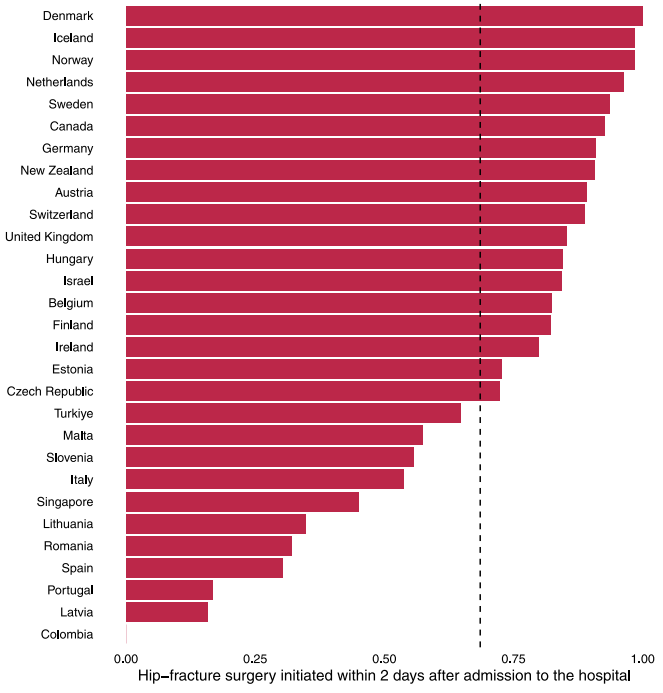


Fig. A.5. Hip-fracture surgery initiated within 2 days after admission to the hospital.

References

[1] Beaussier A, Demeritt D, Griffiths A, Rothstein H. Steering by their own lights: Why regulators across europe use different indicators to measure healthcare quality. Health Policy 2020;124:501–10.

[2] Smith P. Measuring Up: Improving Health System Performance in OECD Countries. OECD Publishing; 2002, p. 295–316, Ch. Developing Composite Indicators for Assessing Health System Efficiency.

[3] Jacobs R, Martin S, Goddard M, Gravelle H, Smith P. Exploring the determinants of NHS performance ratings: lessons for performance assessment systems. J. Health Serv. Res. Policy 2006;11(4):211–7.

[4] Donabedian A. Evaluating the quality of medical care. Milbank Memorial Fund Q. 1966;44(3):166–206.

[5] Kara P, Valentin JB, Mainz J, Johnsen SP. Composite measures of quality of health care: Evidence mapping of methodology and reporting. Plos One 2022;17:1–21.

[6] Jacobs R, Goddard M, Smith P. How robust are hospital ranks based on composite performance measures? Med. Care 2005;43:1177–84.

[7] Jacobs R, Goddard M. How do performance indicators add up? An examination of composite indicators in public services. Public Money & Manage. 2007;27(2):103–10.

[8] Shwartz M, Restuccia J, Rosen A. Composite measures of health care provider performance: A description of approaches. Milbank Q. 2015;93:788–825.

[9] Barclay M, Dixon-Woods M, Lyratzopoulos G. The problem with composite indicators. BMJ Qual. Saf. 2019;28(4):338–44.

[10] Friebel R, Steventon A. Composite measures of healthcare quality: sensible in theory, problematic in practice. BMJ Qual. Saf. 2019;28(2):85–8.

[11] Hofstede SN, Ceyisakar IE, Lingsma HF, Kringos DS, van de Mheen PJM. Ranking hospitals: do we gain reliability by using composite rather than individual indicators? BMJ Qual. Saf. 2019;28(2):94–102.

[12] Carinci F, Van Gool K, Mainz J, Veillard J, Pichora E, Januel J-M, Arispe I, Kim S, Klazinga N. Towards actionable international comparisons of health system performance: Expert revision of the OECD framework and quality indicators. Int. J. Qual. Health Care 2015;27.

[13] Fusco E. Potential improvements approach in composite indicators construction: The multi-directional benefit of the doubt model. Socio-Econ. Plan. Sci. 2023;85:101447.

[14] Nardo M, Saisana M, Saltelli A, Tarantola S, Hoffman A, Giovannini E. Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD Publishing; 2008.

[15] Matos R, Ferreira D, Pedro MI. Economic analysis of portuguese public hospitals through the construction of quality, efficiency, access, and financial related composite indicators. Soc. Indic. Res. 2021;157:361–92.

[16] Pereira MA, Camanho AS, Marques RC, Figueira JR. The convergence of the world health organization member states regarding the united nations' sustainable development goal 'good health and well-being'. Omega 2021;104:102495.

[17] Jacobs R, Smith P, Goddard M. Measuring Performance: An Examination of Composite Performance Indicators, Center for Health Economics. University of York; 2004, CHE Technical Paper Series, n.29.

[18] Cherchye L, Lovell K, Moesen W, Puyenbroeck TV. One market, one number? a composite indicator assessment of eu internal market dynamics. Tech. Rep., Working paper series ces0513, Katholieke Universiteit Leuven, Centrum voor Economische Studien; 2005.

[19] Cherchye L, Moesen W, Rogge N, Puyenbroeck T. An introduction to 'benefit of the doubt' composite indicators. Soc. Indic. Res. 2007;82(1):111–45.

[20] Zhou P, Ang BW, Zhou DQ. Weighting and aggregation in composite indicator construction: A multiplicative optimization approach. Soc. Indic. Res. 2010;96:169–81.

[21] Zanella A, Camanho AS, Dias TG. Undesirable outputs and weighting schemes in composite indicators based on data envelopment analysis. European J Oper Res 2015;245(2):517–30.

[22] Karagiannis G. On aggregate composite indicators. J Oper Res Soc 2017;68(7):741–6.

[23] Van Puyenbroeck T, Rogge N. Geometric mean quantity index numbers with benefit-of-the-doubt weights. European J Oper Res 2017;256(3):1004–14.

[24] Rogge N. On aggregating benefit of the doubt composite indicators. European J Oper Res 2018;264(1):364–9.

[25] Rogge N. Composite indicators as generalized benefit-of-the-doubt weighted averages. European J Oper Res 2018;267(1):381–92.

[26] Verbunt P, Rogge N. Geometric composite indicators with compromise benefit-of-the-doubt weights. European J Oper Res 2018;264(1):388–401.

[27] Aparicio J, Kapelko M. Enhancing the Measurement of Composite Indicators of Corporate Social Performance. Soc. Indic. Res. Int. Interdiscip. J. Qual. Life Meas. 2019;144(2):807–26.

[28] Ferreira DC, Caldas P, Varela M, Marques RC. A geometric aggregation of performance indicators considering regulatory constraints: An application to the urban solid waste management. Expert Syst Appl 2023;218:119540.

[29] Fusco E. Enhancing non-compensatory composite indicators: A directional proposal. European J Oper Res 2015;242(2):620–30.

[30] D'Inverno G, De Witte K. Service level provision in municipalities: A flexible directional distance composite indicator. European J Oper Res 2020;286(3):1129–41.

[31] Pereira MA, Camanho AS, Figueira JR, Marques RC. Incorporating preference information in a range directional composite indicator: The case of Portuguese public hospitals. European J Oper Res 2021;294(2):633–50.

[32] Fare R, Karagiannis G, Hasannasab M, Margaritis D. A benefit-of-the-doubt model with reverse indicators. European J Oper Res 2019;278(2):394–400.

[33] Mahdiloo M, Andargoli AE, Toloo M, Harvie C, Duong T-T. Measuring the digital divide: A modified benefit-of-the-doubt approach. Knowl-Based Syst 2023;261.

[34] Oliveira R, Zanella A, Camanho AS. 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. J Clean Prod 2020;267.
[35] Pereira MA, Marques RC. The 'sustainable public health index': What if public health and sustainable development are compatible? World Dev 2022;149:105708.
[36] Sahoo BK, Singh R, Mishra B, Sankaran K. Research productivity in management schools of India during 1968–2015: A directional benefit-of-doubt model analysis. Omega 2017;66(A):118–39.
[37] Mergoni A, D'Inverno G, Carosi L. A composite indicator for measuring the environmental performance of water, wastewater, and solid waste utilities. Util. Policy 2022;74:101285.
[38] Oliver A. The folly of cross-country ranking exercises. Health Econ. Policy Law 2012;7:7–15.
[39] Street A, Smith P. How can we make valid and useful comparisons of different health care systems? Health Serv. Res. 2021;56:1299–301.
[40] Vidoli F, Mazziotta C. Robust weighted composite indicators by means of frontier methods with an application to European infrastructure endowment. Italian J. Appl. Stat. 2013;23(2):259–82.
[41] Vidoli F, Fusco E, Mazziotta C. Non-compensability in composite indicators: A robust directional frontier method. Soc. Indic. Res. 2015;122(3):635–52.
[42] Fusco E, Vidoli F, Rogge N. Spatial directional robust benefit of the doubt approach in presence of undesirable output: An application to Italian waste sector. Omega 2020;94:102053.
[43] Bogetoft P, Hougaard JL. Efficiency evaluations based on potential (non-proportional) improvements. J. Prod. Anal. 1999;12(3):233–47.
[44] Arah O, Westert G, Hurst J, Klazinga N. A conceptual framework for the OECD health care quality indicators project. Int. J. Qual. Health Care 2006;18(Suppl 1):5–13.
[45] Rogge N, De Jaeger S, Lavigne C. Waste performance of NUTS 2-regions in the EU: A conditional directional distance benefit-of-the-doubt model. Ecol Econom 2017;139:19–32.
[46] Shen Y, Hermans E, Brijs T, Wets G. Data envelopment analysis for composite indicators: A multiple layer model. Soc. Indic. Res. 2013;114(2):739–56.
[47] Tofallis C. On constructing a composite indicator with multiplicative aggregation and the avoidance of zero weights in DEA. J Oper Res Soc 2014;65:791–2.

Francesco Vidoli is Senior Lecturer of Economic Statistics at the University of Urbino, Italy. He holds a Ph.D. in Economics, Mathematics and Statistics for Social Phenomena at the University of Rome, La Sapienza. His research focuses on the role of spatial heterogeneity in influencing firms' strategies and productive efficiency. He has published numerous papers in the fields of operational research and regional science, including Regional Science Policy; Practice, Journal of Regional Science, Local Government Studies, Regional Science and Urban Economics, European Journal of Operational Research and Omega.

Elisa Fusco is Senior Lecturer of Economic Statistics at the University of Florence; she has been statistical models expert in SOGEI S.p.A. at the "Department of Economic modelling and statistical analysis for policy making". and subject expert in Economic statistics for the course "Efficiency and productivity analysis" at the University of Rome La Sapienza. Ph.D. in Methodological Statistics, University of Rome La Sapienza. Her research interests include efficiency and productivity analysis, spatial econometrics and methods of composite Indicators construction through fronter models.

Giacomo Pignataro is Professor of Public Finance at the University of Catania and at Politecnico di Milano, Italy. He holds a Ph.D. in Economics at the University of York. His research interests are in the fields of health economics, especially in the analysis of the efficiency of healthcare organizations, and of cultural economics, with a focus on the efficiency of museums and cultural institutions. He published several articles on these topics on referred journals, including Social Science and Medicine, Socio-Economic Planning Sciences, Health Economics, Health Policy, The European Journal of Political Economy, FinanzArchiv, Journal of Policy Modeling and others.

Calogero Guccio is Professor of Public Finance at the University of Catania, Italy. He holds a Ph.D. in Health Economics at the University of Catania. His research interests include theoretical and empirical aspects of the evaluation of efficiency of private and public subjects. His research output has been published in leading outlets in the field of health economics and public economics.