Health Economics Review

RESEARCH

Open Access

One for all? Assessing the quality of Italian hospital care with the "benefit of the doubt" composite indicator methods

Francesco Vidoli[1], Giacomo Pignataro[2,3*], Sebastiano Battiato[4], Francesco Guarnera[4] and Calogero Guccio[2]

Abstract

Quality assessment in healthcare systems is challenging due to the multidimensional nature of healthcare services. This study evaluates the overall quality provided by hospitals using composite indicators under the Benefit of the Doubt (BoD) approach, which determines the weights of the indicators with minimal assumptions. We used data from 2015-2020 for Italian Local Health Authorities (LHAs) for 21 outcome measures, applying various non-parametric methods to address aggregation and weighting challenges. Our results show that the BoD measures are robust and effectively capture the dynamics of the quality of LHA, even during external shocks such as the COVID-19 pandemic. This research highlights the importance of methodological choices in the construction of composite indicators and demonstrates the effectiveness of the BoD approach in providing a comprehensive measure of healthcare quality.

Keywords Quality of healthcare, Composite indicator, Benefit of the doubt, Italian NHS

JEL: C14, C43, C44, I18

Introduction

Traditionally, healthcare quality has been measured through numerous individual indicators that capture its multifaceted nature[1]. These indicators address various clinical areas and services, reflecting the complex and multidimensional nature of healthcare quality, as conceptualised by Donabedian's tripartite framework of structure, process, and outcome [13]).

Despite the extensive use of single indicators, their limitations are evident when a comprehensive assessment of healthcare quality is needed. Single indicators often fail to provide a comprehensive view and may suffer from reliability issues due to a low number of observations for specific treatments or conditions and/or for

*Correspondence:
Giacomo Pignataro
giacomo.pignataro@unict.it
[1] University of Urbino Carlo Bo, Department of Economics, Society and Politics, Urbino, Italy
[2] University of Catania, Department of Economics and Business, Catania, Italy
[3] Politecnico di Milano, Department of Management, Economics and Industrial Engineering, Milano, Italy
[4] University of Catania, Department of Mathematics and Computer Science, Catania, Italy

[1] Recently, Beaussier et al. [1] surveyed 1,100 indicators, just looking at the work carried out by hospital regulators in four countries (France, England, Germany and The Netherlands).

specific providers. Consequently, composite indicators have been developed to compile multiple indicators into a single comprehensive measure[2]. This approach has been adopted by various international and national institutions, including the World Health Organization[3] and the US Centers for Medicare and Medicaid Services[4]. Efforts to construct a composite measure have faced at least two fundamental challenges: the aggregation of individual indicators into a composite measure and their weighting. Solutions to these issues are crucial for the reliability of the information provided by composite scores about the "real" differences in overall underlying quality of care among different providers. Aggregation of the scores of individual constituent indicators in a composite measure mainly involves assumptions about the compensability of performance between different indicators. Regarding weighting, different methods for assigning weights to individual indicators may correspond to various theoretical constructs or exogenously defined priorities. These methods significantly impact the computed composite scores. In general, the use of composite indicators has shown a limited discussion on the implications of methodological choices for aggregating and weighting individual indicators, leaving a significant degree of discretion in making these choices.

In this study, we apply the *Benefit of the Doubt* (BoD) approach to evaluate the quality of hospital care in Italy using composite indicators. The BoD method allows for the endogenous determination of indicator weights, minimising assumptions, and enhancing robustness. We used data for Italian Local Health Authorities (LHAs) in 2015-2020, covering several outcome measures. The data are collected by the National Agency for Regional Healthcare Services (AGENAS), which is a public entity of the Italian National Health Service, conducting research and providing support to the Minister of Health and the Regions. Data collection is carried out within the programme aimed at measuring the outcomes of different treatments (Programma Nazionale Esiti - PNE) of Italian hospitals. To the best of our knowledge, this research is novel in its application of the BoD family of methods to healthcare quality assessment, specifically focussing on the controversial steps of aggregation and weighting steps in the construction of composite indicators[5].

The practice of aggregating individual indicators by summing their weighted scores, assuming perfect compensability of performance[6], has significant implications. Since individual indicators can refer to different treatments for different patients, the assumption that underperformance in some areas can be compensated by overperformance in others suggests that health losses (or potential health gains) in underperforming treatments are balanced by gains in overperforming care. This is a matter of social evaluation and one cannot simply assume that compensability is always socially acceptable; it must be addressed explicitly. In this paper, we compare two methods - BoD and Robust BoD (RBoD)- which assume perfect compensability, with another method, the Mazziotta-Pareto Index (MPI), which introduces a penalty for unbalanced indicator values. This comparison allows us to illustrate the impact of the perfect compensability assumption on the overall measure of healthcare quality.

Regarding weighting, whatever method is used, the weights should be considered as "value judgments" about the relative importance of the individual constituent indicators. As Jacobs et al. [20] note "*the weights used reflect a single set of preferences, whilst the evidence suggests there exists a great diversity in preferences across policymakers, individual unit actors, and the broader public. There is likely to be considerable variation in the preferences of the respondents*". The methodologies used in this paper are derived from non-parametric frontier analysis, specifically Data Envelopment Analysis (DEA), and they rely on the endogenous determination of weights for each unit under examination. The BoD approach, in particular, allows for the consideration of varying weights for different units. This approach derives the weights for each provider from the assumption that each unit, within its unique context, makes allocation decisions to optimise its performance concerning the different objectives represented by the individual indicators.

Finally, we address a specific problem related to non-parametric frontier analysis within the BoD approach: the robustness of composite indicator scores in the presence of outliers. We use the RBoD method, which is based on the DEA order-*m* efficiency concept proposed by Cazals et al. [6] (for a revised version of this approach to

[2] For a recent survey of composite indicators, developed for measuring the quality of healthcare, see Kara et al. [21].

[3] https://www.who.int/publications/i/item/924156198X; last accessed on 28 August 2023.

[4] https://www.cms.gov/medicare/quality-initiatives-patient-assessment-instruments/hospitalqualityinits/hospitalcompare; last accessed on 28 August 2023.

[5] A notable exception is the work of Lagravinese et al. [23], which employ the Stochastic Multicriteria Acceptability Analysis (SMAA) approach based on a methodological proposal of Greco et al. [18], to compose indicators relative to 17 mortality rates, calculated for Italian regions. However, our work differs markedly, not only for the methodological approach but also for the greater generality of the measure of quality of healthcare care derived in our analysis.

[6] Of course, compensation between the scores of the individual indicators is realised on the basis of a trade-off equal to their relative weights.

composite indicators, see Vidoli et al. [34]). By comparing it with the conventional BoD approach, we demonstrate the discrepancies.

Our findings show that the composite indicators used in this study are robust and effectively capture the dynamics of LHA quality over time, even in the face of external shocks such as the COVID-19 pandemic. The application of these methodologies overcomes critical aspects related to aggregation and weighting, such as the influence of outliers and the need for compensability assumptions. The RBoD method, in particular, ensures that the results remain reliable and accurate by mitigating the impact of outliers. This research contributes to the ongoing development of composite indicators, offering a more nuanced and comprehensive tool to evaluate healthcare quality.

The following sections detail the methodologies used (The methodological approaches section), the data and their application (The data and the application of the methods section), the results of the composite scores (The results section), and the concluding comments (Final remarks section).

The methodological approaches

Composite indicators represent a structured approach to aggregating multiple variables or indicators into a single, more manageable metric. They offer a unique perspective by synthesising diverse data sources and dimensions into a unified assessment. These indicators have been extensively used across economics, environmental sciences, public policy, healthcare, and other fields, enabling decision makers to derive actionable insights from the data. Methodological research in this field in recent years has therefore increasingly sought to move away from subjective and controversial methods, by attempting to propose methods using endogenous criteria to the elementary data, in order to increase the reliability of the composite indicator itself.

The model known as BoD, which was initially presented by Cherchye et al. [8] and earlier by Melyn and Moesen [26] and Karagiannis and Sarris [22], is becoming more widely recognised as a contemporary approach that uses linear optimisation methods to determine the weights of elementary indicators within the model itself. Several adaptations and modifications of the original BoD model have been suggested, often influenced by advances in the DEA literature.

To be more specific, the BoD method can be regarded as a distinct variation of the nonparametric DEA model that exclusively focusses on evaluating "achievements", namely outputs, without going into the details of input factors. From a technical point of view, the BoD model formally aligns with the original input-oriented DEA model introduced by Charnes et al. [7]. Within this framework, all subindicators are treated as output, while input is treated as a constant dummy variable, uniformly set to one for all observations[7]. Formally, for the BoD-setting, the production possibility set Ψ where to search for the relative maximum for each unit can be defined as:

$$\Psi = \{(\mathbf{1}, y) \in \mathbb{R}_{+}^{1+q} | X \equiv \mathbf{1}, Y_j \geq y\}. \quad (1)$$

where q represents the number of basic indicators $y = \{y_1, ..., y_q\}$ collected for $i = 1, ..., n$ observations to be aggregated and one input assumed to be equal to **1** for all units. Each of the q indicators measures a specific aspect of the latent measure under assessment.

Having defined the set Ψ, it is possible to calculate the Farrel-Debreu (output) efficiency score λ as:

$$\lambda(\mathbf{1}, \boldsymbol{y}) = \sup \left\{\lambda > 0 | H(\mathbf{1}, \lambda \boldsymbol{y}) > 0\right\} \quad (2)$$

where H represents the probability function for a unit to be dominated ([10], p. 66). Consequently, the BoD Composite Indicator (CI) can be computed as the reciprocal of $\lambda(\mathbf{1}, \boldsymbol{y})$.

The basic principle of the BoD model is grounded in the recognition that, during the estimation of CI scores, exact information about the correct importance weights for performance indicators is often missing or only known to some extent. To address this issue, the BoD model uses a method where it internally derives the indicator weights from the observed data of the indicators, placing the Decision Making Units (DMUs) in a comparative context and systematically examining them against each other.

However, in its basic version, the BoD model presents several limitations that can limit its practical usefulness and have led several authors to propose specific versions[8]. The main issues dealt with in the various extensions of the BoD model are: *(i)* the lack of robustness of the performance estimates (see e.g. [34]); *(ii)* the perfect compensability among indicators (see e.g. [15, 16, 24, 25, 35]); *(iii)* the inability to treat undesirable indicators (see e.g. [14]); *(iv)* the lack of account and/or correction for background conditions (see e.g. [12] for conditional and [17] for the spatial proposal) and *(v)* the assumption of a linear aggregative model (see e.g. [29–33] nonlinear proposals).

In this specific application, our main focus is on addressing the robustness and the non-compensability issues. As the BoD method operates deterministically, it assumes that

[7] For a deeper understanding of this conceptual framework, readers are encouraged to explore the work of Cherchye et al. [8], providing comprehensive details.

[8] See Greco et al. [19] for a more detailed survey.

the estimated composite performance scores fully reflect the actual performance of DMUs. This assumption does not consider the potential presence of noise or anomalies within the indicator data. However, the mere existence of one unit with outlier or atypical performance data within the sample can significantly lower the composite performance scores of all other units. The BoD production set Ψ can be transformed into a probabilistic framework by using Daraio and Simar [9] proposal; considering a sample of m random variables with replacement $S_m = \{\boldsymbol{Y_i}\}_{i=1}^{m}$ drawn from the density of $\boldsymbol{Y}$, we define the random set $\widetilde{\Psi}_m$ as follows:

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

In this setting, therefore, m represents the number of DMUs dominating the unit under analysis and representing a set of "*potential competing firms*" ([9], p. 71) drawn from the population. In other terms, the single unit is not compared each time with the entire set of units dominating it but with a subset of order m; in this way the effect of any anomalous (super-efficient) units is dampened by the random extraction of a subset of units.

The model is, therefore, implemented through an iterative computation of the sample subset of size m (for $b = 1, ..., B$ times). Estimation of the RBoD measure [34] is represented by the empirical mean over B. Since the benchmarking of each DMU is carried out, not with the entire set of DMUs, but within a smaller subset of size m, the influence of outlier units is diminished.

Regarding the non-compensability issue, we use a non-compensatory method known as the Mazziotta-Pareto index (MPI, [25]). Under the assumption that each component is not substitutable with the others (or only partially so), the simple arithmetic mean (M) is penalised by the coefficient of variation (cv) between the individual indicators (the 'horizontal' variability), in order to reveal units with unbalanced indicator, as in the next equation:

$$MPI_i^{\pm} = M_i(1 \pm cv_i^2) \quad (4)$$

where the symbol ± represents the polarity of the composite indicator.

The MPI's ability to provide a balanced, transparent, and sensitive measure of performance across multiple dimensions makes it a valuable tool in the construction and application of composite indicators. In fact, its non-compensatory nature is crucial in contexts where balanced performance across multiple dimensions is essential, and, considering both the mean and the dispersion of the indicators, helps in identifying not only the average performance, but also the equity of the performance across different dimensions.

Finally, all of the methods present certain inherent limitations for cross-period comparisons; because they are based on comparisons between units in the same set, they are relative measures that do not take into account the effects of changes in production technology and are therefore useful for making comparisons over limited periods of time. In particular, it is assumed that the production frontier against which the units are compared remains stable over time and that changes in the DMUs are solely due to improved efficiency of the individual DMUs. If a co-comparison over a long period of time is required, some extensions to the BoD methods [2, 27] using the Malmquist index and to the MPI method [25] have been proposed.

The data and the application of the methods

The data

Since 2012, the Italian public agency AGENAS has been monitoring various performance indicators, primarily related to health outcomes and admission volumes, for 1,377 accredited public and private hospitals in Italy through the PNE programme. In 2020, the final year of our observation period, AGENAS calculated 164 indicators related to hospital care, 71 of which were specifically related to care outcomes. Data for each hospital and LHA[9] are publicly available on a dedicated website[10].

Given the large number of indicators in the PNE programme, using all of them could bias composite indicator scores, especially under the BoD method, by assigning high weights to the few indicators where a hospital or LHA excels. The risk of this bias potentially increases with the number of indicators used to calculate the composite indicators. Therefore, our analysis focusses on a subset of outcome indicators for selected clinical areas from 2015 to 2020[11]. The selection of individual indicators for the composite computation is crucial. To avoid bias from discretionary choices, we rely on AGENAS's selection of clinical areas and specific indicators from their overall set, chosen by the agency to provide an aggregate representation of hospital performance[12]. Although we recognise that any selection of indicators can affect the values of the

[9] LHAs are responsible for the organisation and the provision of healthcare services to the population resident in their geographical area. For a more detailed and updated report on the organisation and governance of the Italian healthcare system, see De Belvis et al. [11].

[10] https://pne.agenas.it/; last accessed on 28 August 2023.

[11] We downloaded the relevant data from the time series available on the agency's website.

[12] AGENAS has not created a composite indicator to represent the aggregate performance of hospitals over this subset of indicators. Instead, it uses TREEMAP, a graphical tool for displaying large amounts of hierarchically structured data. The latest downloadable illustration of the TREEMAP methodology, as used by AGENAS, can be found in a report downloaded at https://pne.agenas.it/sintesi/sintesi_vis/croc/Treemap_metodi_2021.pdf; last accessed on 28 August 2023.

composite score, including the selection by AGENAS, our measures of overall quality will at least reflect the public agency's value judgement. This judgement is based primarily on the representativeness of the indicators selected within each clinical area.

After checking for missing data and outliers, the clinical areas and indicators used in this paper[13] are:

- Cardiovascular (6 indicators): 30-day Acute Myocardial Infarction (AMI) mortality, percentage of AMI patients treated with Percutaneous Transluminal Coronary Angioplasty (PTCA) within 2 days, 30-day mortality after isolated Coronary Artery Bypass Graft surgery (CABG), 30-day mortality after congestive heart failure, 30-day mortality after valvuloplasty, and 30-day mortality after repair of intact abdominal aortic aneurysm.
- General Surgery (2 indicators): 30-day complications after laparoscopic cholecystectomy and the percentage of admissions with a post-operative length of stay (LOS) of less than 3 days following cholecystectomy.
- Surgical Oncology (5 indicators): 330-day mortality for lung cancer surgery, 30-day mortality for gastric cancer surgery, 30-day mortality for colorectal cancer surgery, 120-day re-operation rate after breast-conserving surgery, and 90-day re-operation rate after breast-conserving surgery.
- Pregnancy and Delivery (3 indicators): primary C-section rate, readmissions after vaginal delivery, and readmissions after cesarean delivery.
- Neurology (2): 30-day mortality after ischaemic stroke and 30-day mortality after craniotomy for a brain tumor.
- Musculoskeletal (2 indicators): femoral neck fracture repair within 2 days and 30-day re-admissions after hip surgery.
- Respiratory (1 indicator): 30-day mortality after Chronic Obstructive Pulmonary Disease (COPD).

For each of these indicators, AGENAS generally provides the raw value of the indicator and, for indicators related to the outcome of healthcare, the risk-adjusted value. We used risk-adjusted values for all selected indicators.

We estimate composite indicators at the LHA level. We chose to measure composite performance at the LHA level instead of the hospital level due to a significant lack of data for several indicators in many hospitals. This was either because of their specialisation, meaning that they did not treat cases relevant to specific indicators, or because they treated too few cases for those indicators. As a result, these hospitals were not assigned a risk-adjusted value for certain indicators, as the risk-adjustment procedure would not have provided a reliable estimate with insufficient cases. However, measurement of composite indicators at the LHA level allows measurement of the quality of the overall service provided by the different structures of the NHS in a given geographical area. We used an unbalanced panel of LHA data for the period 2015-2020, as the AGENAS website does not provide values for some regions' LHAs for the first two years. Although there are 119 observation units for 2017-2020, data are available for only 80 units in 2015 and 98 units in 2016[14].

Finally, because individual indicators use different units of measurement, their values need to be normalised so that they all fall within the same range. Additionally, we need to address polarisation, which means determining whether higher or lower values indicate good or bad performance (e.g., higher mortality rates are bad, but higher percentages of timely treatment are good). Therefore, normalised scores must be adjusted so that their variations represent the same performance implications for all indicators. To address these issues of normalisation and polarisation, we use the scores normalised and polarised by AGENAS for the subset of indicators in their TREEMAP exercise. These scores range from 1 to 5, as illustrated for three cardiovascular indicators in Table 1 (for a complete representation of all TREEMAP scores,

[13] We chose to use the TREEMAP selection of indicators from 2015, the first year of our observation period. Although the 2015 TREEMAP report is no longer available on the AGENAS site, it can still be downloaded from other websites of healthcare institutions (e.g., https://www.asst-mantova.it/documents/338413/1498693/8431.pdf/8cc49e32-2238-3e93-3b58-bead0cf0960f; the site was last accessed on 28 August 2023). The set of indicators used in this paper differs slightly from those used by AGENAS due to data availability issues for certain indicators. Consequently, we make substitutions with indicators from the same clinical area that are as close as possible to the specific outcomes measured by the original indicators. Specifically, we substituted: in general surgery,% of cases of laparoscopic cholecystectomy, in wards with a number of cases > 90, with 30-day complications after laparoscopic cholecystectomy; in surgical oncology,% of breast cancer surgery, in wards with a number of cases > 120, with 90-day reoperation rate after breast preservation surgery; in pregnancy and delivery, proportion of complications during pregnancy or delivery, among women with vaginal delivery, with readmissions after vaginal delivery; in pregnancy and delivery, proportion of complications during pregnancy or delivery, among women with c-section delivery, with readmissions after c-section delivery; in musculoskeletal, waiting time for surgery for tibia and fibula fracture, with 30-day readmissions after hip surgery.

[14] For some regions (Friuli, Marche, Molise, Toscana, and Sardegna), AGENAS provides the values of individual indicators at the provincial level. This is because LHAs in these regions are considered too large (in some cases covering the entire regional area) to provide a performance measure at a sufficiently refined subregional level. For example, in 2019, these five regions had a total of 9 LHAs, but in the AGENAS dataset, they are represented by 29 provinces.

Table 1 Normalisation and polarisation of simple indicators - an extract

Clinical area	Indicator	Very High 1	High 2	Medium 3	Low 4	Very Low 5
Cardio - vascular	Acute myocardial infarction: 30-day mortality	$\leqslant 6$	$6 \dashv 8$	$8 \dashv 12$	$12 \dashv 14$	> 14
	Acute Myocardial Infarction: % treated with PTCA within 2 days	$\geqslant 60$	$45 \vdash 60$	$35 \vdash 45$	$25 \vdash 35$	< 25
	Congestive heart failure: 10% 30-day mortality	$\leqslant 6$	$6 \dashv 9$	$9 \dashv 14$	$14 \dashv 18$	> 18

Notes: The table reports the normalisation and polarisation related to three performance indicators of the cardiovascular clinical area employed by AGENAS in the TREEMAP exercise

Sources: Authors' elaborations on Agenas data

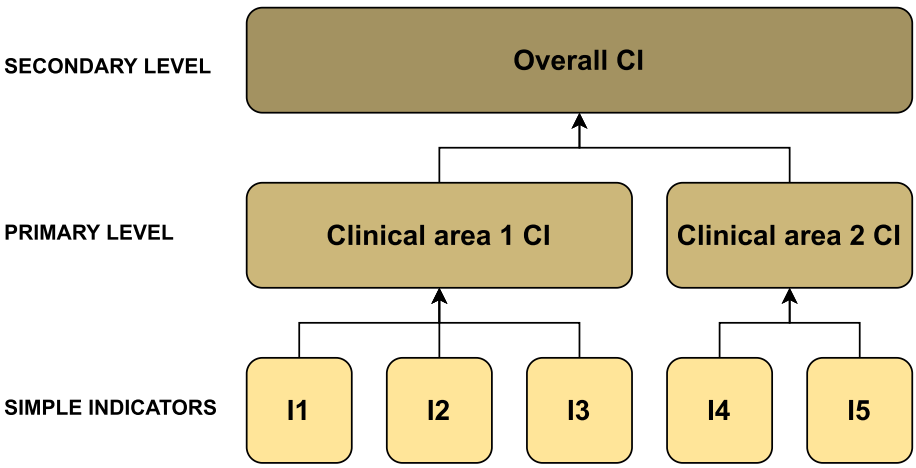

Fig. 1 Hierarchical levels of indicators. Notes: The figure shows the two levels of aggregation used in the analysis to compute the composite indicators. The primary level aggregates at the clinical area level while the secondary level aggregates from the clinical areas to the overall indicator. Sources: our elaboration

see Table 3 in Appendix A)[15]. In terms of polarisation, higher values indicate worse performance.

The application of the different methods for the estimation of composite indicators

We estimate composite indicators at the primary level[16], which are composite indicators for each clinical area, and composite indicators at the secondary level, which aggregate the primary level indicators to represent the overall quality of healthcare provision for any LHA. Figure 1 illustrates the hierarchical aggregation model used to build these composite indicators. The primary level is at the clinical area level, while the secondary level aggregates these clinical areas into an overall indicator.

Following the methods described in The methodological approaches section, we used the `BoD`, `RBoD`, and `MPI` methods. Specifically, for the `RBoD` method, we set $m = 30$ and $B = 100$. We also calculated what would be a composite indicator for AGENAS, following the practice that involves summing the values of individual indicators after weighting them. We did this by taking a simple weighted average of the selected indicators at both the primary and secondary levels, using the AGENAS weights applied in the 2015 TREEMAP exercise for each clinical area and indicator. These weights are reported in Table Appendix B in Appendix A. This allowed us to create a straightforward benchmark to compare the performance of LHAs, based on the importance AGENAS assigns to different indicators and clinical areas.

The results

In this Section, we present the results of our empirical exercise. Since multiple approaches are possible to calculate composite indicators at both the first and second levels, we adopt a data-driven approach. Thus, as previously

[15] For the indicators used in this paper that are not included in the set used by AGENAS, the values have been normalised using the quintiles of their distributions.

[16] The R Compind package was used; https://cran.r-project.org/web/packages/Compind/index.html, [17].

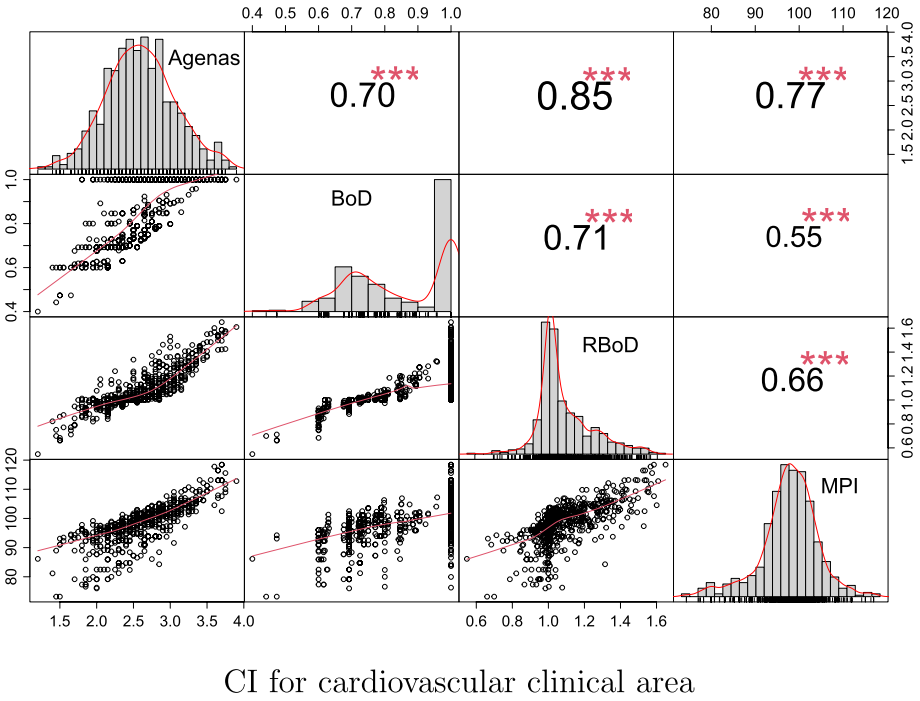

Fig. 2 Correlation among different methodologies. NOTES: The figure presents a correlogram matrix of the composite indicators for the cardiovascular clinical area computed using `Agenas`, `BoD`, `RBoD` and `MPI`. The left side displays the linearity of relationships between the scores of individual methods. The diagonal shows the distribution of individual composite indicators. Finally the right side illustrates the linear correlation between the composite indicators for the cardiovascular clinical area along with their respective statistical significance. SOURCES: Elaboration by the authors on Agenas data

mentioned, we use as a benchmark the indicator (`Agenas`) computed as a simple weighted average of the selected clinical indicators using the weights assigned by AGENAS in the 2015 TREEMAP exercise. Then, to select the methodology (i.e. `BoD`, `RBoD` and `MPI`) that appears more robust compared to our benchmark `Agenas`, we first compare the results of the different methodologies at Level 1 by measuring their Pearson correlation.

To save space, we report the correlation analysis only for the cardiovascular clinical area[17]. The results of the correlation between the composite indicators of level 1 for the cardiovascular clinical area, computed using the three methods (BoD, RBoD, and MPI), and our benchmark AGENAS are presented in Fig. 2.

More specifically, Fig. 2 reports on the left side of the correlogram matrix the linearity of the relationship between the scores of the individual methods; on the diagonal, the distribution of the individual composite indicators and, on the right side of the matrix, the linear correlation between the composite indicators with the relative statistical significance. The correlation scores are generally quite high and significant, particularly for those relative to `Agenas` and `RBoD` indicators. Furthermore, the relatively low correlation between `BoD` and `MPI`, could be due to the presence of outliers, probably exacerbates the implications of the difference in terms of the assumption of compensability between the performances of the individual indicators.

The correlation results are quite encouraging for the use of the methodologies employed in this paper, since the impact of the different assumptions underlying the computation of composite indicators is not so striking in terms of relative scores and ranking of the different units under examination (the LHAs). This is quite important for the issue related to the relevance of the compensability of performance between the different indicators. Instead of making an assumption on compensability, the measurement of correlation shows in this case that, in practice, it does not make such a difference.

Thus, in what follows, given the high correlation among the different indicators, in particular between `RBoD` and `Agenas`, we will focus on `RBoD` scores, because of its robustness to outliers[18].

The summary statistics for the different indicators aggregated at Level 1 and Level 2, and computed using

[17] The correlation results for the other clinical areas overlap with those reported here and are available from the authors on request.

[18] Robustness analyses have been carried out to check the sensitivity of the results to varying the aggregation method at the second level. The analyses showed a very good robustness of the results, thus favouring the choice of the method with the most desirable properties. These analyses are available from the authors on request.

the RBoD methodology are reported in Table 2. Appendix B provides the descriptive statistics by year for the BoD, MPI, and AGENAS Level 1 composite indicators. Since normalised values range from 1 to 5, with 1 being the best and 5 the worst, lower values of the composite indicator indicate better performance.

Table 2 provides a comprehensive overview of the composite indicator (CI) scores in different clinical areas and overall levels for the years 2015 to 2020, calculated using the RBoD methodology. The overall composite indicator at Level 2 remains fairly stable at around 1.1, with little change over the years, reflecting the general consistency of aggregate performance between clinical areas at the aggregate level. The cardiovascular domain shows a gradual improvement in performance, with the average CI score decreasing from 1.2 in 2015 to 1.0 in 2019. Neurology maintains a relatively constant average CI score around 0.7, with minor fluctuations and the lowest score observed at 0.67 in 2018, indicating general stability with some variation. The respiratory domain starts with an average CI score of 0.61 in 2015 but deteriorates to 0.69 in 2020, indicating a steady decline in this area. General surgery shows significant improvement, with the average CI score improving from 0.86 in 2015 to 0.68 in the following years. Surgical oncology maintains consistent performance with an average CI score of around 1.0 over the years, showing no significant changes or trends. The pregnancy sector also remains stable, with little variation, keeping the average CI score around 0.9. The musculoskeletal sector shows some variability, with mean scores fluctuating between 0.79 and 0.66 over the years.

Figure 3 represents the trend in the average quality of care of the different clinical areas, throughout the period of time considered in this paper, across the Italian LHAs. The results in Fig. 3 show a convergence trend in the quality of care provided in the different clinical areas, until 2019: areas where the quality was lower tend to improve, with the exception of pregnancy and delivery, while the respiratory area, where the performance was the best since the beginning of the observation period, tends to maintain its high quality. Even if it is outside the scope of this paper to explain the differences in quality between different services and over time, it can be mentioned that the emphasis on quality of care characterising the policy actions of different healthcare systems, including Italy, may have played a role in this trend. In 2020, the change in the values of the composite indicators shows a general deterioration in the quality of care, with the exception, again, of pregnancy and delivery and of general surgery. Since 2020 was characterised by the well-known shock represented by the Covid-19 pandemic, and we now know, from several studies, the disruption that this shock caused to the provision of healthcare services, the change in our composite indicator may well capture this disruptive effect. This explanation could also be in line with the performance observed, at least for pregnancy and delivery. In fact, it is difficult to think that the pandemic may have had an impact on the mode of primary delivery or on complications during pregnancy and delivery.

Our results for composite indicators can also shed light on a highly debated issue for Italy. The country has traditionally been characterised by a North-South divide in several fields, including the provision of healthcare services.

Figures 4 and 5 show how the quality of care in the different clinical areas has evolved in the main geographical areas of the country.

Table 2 Summary statistics - CI Level 1 and Level 2 - RBoD method

Clinical area	2015		2016		2017		2018		2019		2020	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Level 1												
Cardiovascular	1.2	0.2	1.1	0.18	1.1	0.16	1.1	0.15	1	0.13	1.1	0.14
Neurology	0.78	0.23	0.71	0.24	0.72	0.23	0.67	0.24	0.68	0.21	0.74	0.23
Respiratory	0.61	0.17	0.55	0.17	0.57	0.17	0.55	0.19	0.56	0.18	0.69	0.19
General surgery	0.86	0.24	0.78	0.25	0.71	0.24	0.68	0.26	0.68	0.24	0.68	0.25
Surgical Oncology	1	0.2	1	0.17	1	0.17	1	0.18	0.99	0.18	1	0.18
Pregnancy	0.93	0.21	0.93	0.18	0.91	0.2	0.95	0.14	0.94	0.16	0.84	0.18
Musculoskeletal	0.79	0.25	0.79	0.24	0.7	0.27	0.66	0.24	0.68	0.24	0.75	0.26
Level 2												
Overall CI	1.2	0.12	1.1	0.1	1.1	0.096	1.1	0.097	1.1	0.089	1.1	0.11

Notes: The table reports the descriptive statistics of the composite indicators computed using RBoD method in our sample of Italian LHA in the time span from 2015 to 2020 for both individual clinical areas (Level 1) and at the overall level (Level 2)

Sources: Authors' elaborations on Agenas data

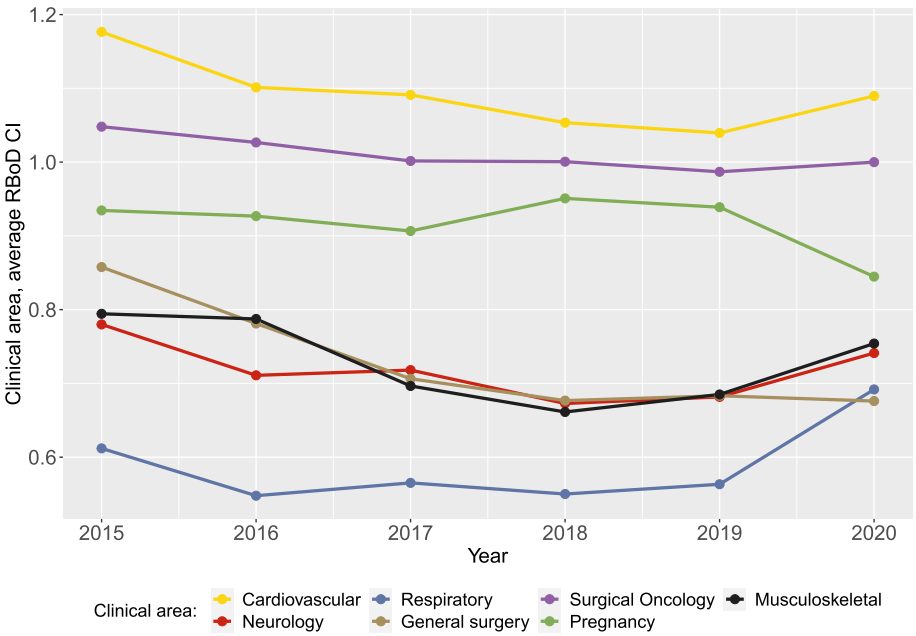

Fig. 3 Primary indicators trend. Notes: Figure reports the temporal dynamics of composite indicators computed at Level 1 with regard to the following clinical areas: Cardiovascular; General Surgery; Surgical Oncology; Pregnancy and Delivery; Neurology; Musculoskeletal; Respiratory. Sources: Elaboration by the authors on Agenas data

The time trend of the primary indicators for the different geographical areas of the country does not show uniform patterns for the different clinical areas, while the geographic trend of the overall quality of care is much clearer.

This can be observed in Fig. 6 which shows the overall composite indicator by geographic areas. More specifically, Fig. 6 reports the results for the mean values of the secondary level composite indicators, for the different geographical areas.

The North-South gap, at the beginning of the period, has been reduced at the end of the same period, even if it still remains substantial. Again, we can say that it is possible that the emphasis on quality in different policy initiatives may have had an impact, especially on LHAs located in the lagging-behind areas of the country. The comparison between time trends for primary and secondary composite indicators also shows the potential advantage of a more aggregate assessment, in terms of clearer information on relevant issues such as, in this case, the dynamics of the quality gap between Northern and Southern healthcare providers in Italy.

More information on the quality trend comes from the clustering of LHAs based on their secondary-level composite indicator trend over time, using Functional Data Analysis (FDA, [28])[19].

In the application proposed in this paper, the funFEM algorithm [5] has been used to cluster the distinct trends and model them as curves within a common and discriminative functional subspace.

We represent the FDA results in Fig. 7. Two clusters of LHAs have been identified, based on their quality behaviour over time: a first group of LHAs that worsen the overall quality of their care over time, even if starting with a better performance; a second one that improves the overall quality of their care over time, even if starting with a worse performance.

The baseline year to represent the performance trend is 2017 (since we missed some observations for the years 2015 and 2016), and the baseline value for the secondary level composite indicator is set equal to 100 for all LHAs in that year[20].

The estimated mean shows the overall trend of the two clusters and, as already observed, the average quality performance for all LHAs worsens in 2020. This clustering improves our understanding of the changes that have occurred in the disparity between quality of care in the North and South regions.

Next, we will examine how Italian Local Health Authorities (LHAs) are distributed between the two

[19] More detailed information on FDA can be found in Appendix C.

[20] When the performance of a given LHA improves, in the following years, because the score of our composite indicator decreases with better performance, the value will be below 100, otherwise it will be above 100.

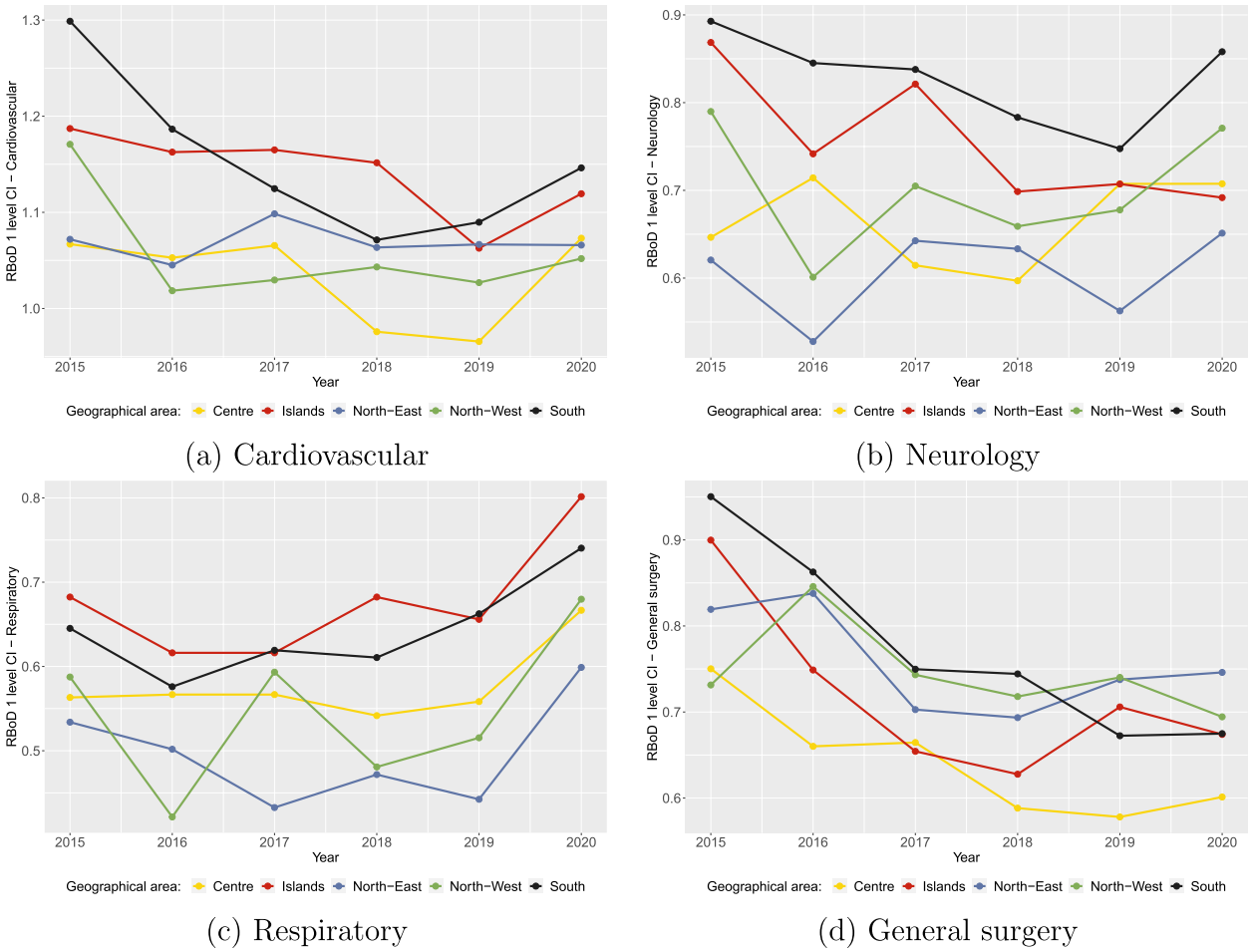

Fig. 4 Average RBoD primary level composite indicators in the main geographical areas

clusters in each region of the country. The distribution is shown in Fig. 8.

The narrowing of the quality gap between the North and the South of the country is due to the fact that the majority of LHAs in the north (more so in the North-East than in the North-West), which started with a better performance than LHAs in other areas of the country, experienced a decline in the overall quality of their services, while LHAs in the South and especially in the two islands (Sicily and Sardinia) showed the opposite behaviour. In conclusion, our findings prove that the composite indicators used in this study are quite robust and effectively capture the dynamics of LHA quality over time, even amidst external shocks such as the COVID-19 pandemic. The applied methodologies address critical issues related to aggregation and weighting, including the influence of outliers and the need for compensability assumptions. The Robust BoD (RBoD) method, in particular, ensures the reliability and accuracy of results by mitigating the impact of outliers. Furthermore, the analysis reveals a narrowing quality gap between the Northern and Southern Italian LHAs. In the next section, we provide concluding remarks on our findings and propose practical applications and extensions of the proposed methods.

Final remarks

The objective of this paper is to present an application of different methodologies for the estimation of composite indicators to the measurement of the quality of healthcare services. The application is related to hospital care provided by Italian LHAs in the period 2015-2020 and has been carried out using data on different quality indicators, provided by the Italian public agency AGENAS. We have compared three different methodologies - BoD, RBoD and MPI, which share the endogenous determination of the weights to be assigned to the single

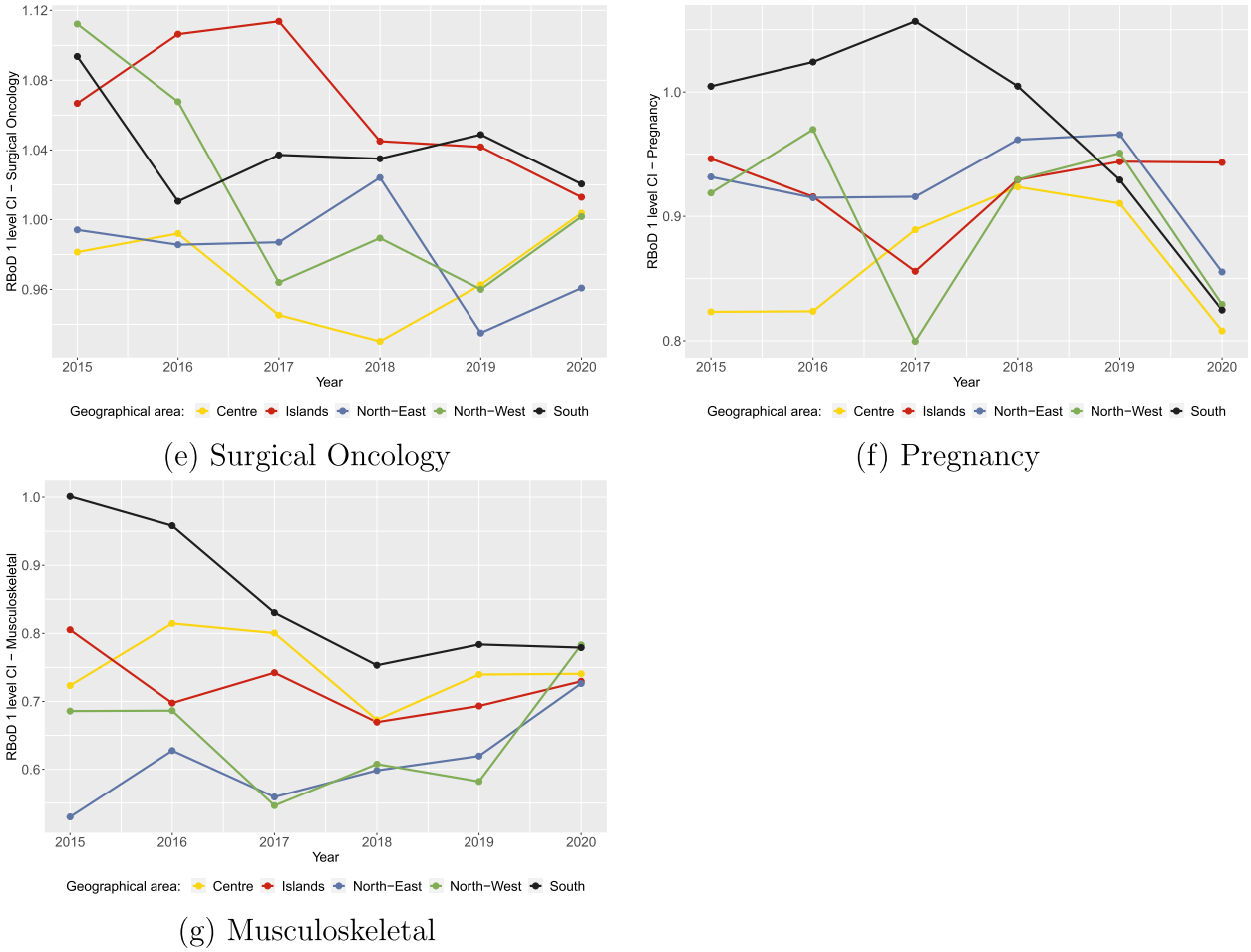

Fig. 5 Average RBoD primary level composite indicators in the main geographical areas. Notes: The figure depicts the dynamics of composite indicators at Level 1 across various geographical areas for the following categories: Cardiovascular (**a**); General Surgery (**b**); Surgical Oncology (**c**); Pregnancy and Delivery (**d**); Neurology (**e**); Musculoskeletal (**f**); and Respiratory (**g**). Sources: Elaboration by the authors on Agenas data

indicators, thus avoiding the need for their exogenous and potentially discriminating (across the different units) identification.

This application arises from the need in the literature and policy debate to comprehensively evaluate the performance of healthcare providers, organisations, and systems. Although individual indicators assess the performance of specific services or dimensions of a healthcare provider, composite indicators provide an integrated perspective on overall performance.

In this respect, we believe that the proposed approach can contribute from several points of view.

For example, healthcare administrators and policy makers could use the proposed approach in their toolkit to improve healthcare service delivery. In fact, the adoption of composite indicators provides a comprehensive understanding of performance in the various clinical areas, allowing quality disparities to be identified. This facilitates targeted interventions to strengthen weaker areas, thus improving the overall delivery of health services. In addition, the specific aspects of the Italian healthcare system, such as the regional disparities between North and South, provide further insight into how geographical and socioeconomic factors influence the quality of healthcare.

However, the policy implications arising from this study are in our opinion helpful not only for Italian NHS, but also for health administrators and policy makers in other countries. In this perspective, the methodologies used in this study are easily applicable and scalable to other health systems and globally. This adaptability ensures that the proposed approach can be used both in different contexts

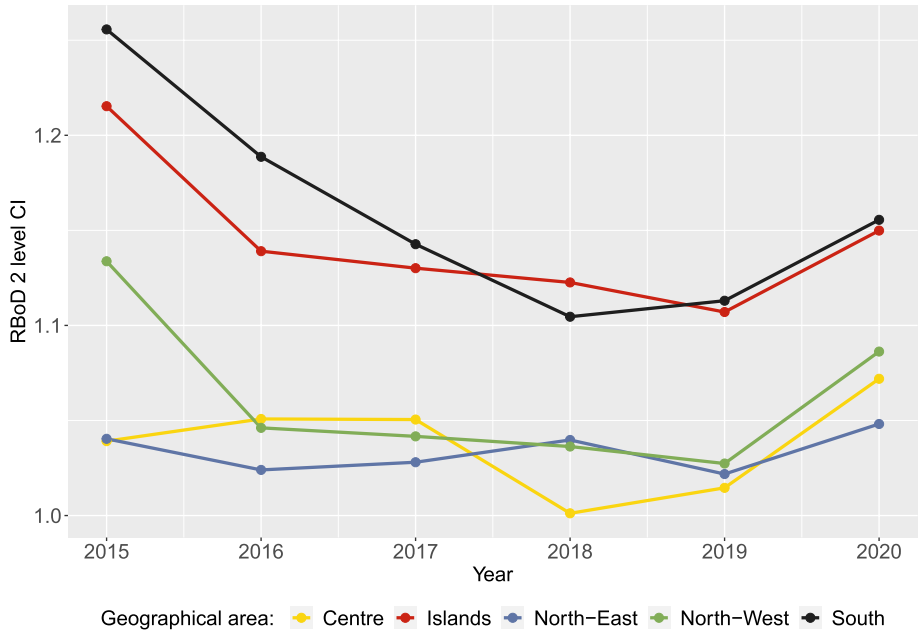

Fig. 6 CI trend by geographical area. NOTES: The figure depicts the dynamics of overall composite indicators at Level 2 across various geographical areas. SOURCES: Elaboration by the authors on Agenas data

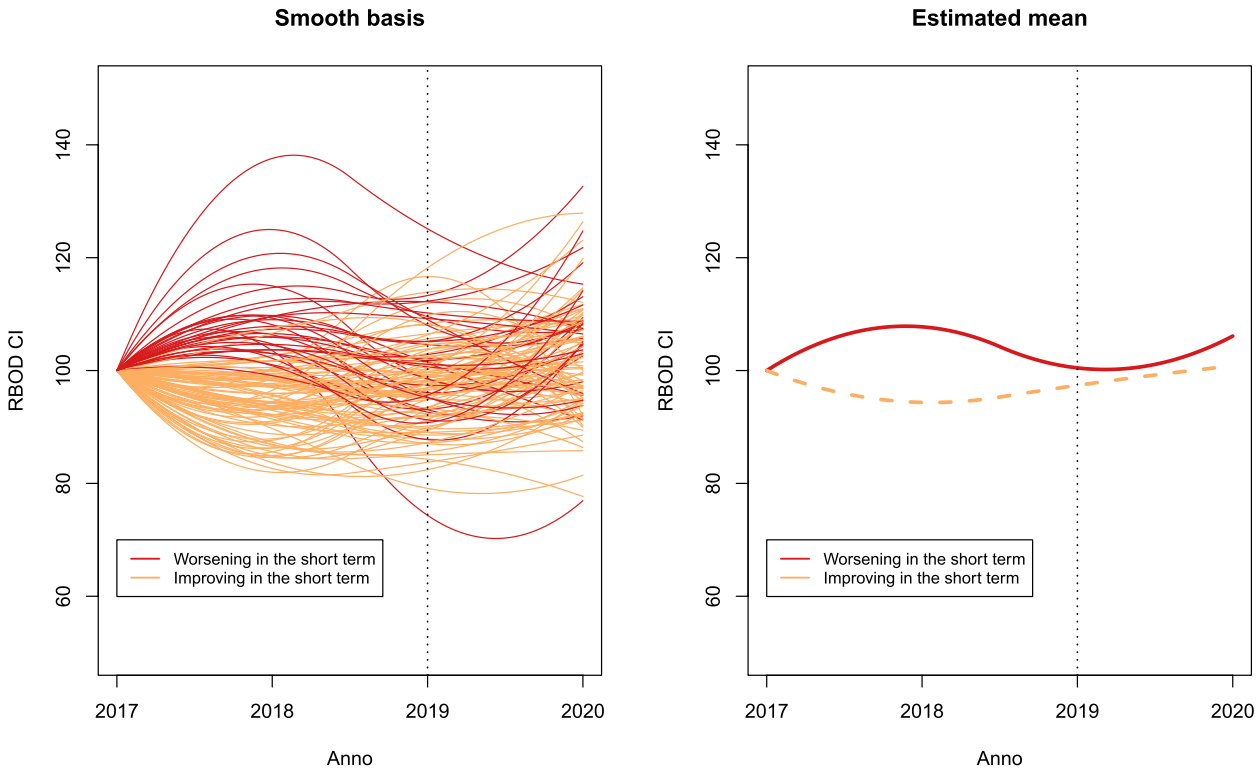

Fig. 7 Functional clusters - overall CI (year 2017 = 100). NOTES: The figure shows the functional cluster between Italian LHAs based on their performance on hospital quality over time identified using the Functional Data Analysis (FDA, [28]). SOURCES: Elaboration by the authors on Agenas data

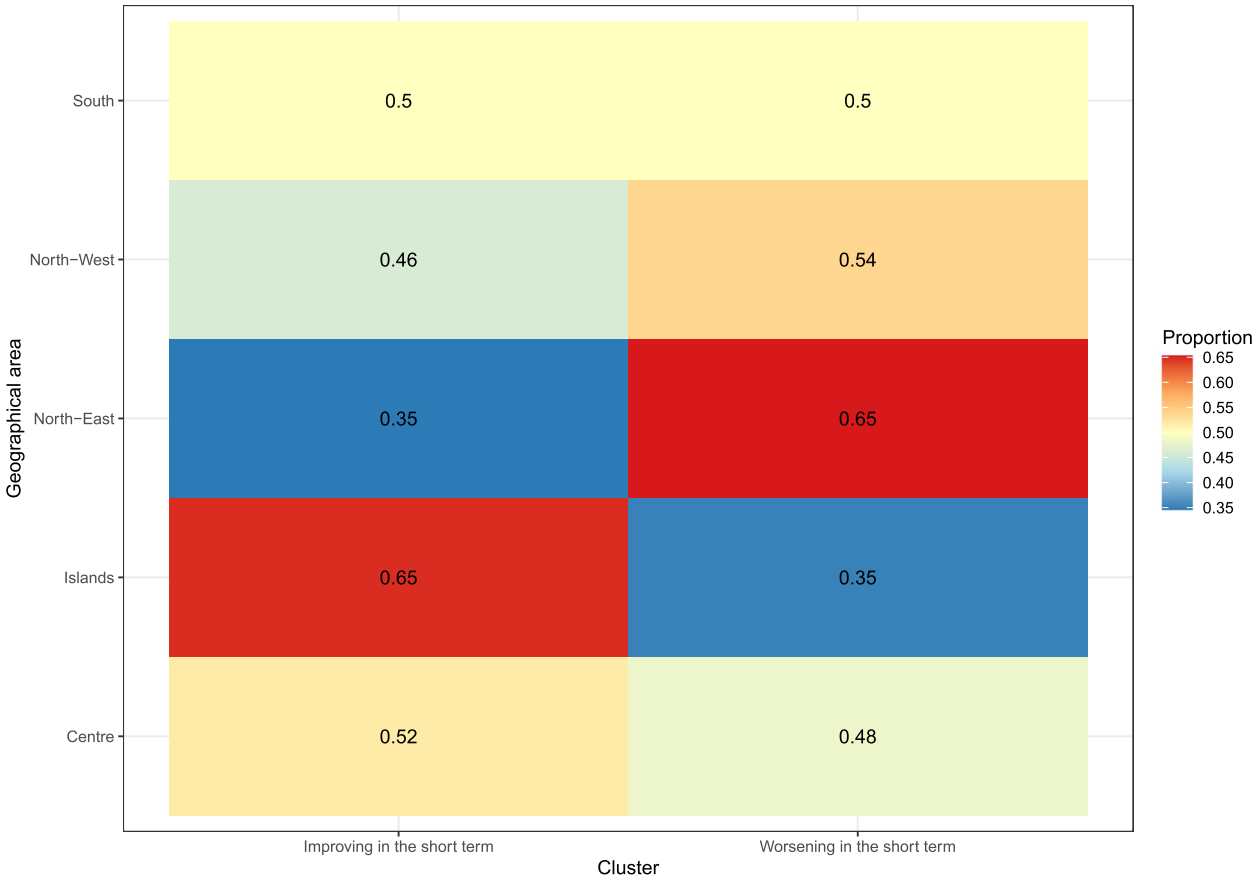

Fig. 8 Functional clusters by geographical area. NOTES: The figure shows the distribution of the functional cluster of Italian LHAs between geographical areas. SOURCES: Elaboration by the authors on Agenas data

and for cross-country evaluation. Indeed, the methodologies discussed - `BoD`, `RBoD` and `MPI` - provide a robust framework for assessing and comparing health performance in different regions and time periods. This comparative approach can improve our understanding of how different healthcare systems can be evaluated against international standards on quality of care.

In this perspective, we believe that our study could contribute to the global discussion of health metrics and assessment of the healthcare system. Indeed, the methods presented can be integrated into global health assessment frameworks, such as those utilised by the World Health Organization (WHO) and the Organization for Economic Co-operation and Development (OECD), supporting the effort to develop more comprehensive and reliable health metrics that can be used to compare the quality of healthcare in different countries and regions.

Finally, future research could explore the application of these methodologies to other healthcare systems and contexts. In this perspective, specific areas for future research could include, for example, examining the impact of different health policies on quality indicators, exploring the role of socio-economic factors in health performance, and analysing the long-term effects of health interventions on quality indicators. Moreover, through international collaborations, future research could also focus on the development of new composite indicators that capture emerging health challenges and priorities. Such collaborations could also improve the comparability of health care quality assessments globally and provide information on the factors that influence health care performance by fostering the development of a more unified approach to evaluation and improvement of health systems.

Appendix A. Simple indicators

Table 3 Simple indicators, Normalisation and polarisation criterion

Clinical area	Indicator	Weight (%)		VERY HIGH 1	HIGH 2	MEDIUM 3	LOW 4	VERY LOW 5
CARDIOVASCULAR	Acute myocardial infarction: 30-day mortality	30	%	⩽ 6	6 ⊣ 8	8 ⊣ 12	12 ⊣ 14	> 14
	Acute myocardial infarction: % treated with PTCA within 2 days	15	%	⩾ 60	45 ⊢ 60	35 ⊢ 45	25 ⊢ 35	< 25
	Congestive heart failure: 30-day mortality	10	%	⩽ 6	6 ⊣ 9	9 ⊣ 14	14 ⊣ 18	> 18
	Aortocoronary bypass: 30-day mortality	20	%	⩽ 1.5		1.5 ⊣ 4		> 4
	Valvuloplasty or heart valve replacement: 30-day mortality	15	%	⩽ 1.5		1.5 ⊣ 4		> 4
	Repair of unruptured abdominal aortic aneurysm: 30-day mortality	10	%	⩽ 1		1 ⊣ 3		> 3
NEUROLOGY	Ischaemic stroke: 30-day mortality	75	%	⩽ 8	8 ⊣ 10	10 ⊣ 14	14 ⊣ 16	> 16
	Surgery for cerebral T: 30-day mortality after craniotomy surgery	25	%	⩽ 1.5		1.5 ⊣ 3.5	3.5 ⊣ 5	> 5
RESPIRATORY	Relapsed COPD: 30-day mortality	100	%	⩽ 5	5 ⊣ 7	7 ⊣ 12	12 ⊣ 16	> 16
GENERAL SURGERY	Laparoscopic cholecystectomy: % admissions with post-operative stay < 3 days	50	%	⩾ 80	70 ⊢ 80	60 ⊢ 70	50 ⊢ 60	< 50
	Ordinary laparoscopic cholecystectomy: 30-day complications val	50	%	=100	80 ⊢ 100	50 ⊢ 80	30 ⊢ 50	< 30
SURG. ONCOLOGY	TM breast surgery: % operations in wards with volume of activity > 135 cases	33	%	=100	80 ⊢ 100	50 ⊢ 80	30 ⊢ 50	< 30
	Proportion of new resections within 120 days after conservative surgery for malignant tumour	17	%	⩽ 5	5 ⊣ 8	8 ⊣ 12	12 ⊣ 18	> 18
	Surgery for TM lung: 30-day mortality	17	%	⩽ 0.5		0.5 ⊣ 3		> 3
	Surgery for TM stomach: 30-day mortality	8	%	⩽ 2	2 ⊣ 4	4 ⊣ 7	7 ⊣ 10	> 10
	Surgery for TM colon: 30-day mortality	25	%	⩽ 1	1 ⊣ 3	3 ⊣ 6	6 ⊣ 8	> 8

Clinical area	Indicator	Weight (%)		VERY HIGH 1	HIGH 2	MEDIUM 3	LOW 4	VERY LOW 5
PREGNANCY	Proportion of deliveries by primary caesarean section	80	%	⩽ 15	15 ⊣ 25	25 ⊣ 30	30 ⊣ 35	> 35
	Vaginal childbirth: Subsequent Admissions During Puerperium	10	%	⩽ 0.20		0.20 ⊣ 0.70		> 0.70
	Caesarean sections: proportion of complications during labour and puerperium	10	%	⩽ 0.30		0.30 ⊣ 1.2		> 1.2
MUSCULOSKELETAL	Femoral neck fracture: surgery within 2 days	90	%	⩾ 70	60 ⊢ 70	50 ⊢ 60	40 ⊢ 50	< 40
	Hip replacement surgery: readmissions at 30 days val	10	gg	< 2	2 ⊢ 4	4 ⊢ 6	6 ⊢ 8	⩾ 8

Notes: The table shows ....

Sources: Authors' elaborations on Agenas data

Appendix B. Other statistics

Table 4 Summary statistics - CI at Level 1, BoD calculation

Clinical area	2015		2016		2017		2018		2019		2020	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Cardiovascular	0.9	0.13	0.86	0.15	0.85	0.15	0.82	0.16	0.82	0.15	0.88	0.14
Neurology	0.75	0.23	0.68	0.23	0.69	0.23	0.65	0.23	0.66	0.21	0.72	0.23
Respiratory	0.61	0.17	0.55	0.17	0.57	0.17	0.55	0.19	0.56	0.18	0.69	0.19
General surgery	0.77	0.23	0.71	0.24	0.65	0.24	0.62	0.26	0.65	0.25	0.63	0.26
Surgical Oncology	0.86	0.18	0.82	0.16	0.81	0.17	0.81	0.18	0.81	0.18	0.83	0.18
Pregnancy	0.78	0.2	0.79	0.19	0.81	0.22	0.83	0.16	0.83	0.17	0.72	0.2
Musculoskeletal	0.76	0.24	0.76	0.24	0.68	0.27	0.64	0.24	0.66	0.24	0.73	0.26

Notes: the table reports the descriptive statistics of the composite indicators computed using BoD method in our sample of Italian LHA in the time span from 2015 to 2020 for individual clinical areas (Level 1)

Sources: Authors' elaborations on Agenas data

Table 5 Summary statistics - CI at Level 1, MPI calculation

Clinical area	2015		2016		2017		2018		2019		2020	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Cardiovascular	101	8.4	100	6.1	100	4.2	92	8.7	98	4	99	5.3
Neurology	101	9.2	99	7.3	100	8.8	98	6.8	98	7.9	99	6.9
Respiratory	103	7.5	100	7.5	101	7.6	100	8.2	101	8.1	106	8.6
General surgery	99	6.3	105	8.1	96	6.1	101	7.6	95	5.1	101	7.1
Surgical Oncology	100	6.1	100	5.1	99	5	98	5.4	98	4.9	97	5.2
Pregnancy	101	7.3	99	5.1	97	5.5	102	6.3	100	5.5	97	4.8
Musculoskeletal	104	7.8	98	7	100	7.1	96	6.5	100	6.4	98	7.2

Notes: the table reports the descriptive statistics of the composite indicators computed using MPI method in our sample of Italian LHA in the time span from 2015 to 2020 for individual clinical areas (Level 1)

Sources: Authors' elaborations on Agenas data

Table 6 Summary statistics - CI at Level 1, composite indicators `Agenas`

Clinical area	2015		2016		2017		2018		2019		2020	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Cardiovascular	2.8	0.54	2.7	0.48	2.6	0.44	2.5	0.46	2.5	0.38	2.6	0.45
Neurology	3	1	2.8	1	2.9	0.99	2.6	0.96	2.5	0.89	2.9	0.98
Respiratory	3.3	0.75	3	0.75	3.1	0.76	3	0.82	3.1	0.81	3.6	0.86
General surgery	3	0.88	2.7	0.88	2.4	0.83	2.3	0.88	2.3	0.74	2.3	0.83
Surgical Oncology	2.9	0.74	2.8	0.65	2.7	0.66	2.7	0.74	2.6	0.64	2.5	0.67
Pregnancy	2.9	0.89	2.8	0.8	2.6	0.74	2.7	0.65	2.6	0.63	2.5	0.67
Musculoskel-etal	3	1.4	2.8	1.3	2.3	1.2	2.2	1.1	2.2	1.1	2.5	1.2

Notes: The table reports the descriptive statistics of the composite indicators `Agenas` for individual clinical areas (Level 1). The indicator `Agenas` is computed as a simple weighted average of the selected clinical indicators utilizing the weights assigned by AGENAS in the 2015 TREEMAP exercise

Sources: Authors' elaborations on Agenas data

Appendix C. Functional data clustering

Functional data analysis (FDA), as introduced by Ramsay and Silverman [28], expands on traditional multivariate techniques by embracing data that can be naturally represented as functions or curves.

One prominent challenge in the functional data approach arises from the assumption that observations exist in an infinite-dimensional space, whereas, in practical scenarios, we only have finite sampled curves observed at discrete time points. In most cases, discrete observations (X_{ij}) are available for each sampled path (X_{ij}) at a finite set of nodes ($t_{ij} : j = 1, ..., m_i$). Consequently, the initial step in an FDA-type analysis frequently involves reconstructing the functional representation of the data from these discrete observations using non-parametric smoothing techniques for functions.

In recent times, this approach has been extended to incorporate classical statistical estimation methods such as factor analysis, regression models, and clustering techniques. This extension is accomplished through non-parametric methods, which typically entail defining specific distances or dissimilarities for functional data and subsequently applying clustering algorithms such as hierarchical clustering or k-means. Alternatively, model-based algorithms, as described by Bouveyron and Jacques [4] and Bouveyron et al. [3], have been employed.

In the application presented in this paper, the funFEM algorithm, as described by Bouveyron and Jacques [5], was used to cluster diverse efficiency trends, modelling them as curves within a shared and distinctive functional subspace.

Acknowledgements
Not applicable.

Authors' contributions
The authors contributed equally to this work.

Funding
This study was funded by the European Union - NextGenerationEU, Mission 4, Component 2, in the framework of the GRINS - Growing Resilient, Inclusive and Sustainable project (GRINS PE00000018 - CUP E63C22002120006). The views and opinions expressed are solely those of the authors and do not necessarily reflect those of the European Union, nor can the European Union be held responsible for them.

Availability of data and materials
The Agenas data used in this study are publicly available at https://pne.agenas.it/indicatori.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 3 April 2024 Accepted: 16 September 2024
Published online: 04 October 2024

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. Ben Lahouel B, Ben Zaied Y, Taleb L, Kočišová K. The assessment of socio-environmental performance change: A Benefit of the Doubt indicator based on Directional Distance Function and Malmquist productivity index. Finance Res Lett. 2022;49:103164.
3. Bouveyron C, Côme E, Jacques J, et al. The discriminative functional mixture model for a comparative analysis of bike sharing systems. Ann Appl Stat. 2015;9(4):1726–60.
4. Bouveyron C, Jacques J. Model-based clustering of time series in group-specific functional subspaces. Adv Data Anal Classif. 2011;5(4):281–300.

5. Bouveyron C, Jacques J. funFEM: an R package for functional data clustering. Quatrième Rencontres R. Grenoble; 2015.
6. Cazals C, Florens J, Simar L. Nonparametric frontier estimation: A robust approach. J Econ. 2002;106(1):1–25.
7. Charnes A, Cooper WW, Rhodes E. Measuring the efficiency of decision making units. Eur J Oper Res. 1978;2(6):429–44.
8. Cherchye L, Knox Lovell CA, Moesen W, Van Puyenbroeck T. One market, one number? A composite indicator assessment of EU internal market dynamics. Eur Econ Rev. 2007;51(3):749–79.
9. Daraio C, Simar L. Introducing environmental variables in nonparametric frontier models: a probabilistic approach. J Prod Anal. 2005;24(1):93–121.
10. Daraio C, Simar L. Advanced robust and nonparametric methods in efficiency analysis. New York: Springer; 2007.
11. De Belvis AG, Meregaglia M, Morsella A, Adduci A, Perilli A, Cascini F, et al. Italy. Health system review 2022. World Health Organization. Regional Office for Europe - European Observatory on Health Systems and Policies. Copenhagen: WHO Regional Office for Europe; 2022.
12. De Witte K, Schiltz F. Measuring and explaining organizational effectiveness of school districts: Evidence from a robust and conditional Benefit-of-the-Doubt approach. Eur J Oper Res. 2018;267(3):1172–81.
13. Donabedian A. Evaluating the Quality of Medical Care. Milbank Memorial Fund Q. 1966;44(3):166–206.
14. Fare R, Karagiannis G, Hasannasab M, Margaritis D. A benefit-of-the-doubt model with reverse indicators. Eur J Oper Res. 2019;278(2):394–400.
15. Fusco E. Potential improvements approach in composite indicators construction: The Multi-directional Benefit of the Doubt model. Socio Econ Plan Sci. 2023;85:101447.
16. 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.
17. Fusco E, Vidoli F, Sahoo BK. Spatial heterogeneity in composite indicator: A methodological proposal. Omega. 2018;77(C):1–14.
18. Greco S, Ishizaka A, Matarazzo B, Torrisi G. Stochastic multi-attribute acceptability analysis (SMAA): an application to the ranking of Italian regions. Reg Stud. 2018;52(4):585–600.
19. Greco S, Ishizaka A, Tasiou M, Torrisi G. On the Methodological Framework of Composite Indices: A Review of the Issues of Weighting, Aggregation, and Robustness. Soc Indic Res. 2019;141:61–94.
20. Jacobs R, Smith P, Goddard MK. Measuring performance: an examination of composite performance indicators: a report for the Department of Health. New York: Centre of Health Economics, University of York; 2004.
21. 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.
22. Karagiannis G, Sarris A. Measuring and explaining scale efficiency with the parametric approach: the case of Greek tobacco growers. Agric Econ. 2005;33(s3):441–51.
23. Lagravinese R, Paolo L, Resce G. Exploring health outcomes by stochastic multicriteria acceptability analysis: An application to Italian regions. Eur J Oper Res. 2019;274(3):1168–79.
24. Lavigne C, De Jaeger S, Rogge N. Identifying the most relevant peers for benchmarking waste management performance: A conditional directional distance Benefit-of-the-Doubt approach. Waste Manag. 2019;89:418–29.
25. Mazziotta M, Pareto A. On a Generalized Non-compensatory Composite Index for Measuring Socio-economic Phenomena. Soc Indic Res. 2016;127(3):983–1003.
26. Melyn W, Moesen W. Towards a synthetic indicator of macroeconomic performance: unequal weighting when limited information is available. Public Econ Res Pap. 1991;1–24.
27. 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;267:121807.
28. Ramsay JO, Silverman BW. Functional data analysis. Springer Series in Statistics. New York: Springer; 2005.
29. Rogge N. Composite indicators as generalized benefit-of-the-doubt weighted averages. Eur J Oper Res. 2018;267(1):381–92.
30. Rogge N. On aggregating Benefit of the Doubt composite indicators. Eur J Oper Res. 2018;264(1):364–9.
31. 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 Econ. 2017;139:19–32.
32. Van Puyenbroeck T, Rogge N. Geometric mean quantity index numbers with Benefit-of-the-Doubt weights. Eur J Oper Res. 2017;256(3):1004–14.
33. Verbunt P, Rogge N. Geometric composite indicators with compromise Benefit-of-the-Doubt weights. Eur J Oper Res. 2018;264(1):388–401.
34. Vidoli F, Fusco E, Mazziotta C. Non-compensability in Composite Indicators: A Robust Directional Frontier Method. Soc Indic Res. 2015;122(3):635–52.
35. Vidoli F, Fusco E, Pignataro G, Guccio C. Multi-directional Robust Benefit of the Doubt model: An application to the measurement of the quality of acute care services in OECD countries. Socio Econ Plan Sci. 2024;93:101877.

Publisher's Note