INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

WILEY

INTERNATIONAL TRANSACTIONS IN OPERATIONAL RESEARCH

Intl. Trans. in Op. Res. 30 (2023) 503–544
DOI: 10.1111/itor.13099

Performance evaluation of emergency department physicians using robust value-based additive efficiency model

Anna Labijak-Kowalska[a], Miłosz Kadziński[a,*], Inga Spychała[a], Luis C. Dias[b,c], Javier Fiallos[d], Jonathan Patrick[e], Wojtek Michalowski[e] and Ken Farion[f]

[a]*Institute of Computing Science, Poznan University of Technology, Piotrowo 2, Poznań 60-965, Poland*
[b]*CeBER, Faculty of Economics, University of Coimbra, Av. Dias da Silva n. 165, Coimbra 3004-512, Portugal*
[c]*INESC Coimbra, Department of Electrical and Computer Engineering, University of Coimbra, Rua Sílvio Lima Polo II, Coimbra 3030-290, Portugal*
[d]*Elizabeth Bruyere Hospital, 43 Bruyere St., Ottawa, ON K1N 5C8, Canada*
[e]*Telfer School of Management, University of Ottawa, 55 Laurier Ave. E, Ottawa, ON K1N 6N5, Canada*
[f]*Quality and Systems Improvement, Children's Hospital of Eastern Ontario, 401 Smyth Road, Ottawa, ON K1H 8L1, Canada*
E-mail: anna.labijak@cs.put.poznan.pl [Labijak-Kowalska]; milosz.kadzinski@cs.put.poznan.pl [Kadziński]; ingaspychala@gmail.com [Spychała]; lmcdias@fe.uc.pt [Dias]; javier.fiallos@gmail.com [Fiallos]; patrick@telfer.uottawa.ca [Patrick]; wojtek@telfer.uottawa.ca [Michalowski]; farion@cheo.on.ca [Farion]

Received 28 October 2020; received in revised form 20 September 2021; accepted 27 November 2021

Abstract

We propose a novel variant of the value-based additive data envelopment analysis model. It conducts a comprehensive robustness analysis of efficiency outcomes for all feasible input and output weights using mathematical programming and the Monte Carlo simulation. We also introduce the original procedures for selecting a common vector of weights and an approach for investigating the stability of results in a multiscenario setting. The presented framework is applied to evaluate the performance of emergency department physicians using data from the Children's Hospital of Eastern Ontario in Ottawa. Our focus is on the physicians' performance when dealing with groups of patients' complaints related to abdominal pain and constipation, fever, extremity injury, head injury, and laceration/puncture. The obtained results emphasize the strong dependence of the physicians' performances on the selected weight vectors. However, they prove helpful in pointing out overall good performers who can serve as universal benchmarks or niche performers being markedly better in providing care to a given complaint group. They also offer a basis for developing an improvement plan for the underperforming physicians, identifying the priorities for a practice-oriented model, and recognizing the most challenging patients' complaints.

Keywords: data envelopment analysis; physician's performance; emergency department; robustness analysis; efficiency analysis; value-based additive efficiency; multiattribute value function; common set of weights

*Corresponding author.

1. Introduction

Measuring the performance in healthcare is a complex, multidimensional problem. At each level—from individual physicians through medical practice and tertiary care to the entire healthcare system—one expects that the properly working unit or institution provides the best possible care efficiently using available resources. The variety of indicators does not allow for a direct performance assessment by monitoring only selected individual measures that should be optimized. In turn, it is required to find a proper trade-off between consumed resources and the quality of provided care.

One of the primary methods for assessing healthcare efficiency is through patient satisfaction surveys, using, for example, a predefined Likert scale (Smith et al., 2004; Jennings et al., 2009). However, such a survey-based approach gives information only about the patients' perceptions while failing to capture how efficiently resources are utilized. Another approach that, in turn, considers multiple performance aspects consists of using the composite indicators to aggregate all the individual indicators into a single quality measure (Goddard and Jacobs, 2009). Nonetheless, this method requires selecting an appropriate aggregation approach and an arbitrary parametrization with weights associated with different indicators. A slight change in these subjective values may vastly influence the relative performance evaluation of healthcare units (Jacobs et al., 2005).

The subjectivity and arbitrariness issues related to setting the weight values are no longer present when using the data envelopment analysis (DEA) (Charnes et al., 1978), that is, a nonparametric efficiency evaluation method. This approach allows measuring the relative efficiency of decision-making units (DMUs), which consume multiple inputs (resources) and produce multiple outputs (effects). DEA allows performing evaluation and measurement without assigning prior weights. In turn, one DMU's efficiency score depends on the input and output values of others. These aspects contribute to DEA's applicability, making the results objective in relation to the scores computed using composite indicators.

Healthcare is, next to banking, agriculture, transportation, and education, one of the most common application areas of DEA (Liu et al., 2013). The most frequently considered DMUs are hospitals. Kohl et al. (2019) provided an in-depth review of DEA applications in healthcare with a particular focus on hospitals. Recent examples include evaluating the Greek National Health Service hospitals (Flokou et al., 2017), investigating an impact of the economic recession on the performance of hospitals in Pennsylvania (Chen et al., 2019), or assessment of the technical efficiency of a few hundred Turkish hospitals (Küçük et al., 2020).

When it comes to other types of DMUs, medical-group practices are becoming increasingly popular. Andes et al. (2002) investigated the organizational factors affecting the overall physician practice efficiency for over one hundred primary care physician practices in the United States. Furthermore, Testi et al. (2013) assessed the primary care physician practices in Italy when treating diabetic patients. In Portugal, primary healthcare units were assessed from a perspective of geographical inequity (Amado and Santos, 2009) and comparing two types of units (Gouveia et al., 2016). DEA has also been used to evaluate individual departments, such as emergency departments (EDs) or operating rooms (ORs). In particular, Kang et al. (2017) examined the efficiency of EDs to help hospitals plan the redesign. Ketabi et al. (2018) and Akkan et al. (2020) evaluated EDs of hospitals in Isfahan (Iran) and Istanbul (Turkey), intending to identify the improvement strategies for

the underperforming units. In turn, Basson and Butler (2006) compared multiple ORs to propose new resource allocations to improve their performance.

DEA has also proved helpful in evaluating the performance of nursing homes. In this context, a combination of different DEA models was used to study the care planning process, determine the best techniques to ensure the quality of care, and identify the determinants that affect the homes' efficiency. Such studies were conducted for the nursing homes in The Netherlands (Lee et al., 2009), Portugal (Veloso et al., 2018), and the United States (Kooreman, 1994; Shimshak et al., 2009). Other types of healthcare institutions evaluated included fire and emergency services (Choi, 2005), visiting nurse service agencies (Kuwahara et al., 2013), and health maintenance organizations (Siddharthan et al., 2000). DEA was also used to evaluate the performance of national healthcare systems (Zehra and Serpil, 2018).

Finally, DEA was used for the assessment of individual physicians working in the hospital. Chilingerian and Sherman (1990) identified the inefficient practice patterns of the physicians treating cardiac patients, whereas Wagner and Shimshak (2000) evaluated the primary care physicians from a managed care organization. Furthermore, Ozcan et al. (2000) compared the resource utilization between medical specialists in the treatment of Medicaid sinusitis patients in Virginia. Also, Johannessen et al. (2017) investigated the impact of hospital reform in Norway on the performance of individual physicians. Finally, Fiallos et al. (2017) developed a model to assess ED physicians' performance taking into account different complaint groups and different types of medical trainees.

The literature on the DEA-based evaluation of healthcare units is rich. Both standard (Basson and Butler, 2006; Kuwahara et al., 2013) and enhanced DEA models (e.g., network DEA; Khushalani and Ozcan, 2017; Gerami et al., 2020 or window-DEA; Flokou et al., 2017) have been used. Moreover, DEA has also been combined with statistical analysis (Chilingerian, 1995; Akkan et al., 2020), multiple criteria decision analysis (MCDA); Rouyendegh et al., 2019), or machine learning (Tosun, 2012). All these applications assess each DMU based on a single vector of input/output weights, namely the vector that yields the most favorable assessment for that unit. Yet, as the choice of any specific vector is open to debate, it is worth analyzing how assessments would change when applying other feasible weight vectors. A noteworthy exception in this regard is the work of Schang et al. (2016) who used ratio-based efficiency analysis (Salo and Punkka, 2011) to evaluate the impact of the chosen weights on the final score of composite indicators applied for evaluating a set of Scottish Health Boards.

This paper introduces a novel robust value-based framework for efficiency analysis. Specifically, we extend the value-based additive DEA (VDEA) model (Gouveia et al., 2008), which combines DEA with the multiattribute value theory (MAVT) (Keeney and Raiffa, 1993). The underlying idea is to convert the relevant inputs and outputs into criteria associated with marginal value functions and aggregate them using an additive model. In the standard VDEA model, each DMU can choose the weights associated with the marginal value functions that minimize the difference of comprehensive value (efficiency score) to the best DMU. In turn, we investigate the robustness of results attained for all feasible input and output weights. We deliver the outcomes referring to the efficiency measures, ranks, and preference relations, and for each of those, we propose methods based on mathematical programming providing information about the extreme values (minimum and maximum) obtained for a given result. As the differences between the extreme bounds are often large, the robustness analysis framework also incorporates stochastic methods based on the Monte

Carlo simulations. They are useful for estimating the distributions of the considered measures or relations. Such distributions are captured by acceptability indices, quantifying the proportion of feasible weights confirming a given result. For the purpose of this work, these methods have been implemented and made available as independent modules on the open-source *diviz* platform (Meyer and Bigaret, 2012).

The proposed methodological framework is also enriched in two ways. On the one hand, we introduce novel procedures for computing a representative vector of common weights that allows ranking all DMUs univocally. Such a vector is chosen to match as well as possible the conclusions obtained through the robustness analysis. In case one unit is robustly better than the other, the selected vector should emphasize this advantage. In turn, for DMU pairs that are indistinguishable in terms of robust results, the chosen vector should make the difference between these DMUs as small as possible. On the other hand, we provide methods for quantifying the results' robustness under different evaluation scenarios. These outcomes consider two levels of robustness. The first level refers to the robustness of outcomes for an individual scenario, whereas the second captures the stability of results given the multiplicity of possible scenarios.

We applied the proposed robust value-based efficiency analysis methods for evaluating the performance of ED physicians using data from the Children's Hospital of Eastern Ontario (CHEO) in Ottawa, Canada. We consider three inputs (the average encounter time per patient visit, the average number of laboratory tests per patient visit, and the average number of radiology orders per patient visit), and one output (rate of nonreturn patient visits within 72 hours). Our primary focus is on a group of patients with primary complaints upon presentation being abdominal pain and constipation. However, in a multiscenario analysis, we also consider two other complaint groups related to fever and lower or upper extremity injury, head injury, and laceration/puncture.

In Fiallos et al. (2017), the performance of the same physicians was evaluated using an original SBM-SWAT VRS efficiency model. The main motivation for its use was to penalize a "compensatory behavior," that is, preventing some physicians from being judged as efficient because of attaining advantageous results on only a single input or output. However, such an approach considers an extremely limited space of symmetric weights, hence clearly favoring the physicians with balanced performance profiles. Also, it involves an arbitrary parameterization of the model with precise values of symmetry factors (β) that are difficult to specify or determine experimentally. Finally, it derives the efficiency measures from analyzing the most favorable weight vector for each physician, providing precise scores that do not offer a common basis for physicians' comparison.

We demonstrate that a more fair and justified way for preventing the above-mentioned "compensatory effect" is to conduct a robustness analysis. It provides meaningful means for comparing physicians based on their performance for diverse scenarios relevant to efficiency analysis. The robust results are less affected by the inclusion or removal of a single physician; they can be derived when the number of DMUs is relatively small compared to the number of inputs and outputs, while highly discriminating between physicians. In this way, we may identify the subsets of the most distinguishing and underperforming physicians while counteracting the "compensatory effect" related to excelling at only a single aspect of their clinical role and performing poorly on the remaining inputs and outputs. Moreover, we demonstrate that the application of our framework is beneficial for directly comparing pairs of physicians, yielding insights for identifying potential outliers, and proposing gradual improvement paths for the DMUs.

The remainder of this paper is organized as follows. Section 2 describes the novel robust value-based methods for efficiency analysis. In Section 3, we describe a case study. Section 4 concludes the paper and discusses the potential implications of the proposed approach.

2. Robustness analysis framework for value-based additive efficiency analysis

In this section, we present the robustness analysis methods for the value-based additive efficiency model. First, we will remind the basic framework that selects for each DMU the input and output weights that minimize the difference of comprehensive value to the best DMU. We will then discuss two streams of procedures for investigating the robustness of efficiency results attained for all feasible input and output weights. Moreover, we will generalize the proposed approaches for investigating the stability of results in multiple scenarios that can be considered for the same DMUs. Finally, we will present the algorithms for selecting a common vector of weights based on robust outcomes.

In what follows, we will use the following notation:

- K—a number of units (DMUs);
- $\mathcal{D}$—a finite set of DMUs, $\mathcal{D} = \{\text{DMU}_1, \ldots, \text{DMU}_K\}$;
- N and M—the number of inputs and outputs, respectively;
- $Q = N + M$—a number of all factors relevant for the analysis;
- w_q—a weight associated with the qth factor (input or output);
- u_q—a marginal value function associated with the qth factor;
- $S_w = \{w = (w_1, w_2, \ldots, w_q)^T | w \geq 0, A_w w \leq 0\}$—a space of feasible weight vectors, where A_w is the coefficient matrix of user-defined linear weight constrains.

2.1. Reminder on value-based additive data envelopment analysis

DEA encompasses several models that can be used to measure the relative efficiency of DMUs. In the most standard approach, the efficiency is expressed as a ratio between a single virtual output and a single virtual input, that is, weighted sums of outputs and inputs, respectively (Charnes et al., 1978). The seminal CCR (Charnes et al., 1978) and BCC models (Banker et al., 1984) belong to this category of radial models, in which the weights involved in the efficiency measure are established by identifying the most advantageous scenario for the DMU under evaluation. Later, several nonradial models have been proposed, such as the directional distance function (Färe and Grosskopf, 2000) and the additive model (Charnes et al., 1985). All these methods share the core DEA features of considering an empirical production possibility set and allowing each DMU under evaluation to select the weights involved in the definition of efficiency in a way that makes its efficiency score as good as possible. When using an additive efficiency model (Charnes et al., 1985), the underlying idea is to maximize the L_1 distance of each DMU to the efficient frontier. A few issues can be associated with this model: the comparability of the scales on which the inputs and outputs are expressed, the very pessimistic character of the derived efficiency

measures, and the lack of their intuitive interpretation. To address these issues, Gouveia et al. (2008) proposed a variant of an additive DEA model, exploiting the links between DEA and MAVT. In this approach, the DMUs are treated as decision alternatives evaluated in terms of multiple relevant criteria. Each criterion corresponds to an input or an output factor in the traditional efficiency model. Specifically, a comprehensive value E_o of DMU_o is computed using an additive value function, that is, a weighted sum of the marginal values assigned to the performance on each factor:

$$E_o = \sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_o), \tag{1}$$

where w_q is the weight, interpreted as a scale coefficient of the marginal value functions u_j, such that w_q, $q = 1, \ldots, Q$, and $\sum_{q=1}^{Q} w_q = 1$. Moreover, a preference direction is associated with each factor q, $q = 1, \ldots, Q$. Function u_j takes values between 0 and 1, being nonincreasing for the criteria corresponding to inputs and nondecreasing for outputs. In this way, lesser inputs and greater outputs are more preferred and all inputs and outputs are express in comparable value scales. Overall, the comprehensive value lies in the range of [0, 1]. Using the above model, the efficiency of DMU_o relatively to the set of DMUs can be verified by solving the following linear programming (LP) problem:

$$\text{Minimize } d_o \tag{2}$$

s.t.

$$\left.\begin{array}{l} \left.\begin{array}{l} \displaystyle\sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_k) - \sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_o) \le d_o, \text{ for } k = 1, \ldots, K, \\ d_o \ge 0, \end{array}\right\} \\ \left.\begin{array}{l} \displaystyle\sum_{q=1}^{Q} w_q = 1, \\ w_q \ge 0, \ q = 1, \ldots, Q, \\ \mathbf{w} \in S_w. \end{array}\right\} \mathcal{W} \end{array}\right\}.$$

This LP minimizes the distance d_o of analyzed $\mathrm{DMU}_o \in \mathcal{D}$ to the unit with the greatest comprehensive value. If the least distance ($d_{*,o}$) is equal to 0, then DMU_o is considered efficient. It means that there exists some feasible weight vector for which DMU_o attains a comprehensive value not worse than the value of all other units. Otherwise, that is, if $d_{*,o} > 0$, DMU_o is not efficient, and $d_{*,o}$ reflects a "min-max regret" perspective. In the following sections, we will denote a set of constraints specifying all feasible, nonnegative, and normalized weights by $\mathcal{W}$.

The assessment of a DMU_o with E_o and $d_{*,o}$ reflects two different perspectives. On the one hand, E_o might be called an absolute efficiency score, as it is independent of the other DMUs. It indicates a score in [0, 1], where 1 corresponds to an ideal situation in which a DMU has a value of 1 on

every criterion, that is, it produces the maximum amount of outputs with the minimum amount of inputs, whereas 0 corresponds to a value of 0 on every criterion, that is, it produces the minimum amount of outputs with the maximum amount of inputs. On the other hand, $d_{*,o}$ corresponds to a DEA relative efficiency, that is, relative to the empirically observed efficient frontier, which could change if other DMUs were added or excluded from $\mathcal{D}$.

Let us emphasize that in terms of MCDA, DMUs with $d_{*,o} = 0$ would be formally called "weakly efficient." It is possible that a dominated unit would attain a comprehensive value that is at least as good as all other units' scores. If this effect was undesired, one could either assume that the weights w_q, $q = 1, \ldots, Q$, should be positive or solve a second LP problem to maximize the minimal weight values for $d_o = 0$ (Gouveia et al., 2008).

When computing $d_{*,o}$, only the input/output weight vector most favorable to DMU_o is taken into account, which limits the insights that can be obtained from the analysis. First, it makes the comparison of efficiency scores questionable due to the nonuniqueness of the weight vectors favorable to each DMU, that is, the analyst lacks a common basis to analyze the attained efficiencies. Second, such an analysis neglects other weight vectors that could provide a realistic setting for the comparison of DMUs, potentially leading to useful information on the variety of efficiency scores under a variety of scenarios. Third, it provides limited means for discriminating between the units. This is particularly true when the number of considered factors is large, implying that a large subset of DMUs can be deemed efficient.

The limitations of using a single weight vector motivated the development of methods for robustness analysis (Lahdelma and Salminen, 2006; Salo and Punkka, 2011; Kadziński et al., 2017). Their essence consists of investigating the stability of outcomes for all feasible weights associated with the inputs and outputs. In what follows, we discuss the methods that incorporate the mathematical programming techniques to capture the exact, extreme outcomes, or the Monte Carlo simulation to estimate the distribution of results observed for feasible weights. When doing so, we assume a uniform distribution of weights. In this way, each weight vector has equal chances ($= 1/\text{vol}(W)$, where $\text{vol}(W)$ is the volume of the feasible weight space) to be considered within a sample of weights derived in the simulation. However, it is also possible to use the method with some exogenously given weight distribution.

2.2. Robustness analysis with mathematical programming

In this section, we discuss the mathematical models for computing the extreme efficiency results observed in the set of all feasible weights. We refer to three types of outcomes: efficiency scores, ranks, and pairwise preference relations.

When it comes to the efficiency scores, we may consider the relative distances or absolute values. For the former (Gouveia et al., 2008), we are interested in the range $[d_{*,o}, d_o^*]$ delimited by the least $d_{*,o}$ and the greatest d_o^* possible distance of DMU_o from the efficient unit that attains the maximal comprehensive value for a given weight vector. The minimal distance $d_{*,o}$ can be computed as explained in Section 2.1, whereas the maximal one, d_o^*, can be obtained by solving the following mixed-integer linear programming (MILP) model:

$$\text{Maximize } d_o \tag{3}$$

s.t.

$$\left.\begin{aligned}
&\sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_k) - d_o \geq \sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_o) - C(1-b_k), \text{ for } k = 1, \ldots, K,\ k \neq o,\\
&\sum_{k=1,\ldots,K;k\neq o} b_k = 1,\\
&b_k \in \{0,1\}, \text{ for } k = 1, 2, \ldots, K; k \neq o,\\
&d_o \geq 0,\\
&\mathcal{W},
\end{aligned}\right\}$$

where C is a large positive constant. The above model maximizes the distance of DMU_o from some other DMU. The first four constraints guarantee that d_o is equal to the difference between E_k and E_o for some $k \in \{1, \ldots, K\}$ and $k \neq o$. Note that if a binary variable $b_k \in \{0, 1\}$ is equal to 0, then the first constraint is always satisfied, whereas in case $b_k = 1$, then $C(1 - b_k) = 0$ and $d_o = E_k - E_o$. We require that the latter holds for some DMU_k, $k \in \{1, \ldots, K\}$ and $k \neq o$.

In turn, the interval $[E_{*,o}, E_o^*]$, delimited by the least $E_{*,o}$ and the greatest E_o^* efficiency scores, can be determined by optimizing the comprehensive value of DMU_o subject to the constraints defining a set of admissible inputs and output weights, that is,

$$\text{Minimize/maximize } \sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_o), \text{ s.t. } \mathcal{W}. \tag{4}$$

The rank-oriented perspective offers greater stability because it is based on ordinal rather than cardinal comparisons (Salo and Punkka, 2011). Note that some small changes in the data that might change DMU scores might still keep the ranking of the DMUs unchanged (Kadziński et al., 2017). To compute the best (minimal) possible $R_{*,o}$ rank for DMU_i, the following MILP problem needs to be solved:

$$\text{Minimize } 1 + \sum_{k=1,\ldots,K;k\neq o} b_k \tag{5}$$

s.t.

$$\left.\begin{aligned}
&\sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_k) - \sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_o) \leq C b_k, \text{ for } k = 1, \ldots, K; k \neq o,\\
&b_k \in \{0,1\}, \text{ for } k = 1, 2, \ldots, K; k \neq o,\\
&\mathcal{W}.
\end{aligned}\right\}$$

The above problem minimizes the number of DMUs that, for some feasible weight vector, attain greater (absolute) efficiency than DMU_o. Such a number increased by one is equal to

$R_{*,o}$ (Kadziński et al., 2012b). To compute the worst (maximal) possible R_o^* rank for DMU_o, we need to maximize the number of DMUs with the efficiency scores greater than the efficiency of DMU_o. This can be attained by solving the following MILP problem:

$$\text{Maximize } 1 + \sum_{k=1,\ldots,K;k\neq o} b_k$$

s.t.

$$\left.\begin{array}{l} \sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_o) - \sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_k) \leq C(1-b_k), \text{ for } k=1,\ldots,K; k\neq o, \\ b_k \in \{0,1\}, \text{ for } k=1,2,\ldots,K; k\neq o, \\ \mathcal{W}. \end{array}\right\} \tag{6}$$

The extreme relative distances, absolute values, and ranks indicate the performance of each DMU in the least and the most favorable scenarios that correspond to the pessimistic and optimistic settings, respectively. When referring to the latter concepts, we will mean the weight vectors for which a DMU attains the worst or the best results in the entire space of feasible input and output weights from a particular outcome perspective.

It is also possible to compare DMUs in a pairwise fashion concerning their efficiencies for all feasible weights. This efficiency-based binary relation, which we call pairwise preference relation, is defined for any pair of DMUs, being independent of the remaining DMUs. Given a set of feasible weights associated with different factors, two certainty levels can be considered (Greco et al., 2008). On the one hand, the possible preference relation $\succsim_E^P$ holds for a pair (DMU_o, DMU_k) if $E_o \geq E_k$ for at least one feasible weight vector. To verify its truth, the following LP model needs to be solved:

$$\text{Maximize } d_{o,k}, \quad s.t. \quad \sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_o) - \sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_k) \geq d_{o,k} \text{ and } \mathcal{W}. \tag{7}$$

If the maximal attained value of $d_{o,k}$ is not lesser than 0, there exists at least one feasible vector $\mathbf{w}$ for which $E_o \geq E_k$, and thus $\mathrm{DMU}_o \succsim_E^P \mathrm{DMU}_k$. On the other hand, the necessary preference relation $\succsim_E^N$ holds for a pair (DMU_o, DMU_k) if $E_o \geq E_k$ for all feasible weight vectors. Its truth can be verified be considering the following LP problem:

$$\text{Minimize } d_{o,k}, \quad s.t. \quad \sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_o) - \sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_k) \leq d_{o,k} \text{ and } \mathcal{W}. \tag{8}$$

When the minimal value of $d_{o,k}$ is greater than or equal to 0, there is no feasible weight vector for which $E_k > E_o$. This, in turn, implies that $E_o \geq E_k$ holds for all feasible weights, and thus $\mathrm{DMU}_o \succsim_E^N \mathrm{DMU}_k$.

Analyzing exact robust results is fundamental in decision problems with high stakes, where the specification of weight constraints is impossible or hampered by significant uncertainties. Then,

the variety of results is greater, and one may implement more "precautionary" rules based on the analysis of the worst possible results (in case of efficiency scores or ranks) or the necessary outcomes (in case of preference relations).

2.3. Robustness analysis methods for multiple scenarios of efficiency evaluation

The robustness analysis methods for DEA have been initially designed for dealing with a single scenario (Kadziński et al., 2017; Salo and Punkka, 2011), representing a particular evaluation context for a set of homogeneous DMUs. In such a scenario, the units are characterized by precise values of inputs and outputs. However, in some situations, the same set of DMUs could be evaluated under multiple scenarios. Let us denote a set of such scenarios by $\mathcal{S}$. For each DMU, the input and output values may differ from one scenario to another, hence potentially leading to different efficiency results. For example, in the study discussed in this paper, the scenarios correspond to different complaint groups, with complaints in each group forming a relatively homogenous population regarding ED management. It does not make sense to jointly consider different clinical and diagnostic categories, as this would lead to an averaging effect. Practice variations are expected and observed across presenting complaints due to the difference in resource utilization patterns for each type of complaint. This motivated accounting for each group separately and producing performance evaluations per type of complaint.

In this section, we extend the robust methods to address such multi-scenario settings. This is attained by adopting the approaches proposed initially for dealing with group decision-making problems (Greco et al., 2012). The multiscenario robust results consider two levels of certainty for the efficiency outcomes. The first level refers to the robustness analysis results for each scenario $S \in \mathcal{S}$. In what follows, we focus only on the exact outcomes computed with mathematical programming (see Section 2.2). Let us denote the extreme distances to the efficient unit by $[d_{*,o,S}, d^*_{o,S}]$, the extreme efficiency scores by $[E_{*,o,S}, E^*_{o,S}]$, the extreme ranks by $[R_{*,o,S}, R^*_{o,S}]$, and the necessary and possible preference relations by $\succsim^N_{E,S}$ and $\succsim^P_{E,S}$, respectively. The other level concerns the support given to some robust results by different scenarios. For this purpose, we consider the necessary and possible support depending on whether some outcome is confirmed by all or at least one scenario, respectively. Without loss of generality, we define the considered results only in the context of the necessary preference relation and extreme efficiency scores, and they can be generalized analogously to the possible relation, extreme distances, and scores:

- the necessary-necessary preference relation $\succsim^{N,N}_{E,S}$ holds for $(\mathrm{DMU}_o, \mathrm{DMU}_k) \in \mathcal{D} \times \mathcal{D}$ if for all $S \in \mathcal{S}$, $\mathrm{DMU}_o \succsim^N_{E,S} \mathrm{DMU}_k$;
- the necessary-possible preference relation $\succsim^{N,P}_{E,S}$ holds for $(\mathrm{DMU}_o, \mathrm{DMU}_k) \in \mathcal{D} \times \mathcal{D}$ if for at least one $S \in \mathcal{S}$, $\mathrm{DMU}_o \succsim^N_{E,S} \mathrm{DMU}_k$;
- the set of possible-necessary efficiency ranks $[R^N_{*,o,S}, R^{*,N}_{o,S}]$ is a set of ranks attained for all $S \in \mathcal{S}$, that is, $[R^N_{*,o,S}, R^{*,N}_{o,S}] = \bigcap_{S\in\mathcal{S}}[R_{*,o,S}, R^*_{o,S}]$;
- the set of possible-possible efficiency ranks $[R^P_{*,o,S}, R^{*,P}_{o,S}]$ is a set of ranks attained for at least one $S \in \mathcal{S}$, that is, $[R^P_{*,o,S}, R^{*,P}_{o,S}] = \bigcup_{S\in\mathcal{S}}[R_{*,o,S}, R^*_{o,S}]$.

Note that in the case of great divergence of results for various scenarios, $[R^{N}_{*,o,S}, R^{*,N}_{o,S}]$ can be empty, whereas $[R^{P}_{*,o,S}, R^{*,P}_{o,S}]$ does not need to be continuous, that is, there can be some holes in the range delimited by $R^{P}_{*,o,S}$ and $R^{*,P}_{o,S}$. In any case, these outcomes are useful for verifying the stability of performance under multiple scenarios, indicating the spaces of agreement and discordance for the same unit or pair of DMUs.

2.4. Robustness analysis with the Monte Carlo simulation

In most decision problems, the difference between the extreme distances, scores, or ranks is large, the possible relation is rich, whereas the necessary one is relatively poor. Thus, it is important to determine the distribution of distances, scores, ranks, and relations over the feasible weight space. Such a probability distribution can be estimated with Monte Carlo simulations. To generate a random sample of weights, we apply the hit-and-run algorithm (Ciomek and Kadziński, 2021). In general, it is possible to use any arbitrarily chosen probability distribution on the joint density function in the feasible weight space. When it can be reliably defined, the evaluation model reflects the DM's preferences more faithfully. However, elicitation of a fully specified probability distribution calls for a major effort. When it is not possible, a standard assumption—also made in this paper—is to consider weights that are uniformly distributed in the feasible space (Lahdelma and Salminen, 2006). As noted in Kadziński et al. (2017), it is in line with the spirit of robustness analysis, where each feasible weight vector is equally authorized to make some outcome nonnecessary or possible, or shift the extreme bounds.

The distribution of different efficiency results can be captured with the stochastic acceptability indices quantifying the shares of feasible weights confirming a given outcome. We consider the following four types of indices:

- *Distance acceptability interval index* ($DAII$) (DMU_o, b_i) is the share of feasible weights for which the distance (to the efficiency frontier) of DMU_o to the best unit belongs to the interval $b_i = (b_{i,*}, b^*_i]$, being one of the B buckets partitioning the range [0,1] so that $\bigcup_{i=1}^{B} b_i = [0, 1]$, $b_i \cup b_j = \emptyset$, $i \neq j$, and b_1 is left-closed, that is, $b_1 = [b_{1,*} = 0, b^*_1]$ (by default, we assume that $b^*_i - b_{i,*} = b^*_{i+1} - b_{i+1,*}$, $i = 1, \ldots, B-1$).
- *Efficiency acceptability interval index* ($EAII$) (DMU_o, b_i) is the share of feasible weights for which the efficiency (in terms of comprehensive score) of DMU_o, E_o, belongs to the interval b_i.
- *Efficiency rank acceptability index* ($ERAI$) (DMU_o, r) is the share of feasible weights for which DMU_o attains rth rank (in terms of comprehensive score).
- *Pairwise efficiency outranking index* ($PEOI$) (DMU_o, DMU_k) is the share of feasible weights for which DMU_o attains at least as good efficiency as DMU_k ($E_o \geq E_k$) (in terms of comprehensive score).

Also, by averaging the measures observed for all feasible weight vectors derived with the Monte Carlo simulations, we may estimate for DMU_o its expected distance $Ed(\mathrm{DMU}_o)$ to the efficient DMU, expected efficiency $EE(\mathrm{DMU}_o)$, and expected rank $ER(\mathrm{DMU}_o)$. These measures can be used to impose a complete order on the considered set of DMUs (Labijak-Kowalska and Kadziński, 2021). Their analysis is beneficial in decision problems with modest stakes or relatively

rich weight constraints, when the average performance or expected values may be used for deriving a decision recommendation representative for the entire set feasible weights.

In Section A1, we illustrate how such stochastic acceptability indices and expected efficiency measures are computed for the study considered in Section 3. To keep this illustration concise, we use a limited set of 10 samples. On the contrary, the results reported for the case study in the main paper are derived from the analysis of 10,000 uniformly distributed weight vectors.

The proposed robustness analysis based on mathematical programming and the Monte Carlo simulations have been made available on the open-source software platform *diviz* (Meyer and Bigaret, 2012). Each method was implemented as an independent module. These modules accept inputs and provide results in the XMCDA standard, enabling combining them into complex workflows and visualizing the results using other modules available on *diviz*.

2.5. Selection of a common vector of weights based on the outcomes of robustness analysis

In the traditional DEA models, for each DMU, we select a potentially different weight vector that reflects the most advantageous performance scenario for this unit. While this way of proceeding is useful for verifying the efficiency status of different DMUs, it may prevent a justifiable ranking or a selection of the best units due to the lack of a common base for their comparison (Contreras, 2020). In turn, robustness analysis is oriented toward summarizing the results of comparing the DMUs on all feasible input and output weights, hence offering multiple, possibly infinitely many, bases for joint consideration of all units. Even though such results are useful for understanding the stability of results, some users may find them challenging to understand, mainly due to the multiplicity of weight vectors that serve as the basis for conducting the robustness analysis.

In some applications, it might be more appropriate to consider the same basis for evaluating the DMUs, namely by selecting a common vector of weights for evaluating all DMUs. In this way, all units can be ranked on a unified scale, which increases the discrimination power compared to the classical DEA models. The idea of selecting a common vector of weights was introduced by Charnes et al. (1989), quickly finding its first applications in the evaluation of highway maintenance patrols (Cook et al., 1990) and farms in Kansas (Thompson et al., 1990). Over the last decades, multiple methods for determining a common vector of weights have been proposed. These approaches build on the concepts of ideal and anti-ideal alternatives, weighting schemes, cross-efficiency analysis, incorporating the DM's preferences, evaluating only a proper subset of DMUs, statistical analysis, or game theory (Contreras, 2020).

This section introduces the novel procedures for selecting a common vector of weights based on the analysis of results derived with robustness analysis. Overall, we aim at selecting a single weight vector representing the whole set of feasible input and output weights. Our purpose is to find a vector that matches as well as possible the results deemed to be robust. In particular, if the robust results warrant concluding that some DMU_o is better than some DMU_k, then the difference between the efficiency scores of these two DMUs should be enhanced. This will depend on the truth of a specific robust relation (let us denote it by $\succ^{\mathcal{W}}$), confirming the evident advantage of one DMU over another given the results attained for all feasible weights. On the other hand, we can point out the pairs of DMUs for which the efficiency difference should be small due to the ambiguity in their comparison, given all input and output weights. Such pairs are incomparable ($R^{\mathcal{W}}$) in terms of the

Table 1
Conditions justifying the truth of the robust preference $\succ^{\mathcal{W}}$ and incomparability $R^{\mathcal{W}}$ relations

Result	$\mathrm{DMU}_o \succ^{\mathcal{W}} \mathrm{DMU}_k$	$\mathrm{DMU}_l R^{\mathcal{W}} \mathrm{DMU}_p$
$\succsim_E^N$	$\mathrm{DMU}_o \succsim_E^N \mathrm{DMU}_k$ and not $(\mathrm{DMU}_k \succsim_E^N \mathrm{DMU}_o)$	not $(\mathrm{DMU}_l \succsim_E^N \mathrm{DMU}_p)$ and *not*$(\mathrm{DMU}_p \succsim_E^N \mathrm{DMU}_l)$
EE	$EE(\mathrm{DMU}_o) - EE(\mathrm{DMU}_k) > t_{EE}$	$\lvert EE(\mathrm{DMU}_o) - EE(\mathrm{DMU}_k)\rvert \leq t_{EE}$
ER	$ER(\mathrm{DMU}_o) - ER(\mathrm{DMU}_k) > t_{ER}$	$\lvert ER(\mathrm{DMU}_o) - ER(\mathrm{DMU}_k)\rvert \leq t_{ER}$
$PEOI$	$PEOI(\mathrm{DMU}_o, \mathrm{DMU}_k) - PEOI(\mathrm{DMU}_k, \mathrm{DMU}_o) > t_{PEOI}$	$\lvert PEOI(\mathrm{DMU}_l, \mathrm{DMU}_p) - PEOI(\mathrm{DMU}_p, \mathrm{DMU}_l)\rvert \leq t_{PEOI}$

robust relation $\succ^{\mathcal{W}}$. Thus interpreted, the selected common vector of weights is representative for all feasible weight vectors in the sense of the robustness concern.

The outcomes discussed in Sections 2.2 and 2.4 provide diverse bases for defining the conditions underlying the truth or falsity of the robust relation $\succ^{\mathcal{W}}$. In this paper, we will refer to four possibilities that build on the necessary preference relation ($\succsim_E^N$), expected efficiency scores (EEs) and ranks (ERs), and $PEOI$s. The respective conditions needed for establishing relations $\succ^{\mathcal{W}}$ and $R^{\mathcal{W}}$ are defined in Table 1. For example, when referring to $\succsim_E^N$, one unit can be judged as univocally more advantageous than another if it is necessarily preferred to it, confirming that its efficiency is at least as good for all feasible weights. On the contrary, the comparison based on $\succsim_E^N$ can be judged ambiguous if a given pair of units is incomparable in terms of $\succsim_E^N$. This means that for at least one feasible weight vector, one unit is judged more efficient, whereas, for some other input and output weights, the relation is inverse. Furthermore, when referring to the EEs and ERs, we can judge one unit as stochastically preferred to another if its expected efficiency or rank is better by some pre-defined threshold, t_{EE} or t_{ER}, specifying the minimal difference in expected results justifying an evident advantage. When such a threshold is not exceeded, we may assume that the difference is negligible. Finally, as far as $PEOI$s are concerned, the truth of a robust preference relation $\succ^{\mathcal{W}}$ is well motived when the share of feasible weights for which DMU_o is more efficient than DMU_k is greater than the share of weights for which the relation is inverse by more than threshold t_{PEOI}. By default, thresholds t_{EE}, t_{ER}, and t_{PEOI} are set to zero. However, the user can also set them to some positive values, hence imposing more demanding requirements for instantiating $\succ^{\mathcal{W}}$ as well as a greater tolerance for establishing $R^{\mathcal{W}}$.

The selection of a common vector of weights is conducted by attaining the two targets lexicographically. First, we maximize the minimal difference between efficiency scores for pairs of units related by $\succ^{\mathcal{W}}$, that is,

$$\text{Maximize } \alpha \tag{9}$$

s.t.

$$\left.\begin{array}{l} \text{for } (\mathrm{DMU}_o, \mathrm{DMU}_k) \in \mathcal{D} \times \mathcal{D} : \mathrm{DMU}_o \succ^{\mathcal{W}} \mathrm{DMU}_k : \\ \displaystyle\sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_o) - \sum_{q=1}^{Q} w_q u_q(\mathrm{DMU}_k) \geq \alpha, \\ \mathcal{W}. \end{array}\right\}$$

Let us denote the optimal solution of the above LP problem by α^*. Second, we minimize the maximal difference between efficiency scores for pairs of units related by $R^{\mathcal{W}}$, that is,

$$\text{Minimize } \beta \tag{10}$$

s.t.

$$\left.\begin{array}{l}
\text{for } (\text{DMU}_l, \text{DMU}_p) \in \mathcal{D} \times \mathcal{D} : \text{DMU}_l R^{\mathcal{W}} \text{DMU}_p : \\
\displaystyle\sum_{q=1}^{Q} w_q u_q(\text{DMU}_l) - \sum_{q=1}^{Q} w_q u_q(\text{DMU}_p) \le \beta, \\
\displaystyle\sum_{q=1}^{Q} w_q u_q(\text{DMU}_p) - \sum_{q=1}^{Q} w_q u_q(\text{DMU}_l) \le \beta, \\
\text{for } (\text{DMU}_o, \text{DMU}_k) \in \mathcal{D} \times \mathcal{D} : \text{DMU}_o \succ^{\mathcal{W}} \text{DMU}_k : \\
\displaystyle\sum_{q=1}^{Q} w_q u_q(\text{DMU}_o) - \sum_{q=1}^{Q} w_q u_q(\text{DMU}_k) \ge \alpha^*, \\
\mathcal{W}.
\end{array}\right\}$$

The common vector of weights selected in this way can be used to order all units from the best to the worst. The obtained ranking emphasizes the outcomes following the use of all feasible weights, which contributed to the selection of an underlying representative vector of weights. It positively affects the accuracy of the provided results while extending the robustness analysis in the capacity to explain its outcomes. The user can analyze the computed weights and efficiency scores, which is more understandable than examining the necessary, extreme, expected, or stochastic outcomes. Note that this idea has not yet been explored in the context of DEA, even though it has been successfully applied in MCDA (Kadziński et al., 2012a).

3. Case study: efficiency evaluation of emergency department physicians

In this section, we discuss the application of the proposed method for evaluating the performance of 20 full-time ED physicians. Data used in the study came from a sufficiently long period of time (12 months) and was controlled for a case-mix. A detailed description of the case study setting can be found in Fiallos et al. (2017).

We consider the following three inputs, reflecting the essential resources consumed by the physicians in the process of managing patients in the ED:

- i_1—an average encounter time per patient visit (AVG_MDTIME_PAT), which is defined as an average number of minutes between the first contact of the physician with the patient and the moment a disposition decision is made and recorded on a patient's chart;
- i_2—an average number of laboratory tests per patient visit (AVG_LAB_PAT) when diagnosing a patient;

Table 2
Input and output values for the 20 physicians given complaint group $G1$ (abdominal pain and constipation) (Fiallos et al., 2017)

MD	i_1—AVG_MDTIME_PAT	i_2—AVG_LAB_PAT	i_3—AVG_RAD_PAT	o_1—RATE_NR72
MD1	2.026	2.760	0.920	1.000
MD2	1.959	2.381	0.774	0.961
MD3	2.223	2.333	0.643	0.905
MD4	1.884	1.823	0.661	0.952
MD5	1.511	0.857	0.487	0.952
MD6	1.456	1.330	0.648	0.978
MD7	1.903	1.877	0.596	0.956
MD8	1.704	1.730	0.678	0.939
MD9	1.708	1.927	0.657	0.968
MD10	1.979	1.508	0.820	0.922
MD11	1.652	1.618	0.592	0.981
MD12	2.169	1.863	0.608	0.961
MD13	1.634	1.538	0.786	0.979
MD14	1.745	2.117	0.738	0.942
MD15	1.594	1.548	0.602	0.957
MD16	2.311	1.538	0.462	0.974
MD17	1.962	1.748	0.557	0.948
MD18	1.804	1.590	0.723	0.977
MD19	1.567	1.487	0.601	0.937
MD20	1.435	1.198	0.568	0.969

- i_3—an average number of radiology orders per patient visit (AVG_RAD_PAT) used in the diagnosis.

Indeed, one can expect that an efficiently working physician arrives at the correct diagnosis in a shorter time and ordering fewer laboratory tests and radiology orders than a less efficient one. As an output (o_1), we will consider each physician's quality of care measured by the rate of nonreturn patient visits within 72 hours of discharge (RATE_NR72). Such a value has been traditionally considered one of the most informative indicators of the physicians' performance (Hung and Chalut, 2008).

Patients have a variety of reasons for visiting an ED. Given different complaint groups, one may observe variations in the clinical practices and different levels of the available resources such as time, tests, or orders. For this reason, the efficiency of physicians should be evaluated individually for each complaint type, representing a different clinical and diagnostic category. In this case study, our primary focus is on a group of patients complaining ($G1$) about abdominal pain and constipation. The input and output values for this group are presented in Table 2. In this context, we will discuss the results of robustness analysis obtained with mathematical programming, the Monte Carlo simulation, and common sets of weights selected using different procedures. We will also consider two other complaint groups—fever ($G2$) and lower or upper extremity injury, head injury, and laceration/puncture ($G3$). The three groups will serve as the basis for the multiscenario robustness analysis. The descriptive statistics of inputs and outputs for all considered groups are

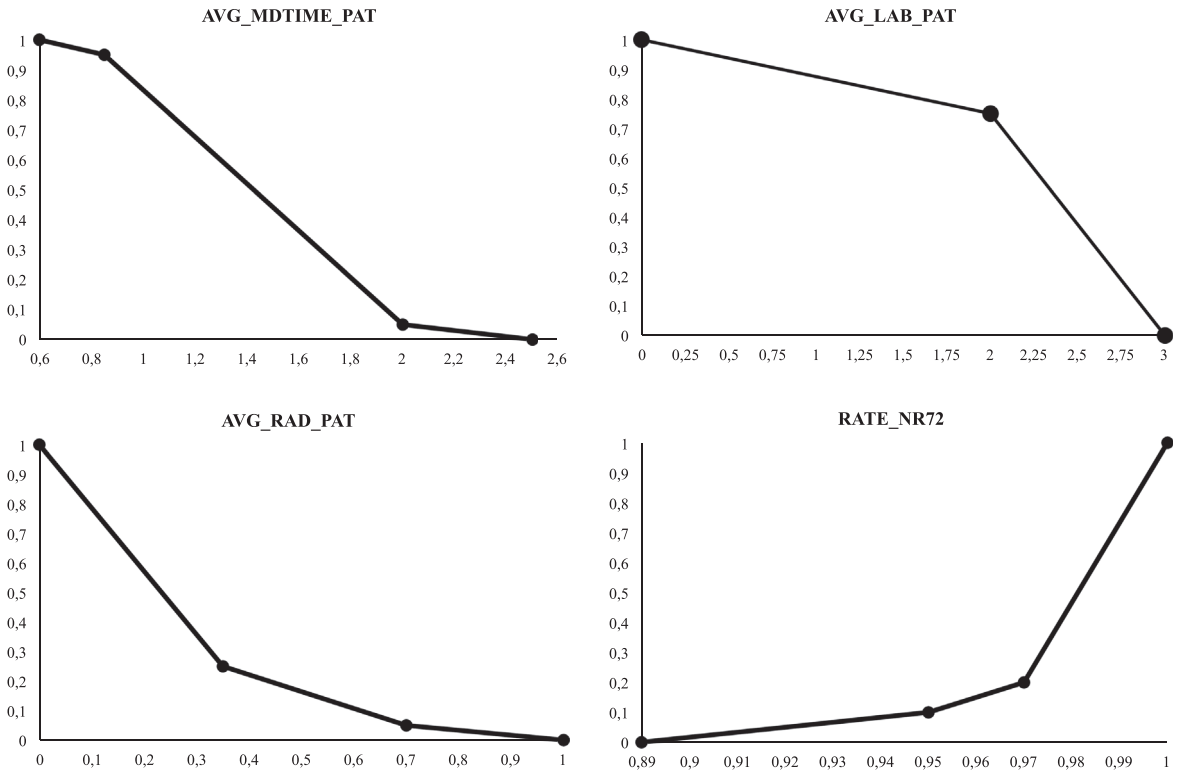

Fig. 1. Marginal value functions for the inputs and output used to evaluate the performance of ED physicians (x-axis—performances; y-axis—marginal value).

given in Section A2. To keep the main paper concise, some other data and results are also presented or discussed in the Appendices.

The marginal functions that will be used in the value-based efficiency analysis are presented in Fig. 1. They have been elicited from an independent medical expert using a direct questioning technique. He took into account performance ranges for each factor, the per-factor preferences, and the performances' distribution. This led to defining the convex functions for i_3 and o_1, a concave function for i_2, and a sigmoid-like function for i_1. Moreover, to prevent the dominating role of any factor on the final results, their weights have been constrained to at most 0.5 (i.e., $w_q \leq 0.5$, $q \in \{i_1, i_2, i_3, o_1\}$). We incorporated the latter assumption to avoid scenarios in which a physician is deemed efficient simply because of excelling at only one aspect of the clinical role while being ineffective at all other aspects.

Note that in the original case study, physicians' performance was analyzed using a more traditional SBM-SWAT VRS model that considers a single most advantageous weight vector for each DMU (Fiallos et al., 2017). The results presented in Fiallos et al. (2017) take the form of precise efficiency scores for each physician and each complaint group, providing somewhat limited and straightforward insights. In the following subsections, we discuss the insights derived from the analysis of all feasible weights offering means for identifying overall good or bad performers and

applying individual common sets of weights forming the basis for deriving univocal and well-justified ranking of physicians. In this perspective, we increase the discriminative power of efficiency results compared to the traditional methods and address the criticism leveled against the way the efficiency scores are computed in these approaches when analyzing only the input/output weights, which are the most favorable to each DMU. Moreover, we focus on different perspectives on the efficiency of physicians while acknowledging that ranks and pairwise preference relations are more interpretable to nonspecialists in DEA. When it comes to multiple scenarios, we present the aggregated results summarizing physicians' performance for different complaint groups instead of simply displaying the numerical outcomes for each considered scenario individually.

3.1. Robust efficiency results for complaint group G1: abdominal pain and constipation

This section presents the robust results for complaint group $G1$ concerning abdominal pain and constipation. First, we discuss robustness analysis outcomes referring to the efficiency scores (the discussion on the rank-related perspective and pairwise preference relations is provided in the Appendices). Second, we present the common sets of weights and the underlying rankings of physicians.

3.1.1. Distances to the efficient unit and efficiency scores

This section discusses the robustness of distances to the efficient physician and efficiency scores for the set of 20 physicians (further referred to as MD1, etc.). In Table 3, we present the extreme distances (columns d_* and d^*) and scores (columns E_* and E^*). The minimal distance d_* is equal to 0 for six physicians: MD1, MD5, MD6, MD11, MD16, and MD20. They are deemed efficient because they attain the greatest efficient score for at least one feasible weight vector. On the other extreme, even for the best scenario for MD2 and MD3, their minimal distances to the efficient physician are quite large (0.1836 and 0.1911, respectively). This implies that they are far from working efficiently.

The maximal efficiency score E^* is strongly correlated with the minimal distance d_*. This is understandable because if some physician acts efficiently or (s)he is close to being efficient, this should be due to attaining a relatively high efficiency score in the most favorable scenario. The greatest efficiency scores are attained by MD20 (0.6712) and MD6 (0.6547). It is worth noting that E^* for the efficient physician MD1 (0.5900) is lesser than E^* for the inefficient physicians: MD13 (0.6239), MD18 (0.5940), and MD19 (0.6015). This confirms the importance of analyzing the relative distances rather than absolute scores when deciding about efficiency.

When considering the least favorable scenarios, the best maximal distances d^* are between the two efficient physicians, MD11 (0.2041) and MD6 (0.2601), and inefficient MD13 (0.2688). This indicates that even in the most pessimistic scenarios for these physicians, the differences in their efficiencies are relatively small in terms of their scores on a scale of comprehensive value. The worst maximal distances d^* are for MD3 (0.4974) and MD1 (0.5575), being about twice as large as for the best performing physicians. When it comes to the minimal efficiency scores E_*, the best physicians are also MD6 (0.2465) and MD11 (0.2170). In turn, the least performing ones are MD1 (0.0304) and MD3 (0.0264), with efficiency scores very close to zero in the most pessimistic scenario.

Table 3
Extreme and expected values of distances to the efficient unit and efficiency scores for all considered physicians

MD	d_*	d^*	Ed	E_*	E^*	EE
1	0.0000	0.5575	0.1601	0.0304	0.5900	0.3169
2	0.1836	0.4103	0.2844	0.0599	0.3096	0.1882
3	0.1911	0.4974	0.3096	0.0264	0.2914	0.1619
4	0.0921	0.4155	0.1945	0.0911	0.4565	0.2785
5	0.0000	0.3658	0.0675	0.1409	0.6628	0.4048
6	0.0000	0.2601	0.0196	0.2465	0.6547	0.4552
7	0.0764	0.3957	0.1854	0.1177	0.4477	0.2873
8	0.0950	0.4345	0.1667	0.0721	0.5327	0.3061
9	0.0812	0.3744	0.1436	0.1323	0.5188	0.3297
10	0.0967	0.4650	0.2275	0.0417	0.4390	0.2457
11	0.0000	0.2041	0.0380	0.2170	0.6455	0.4370
12	0.0755	0.4297	0.2031	0.0678	0.4611	0.2699
13	0.0153	0.2688	0.0646	0.1861	0.6239	0.4108
14	0.1377	0.4415	0.2089	0.0652	0.4559	0.2638
15	0.0543	0.3862	0.1157	0.1205	0.5871	0.3572
16	0.0000	0.3609	0.1373	0.1025	0.5572	0.3361
17	0.0634	0.4355	0.1953	0.0882	0.4566	0.2772
18	0.0382	0.2903	0.1094	0.1248	0.5940	0.3656
19	0.0563	0.4142	0.1226	0.0925	0.6015	0.3499
20	0.0000	0.3465	0.0543	0.1602	0.6712	0.4188

To judge the stability of efficiency results for all feasible weights, we can refer to the distance and efficiency intervals' widths. On the one hand, the difference between d^* and d_* is the smallest for MD11 (0.2041), confirming the robustness of its relatively high-performance evaluation. On the other hand, for MD1, this difference is the greatest (0.5575), indicating high dependence of results on the selected input and output weights.

To expand the analysis of extreme distances and efficiency scores, we will estimate their distributions using Monte Carlo simulation (see Section A3 for the detailed results), considering 10 equally distributed buckets, from [0.0, 0.1] to (0.9, 1.0]. Note that the methods would work with any other arbitrarily specified subranges. Such distributions are useful for identifying the physicians consuming all their inputs and producing outputs efficiently, independently of the selected factor weights, or those physicians who are more oriented toward optimizing an individual input or output. Let us emphasize that smaller values are better when considering the distances, and larger values are better when considering the efficiency scores.

The distance of MD6 and MD11 to the efficient physician is lower than 0.1 for more than 95% weight vectors. This confirms that these physicians perform efficiently or are very close to being efficient for the vast majority of scenarios. Furthermore, even though MD16 is efficient, its distance from the efficient physician is most often between 0.1 and 0.2 (51.2%), and only for 29.5% weights, it lies in the interval [0.0, 0.1]. This suggests that MD16 cannot optimize all inputs and outputs equally well. The analysis of $DAII$s and $EAII$s is also helpful to identify the underperforming physicians. For example, the efficiency scores for MD2 and MD3 are at most 0.2 for, respectively, 57.3% and 68.4% feasible weight vectors, hence confirming their low performance in terms of the efficiency of provided care.

Table 4
Common sets of weights selected using four different procedures

Procedure	w_{i_1}	w_{i_2}	w_{i_3}	w_{o_1}
$\succsim_E^N$	0.46541	0.21630	0.11753	0.20076
ER	0.23715	0.24646	0.26183	0.25456
EE	0.36510	0.30947	0.00000	0.32543
$PEOI$	0.25502	0.19246	0.29040	0.26213

To construct a complete ranking of physicians without using a common vector of weights, we can use the expected distances to the efficient unit (Ed) and expected efficiencies (EE). These metrics are summarized in Table 3. They impose the same orders on the set of physicians under consideration. On the one hand, the top-ranked physicians are MD6 ($Ed = 0.0196$ and $EE = 0.4552$) and MD11 ($Ed = 0.038$ and $EE = 0.437$). For them, the difference to the best physician is, on average, very low, which confirms their position as overall good performers. On the other hand, the bottom-ranked physicians are MD2 ($Ed = 0.2844$ and $EE = 0.1882$) and MD3 ($Ed = 0.3096$ and $EE = 0.1619$), characterized by larger expected distances to the best physicians and lower expected efficiencies.

In general, the analysis of extreme distances and efficiency scores allows distinguishing the MDs exhibiting universal good practices to follow. These include units that attain favorable results for the wide spectrum of feasible weights. In this perspective, MD6 and MD11 can be considered for others as the benchmarks. Other MDs that are efficient only under specific conditions can be judged more niche (see, e.g., MD1 and MD16). These results are also helpful in discriminating between the inefficient DMUs. On the one hand, MDs with favorable extreme distances and scores have the most significant potential for becoming efficient. Therefore, the management may implement the corrective plan for units such as MD13 and MD18 in the first order. On the other hand, high distances and low scores indicate the MDs for which becoming efficient would be the most challenging, and the corrective actions need to be distributed over a longer-term (see, e.g., MD2 and MD3).

An analogous discussion on the robustness of efficiency ranks and pairwise preference relations is presented in Sections A4 and A5.

3.1.2. Analysis of rankings obtained by applying the common sets of weights

This section reports the results obtained using four procedures for selecting the common vector of weights presented in Section 2.5. They build on the expected efficiencies EEs (see Section 3.1.1), expected ranks ERs (see Section A4), the necessary preference relation $\succsim_E^N$, or $PEOI$s (see Section A5). We parameterize the procedures with the following thresholds justifying the truth of a robust preference relation: $t_{ER} = 0.5$, $t_{EE} = 0.1$, and $t_{PEOI} = 0.15$. Hence, to justify an evident advantage in performance of one physician over another, his/her expected rank should be better by more than 0.5, or the expected efficiency should be greater by more than 0.1, or the share of feasible input/output weights confirming better performance should be greater by more than 15% than the share of weights confirming worse performance.

In Table 4, we present the common sets of weights selected using four different procedures. For example, when considering the weights chosen based on the analysis of $\succsim_E^N$, the highest priority

Table 5
Efficiency scores and ranks attained by physicians for the common sets of weights selected using four different procedures

Procedure	$\succsim_E^N$		ER		EE		$PEOI$	
MD	Efficiency	Rank	Efficiency	Rank	Efficiency	Rank	Efficiency	Rank
1	0.2633	13	0.3137	11	0.3984	10	0.3127	8
2	0.1742	19	0.1832	19	0.2241	19	0.1618	19
3	0.1358	20	0.1578	20	0.1731	20	0.1339	20
4	0.2631	15	0.2706	14	0.3261	13	0.2343	14
5	0.4368	3	0.3956	5	0.4701	5	0.3609	5
6	0.4941	1	0.4444	1	0.5662	1	0.4133	1
7	0.2631	14	0.2802	13	0.3251	14	0.2453	13
8	0.3244	10	0.2971	12	0.3720	11	0.2622	12
9	0.3407	9	0.3210	10	0.3984	9	0.2886	11
10	0.2207	18	0.2372	18	0.2927	18	0.1958	18
11	0.4347	4	0.4279	2	0.5251	2	0.3975	2
12	0.2245	17	0.2632	16	0.2999	17	0.2265	17
13	0.4238	5	0.4002	4	0.5160	3	0.3669	4
14	0.2819	11	0.2559	17	0.3243	15	0.2265	16
15	0.3852	7	0.3481	7	0.4278	7	0.3152	7
16	0.2669	12	0.3303	9	0.3567	12	0.2947	10
17	0.2410	16	0.2706	15	0.3024	16	0.2343	15
18	0.3510	8	0.3562	6	0.4481	6	0.3208	6
19	0.3853	6	0.3407	8	0.4194	8	0.3073	9
20	0.4669	2	0.4088	3	0.5063	4	0.3767	3

is assigned to i_1, whereas the lowest priority is attributed to i_3. On the contrary, the values of weights selected based on ERs are more balanced, ranging between 0.23715 (for i_1) and 0.26183 (for i_3).

The respective efficiency scores and ranks for the 20 physicians are given in Table 5. These scores are derived from the lexicographic optimization of two targets—maximization of the efficiency difference for pairs of physicians related by the robust preference relations and minimization of such a difference for pair incomparable in terms of this relation.

Let us discuss in detail the results built on ER and $\succsim_E^N$. When it comes to the expected ranks (see Section A4), the three best performing physicians are MD6 (1.860), MD11 (2.914), and MD20 (3.682), whereas the three bottom-ranked physicians are MD10 (17.347), MD2 (18.669), and MD3 (19.640). The expected ranks' analysis is the basis for selecting a common vector of weights. For example, MD6 should be preferred to MD20, which, in turn, should be judged better than MD5, etc., according to the common weight vector to be chosen. Solving the LP problem (Section 2.5), the minimal efficiency difference for pairs with expected ranks differing by more than 0.5 is positive (0.00738). This means that the derived rankings reflect the order of physicians implied by ERs. For example, MD6 is ranked first with an efficiency of 0.444, and MD3 is ranked last with an efficiency of 0.1578. Thus, the expected results derived from the analysis of all feasible weights have been captured with a single common weight vector: [0.23715, 0.24646, 0.26183, 0.25456]. Moreover, the derived ranking can be seen as a synthetic representation of the expected results derived from the stochastic analysis.

Table 6
Values of Kendall's τ coefficient for all pairs of rankings obtained with different procedures

Procedure	$\succsim_E^N$	ER	EE	$PEOI$
$\succsim_E^N$	1.000	0.842	0.853	0.821
ER	0.842	1.000	0.926	0.958
EE	0.853	0.926	1.000	0.926
$PEOI$	0.821	0.958	0.926	1.000

In the same spirit, the efficiency scores built on $\succsim_E^N$ (see Section A5) allowed flattening, in a reasonable way, the graph of the necessary preference relation determined with mathematical programming. The minimal efficiency difference for pairs related by $\succsim_E^N$ is 0.04243. Hence, the procedure succeeded in reflecting the preference confirmed by all feasible weights in a complete order imposed by applying a single weight vector: [0.46541, 0.21630, 0.11753, 0.20076]. For example, such an advantage can be observed for the following pairs: (MD6, MD15), (MD15, MD8), (MD8, MD14), and (MD14, MD3). Furthermore, the physicians who are necessarily preferred to many other physicians attain the best scores and ranks according (see MD6 (1), MD20 (2), MD5 (3), MD11 (4), and MD13 (5)). On the contrary, the physicians necessarily outperformed by many others are ranked at the bottom (see MD10 (18), MD2 (19), and MD3 (20)). Interestingly, MD1, being incomparable in terms of $\succsim_E^N$ with any other physician, is ranked 13th, hence attaining an intermediate position. Overall, the analysis of such a ranking supports the comprehension of the necessary preference relation, making comparisons among the physicians more clear and the entire order well justified due to its roots in the outcomes observed for all feasible weight vectors.

The rankings constructed by the four procedures (see Table 5) are very similar. In Table 6, we present the values of Kendall's τ coefficient (Winkler and Hays, 1985) for all pairs of obtained rankings. For example, MD6 and MD3 are ranked at, respectively, the very top and very bottom by all procedures. The slight differences between the ranking produced by different procedures are the result of different tolerance levels that were used. On the one hand, the necessary preference relations left many pairs of physicians incomparable, whereas the expected ranks coupled with $t_{ER} = 0.5$ allowed comparing almost all pairs of physicians. On the other hand, requiring that physician's expected efficiency is better than another by more than 0.1 is clearly more limiting than requiring the difference in expected ranks to be greater than 0.5. Consequently, different numbers of pairs of physicians were considered in the two phases of lexicographic optimization. While this had an impact on the rankings, the subsets of the best, medium, and the worst performers stay the same.

The discussed rankings are also strongly correlated with the one presented in Fiallos et al. (2017), derived using the SBM-SWAT VRS model. The correlation coefficients range from 0.611 to 0.723 when considering the ranking based on EE or $\succsim_E^N$, respectively. When comparing the four rankings with the order reported in Fiallos et al. (2017), the positions attained by MD2, MD6, MD10, MD12, MD15, and MD20 differ by at most 2. The greatest differences are observed for MD1, MD3, MD11, and MD16 (up to 10, 6, 7, and 6 positions, respectively). The reasons underlying these differences have various origins. For example, we demonstrated that the performance of MD1 highly depends on the selected weight vector, while it was ranked at the bottom in Fiallos et al. (2017). Moreover, MD3 was judged as the worst performing physician according to all ranking

Table 7
The possible-necessary and possible-possible intervals of distances to the efficient physician, efficiency scores, and ranks based on the analysis of three complaint groups

MD	$[d^{P}_{*,o,S}, d^{*,P}_{o,S}]$	$[d^{N}_{*,o,S}, d^{*,N}_{o,S}]$	$[E^{P}_{*,o,S}, E^{*,P}_{o,S}]$	$[E^{N}_{*,o,S}, E^{*,N}_{o,S}]$	$[R^{P}_{*,o,S}, R^{*,P}_{o,S}]$	$[R^{N}_{*,o,S}, R^{*,N}_{o,S}]$
1	[0.0000, 0.5575]	[0.0631, 0.3784]	[0.0304, 0.9623]	[0.3095, 0.5900]	[1, 20]	[12, 18]
2	[0.0861, 0.4292]	[0.1836, 0.3437]	[0.0599, 0.7936]	[0.2861, 0.3096]	[13, 20]	[14, 20]
3	[0.0607, 0.4974]	[0.1911, 0.4653]	[0.0264, 0.8663]	[0.1735, 0.2914]	[9, 20]	[18, 20]
4	[0.0072, 0.4938]	[0.1277, 0.1317]	[0.0911, 0.9928]	[0.3850, 0.4565]	[2, 20]	[12, 18]
5	[0.0000, 0.4423]	[0.0839, 0.3174]	[0.1409, 0.8954]	[0.2710, 0.6628]	[1, 17]	[7, 8]
6	[0.0000, 0.4358]	[0.0000, 0.1670]	[0.2465, 0.9486]	[0.4215, 0.6547]	[1, 14]	[1, 5]
7	[0.0444, 0.4231]	[0.1057, 0.1830]	[0.1177, 0.8876]	[0.4054, 0.4477]	[6, 18]	[11, 12]
8	[0.0019, 0.4345]	[0.0950, 0.3191]	[0.0721, 0.9359]	[0.3579, 0.5327]	[2, 17]	[8, 10]
9	[0.0461, 0.4367]	[0.0812, 0.2946]	[0.1323, 0.9001]	[0.2938, 0.5188]	[5, 18]	[9, 15]
10	[0.0172, 0.4650]	[0.0967, 0.2591]	[0.0417, 0.8471]	[0.3318, 0.4390]	[3, 20]	∅
11	[0.0000, 0.4421]	[0.0706, 0.2041]	[0.2170, 0.8971]	[0.3515, 0.6455]	[1, 20]	[6, 8]
12	[0.0000, 0.4445]	[0.0755, 0.1401]	[0.0678, 0.9340]	[0.4483, 0.4611]	[1, 18]	∅
13	[0.0153, 0.3145]	[0.0257, 0.1597]	[0.1861, 0.9150]	[0.4287, 0.6239]	[2, 14]	[3, 7]
14	[0.0507, 0.4415]	[0.1377, 0.1791]	[0.0652, 0.8690]	[0.4106, 0.4559]	[4, 19]	[10, 16]
15	[0.0000, 0.3862]	[0.0543, 0.3312]	[0.1205, 0.9686]	[0.3120, 0.5871]	[1, 18]	[5, 11]
16	[0.0000, 0.4585]	∅	[0.1025, 1.0000]	[0.5094, 0.5572]	[1, 20]	[2, 6]
17	[0.0602, 0.4355]	[0.0910, 0.1881]	[0.0882, 0.8724]	[0.4003, 0.4566]	[3, 18]	[6, 18]
18	[0.0382, 0.4230]	[0.0915, 0.2903]	[0.1248, 0.8061]	[0.3039, 0.5940]	[4, 19]	[12, 14]
19	[0.0269, 0.4142]	[0.0584, 0.1659]	[0.0925, 0.9056]	[0.4225, 0.6015]	[3, 15]	[4, 9]
20	[0.0000, 0.4541]	[0.0139, 0.3465]	[0.1602, 0.9711]	[0.2278, 0.6712]	[1, 18]	[2, 8]

methods considered in this paper. This is implied by its unfavorable evaluation for the vast majority of feasible weights, which follows the transformation of its performances into marginal values using the functions presented in Fig. 1. However, according to Fiallos et al. (2017), five other physicians were judged worse than MD3.

3.2. Multiscenario robustness analysis for different complaint groups

In this section, we present the aggregated results of robustness analysis for the three complaint groups related to abdominal pain and constipation ($G1$), fever ($G2$), and lower or upper extremity injury, head injury, and laceration/puncture ($G3$). The input and output values for groups $G2$ and $G3$ are given in Section A6. The analysis of pairwise-oriented outcomes is provided in Section A7. In the main paper, we focus on the robust intervals of distances to the best physician, efficiencies, and ranks.

To derive the aggregated score- and rank-related results for three complaint groups, we conducted a robustness analysis for each of them individually and introduced a second level of certainty to capture the stability of outcomes for physicians treating patients from different groups. In Table 7, we present the extreme distances to the efficient physician, efficiency scores, and ranks obtained in that way. These marked as necessary (N) indicate the values obtained for all complaint groups, while the ones denoted as possible (P) specify the values obtained for at least one group.

The lower bound of the possible distance interval $[d^{P}_{*,o,S}, d^{*,P}_{o,S}]$ is equal to 0 for eight physicians: MD1, MD5, MD6, MD11, MD12, MD15, MD16, and MD20. These physicians perform efficiently, treating at least one complaint group. Moreover, MD6 is the only physician for whom the lower bound of the necessary distance interval $[d^{N}_{*,o,S}, d^{*,N}_{o,S}]$ is 0. This confirms its efficiency for all three complaint groups. The next two best results are attained by MD20 (0.0139) and MD13 (0.0257), which means that they are nearly efficient for all considered settings. In turn, for MD16, the intersection of the distances to the efficient physician over all groups is empty. Such an outcome indicates that MD16's performance strongly depends on the group. (S)he performed quite well for one group and all feasible input and output weights and poorly for some other group.

The possible-possible intervals of efficiency scores $[E^{P}_{*,o,S}, E^{*,P}_{o,S}]$ are wide for all physicians. The minimal width is for MD11 (0.6801), whereas the maximal difference between the extreme scores for different complaint groups is equal to 0.9319 (see MD1). When it comes to the width of the possible-necessary efficiency score interval $[E^{N}_{*,o,S}, E^{*,N}_{o,S}]$, it is minimal for MD12 (0.0128) and maximal for MD20 (0.4434). The physicians with the greatest width of the possible-necessary interval and the least width of the necessary-necessary interval are the most specialized ones, attaining highly variable results for different complaint groups.

Similar conclusion can be derived from the analysis of multiscenario rank intervals (see $[R^{P}_{*,o,S}, R^{*,P}_{o,S}]$ and $[R^{N}_{*,o,S}, R^{*,N}_{o,S}]$). The relative performance of MD2 and MD3 is rather poor for all complaint groups. Their best rank for any group is 13 and 9, respectively. For other physicians, the possible-possible rank intervals are rather wide, again confirming their varied performance. In particular, there are three physicians (MD1, MD11, and MD16) who attributed all ranks when considering the three complaint groups.

When considering the possible-necessary rank intervals $[R^{N}_{*,o,S}, R^{*,N}_{o,S}]$, we can observe that for MD10 and MD12, there is no single rank attained for all complaint groups. The best results are observed for MD5 who attained ranks in the interval [1, 5] for all scenarios. Similarly, in the most favorable scenario, MD16 and MD20 are ranked at least second ($R^{N}_{*,o,S} = 2$) for all complaint groups. On the contrary, MD1 is ranked only 12th for one group ($R^{N}_{*,o,S} = 12$). Given its efficiency for some other group ($R^{P}_{*,o,S} = 1$), this means that the performance of MD1 mostly depends on the selected priorities and evaluation scenario.

In Section A8, we summarize the results derived for each physician with the proposed robustness analysis framework for a single scenario referring to complaint group $G1$ and multiple scenarios concerning groups $G1$, $G2$, and $G3$. Specifically, we refer to the ranks attained by each physician according to various measures.

4. Conclusions and implications

We presented a novel robustness analysis framework for DEA incorporating a value-based additive efficiency model. The basic framework incorporates mathematical programming techniques and the Monte Carlo simulation to exploit all feasible input and output weights. These methods derive two types of results concerning four perspectives relevant to the analysis. One type of results, extreme outcomes, captures exact outcomes observed for the most and the least advantageous weight vectors for a given DMU or instantiated for all or at least one feasible weight vector. Another type

of results, stochastic acceptability indices, quantify the share of feasible weight vectors supporting some conclusions. The four accounted perspectives concern efficiency scores, distances from the efficient unit, ranks, and pairwise efficiency preference relations. Such outcomes provide rich information on the stability of efficiency outcomes from the complementary perspectives that focus on the DMUs assessed individually, compared pairwise, or collated with all remaining units in the analyzed set. To facilitate the application of these methods in practice, we created an open-source system implementing them on the *diviz* platform.

In addition, the primary framework was extended in two ways. On the one hand, we introduced the procedures for selecting the common vector of weights. These procedures incorporate robustness by exploiting stability analysis outcomes to define the score differences that should be emphasized in the ranking constructed with the chosen weight vector. One may either maximize the differences between efficiencies for pairs of DMUs for which an evident advantage of either of them can be observed given results attained for all feasible weight vectors or minimize such a difference if the results of such a comparison are not univocal. Specifically, we discussed the procedures exploiting the necessary efficiency preference relation, expected efficiencies, expected ranks, or *PEOI*s. On the other hand, we adjusted the robustness analysis framework to a multiscenario setting, in which the same DMUs are evaluated under different conditions or from various perspectives. The main innovation consisted of accounting for the second level of certainty, referring to the necessity or possibility of some robust conclusion given multiple relevant scenarios.

The proposed approach was applied to evaluating the performance of the ED physicians, assuming time, laboratory tests, and radiology orders as inputs, and rate of nonreturn visits to the ED within 72 hours as a single output that is a proxy for physicians' performance and the quality of the provided care. The robust results provide multiple implications for both individual physicians and hospital managers. Let us emphasize that due to the specificity of DEA, these conclusions are limited by considering a specific setup involving a particular group of analyzed peers, factors selected as relevant for the analysis, and an adopted efficiency model. Thus, they do not refer to any external standards.

First, the wide intervals of efficiency scores, distances to the efficient physician, and ranks, observed for most physicians for a single complaint group, serve as the evidence for the strong dependence of the physicians' performances on the selected weight vector (i.e., priorities assigned to different inputs and outputs). Such a high variability of results should make the analysts careful with some definitive judgments about the physicians' performance and might help identify the outliers. This variability also puts into question the results obtained with traditional DEA methods taking into account only the most advantageous scenario for each DMU, MCDA approaches, or composite indicators, due to their reliance on a single, often user-defined subjective weight vector or a limited subset of weight vectors.

Second, even though one should not draw strict conclusions about individual physicians' efficiency, the robust results serve as a good starting point for an in-depth investigation. In particular, these outcomes can be used to identify physicians who are markedly better in providing care to a given complaint group. These best performers should have low distances to the efficient physician, high efficiency scores or ranks for most feasible weight vectors, and not be outperformed by one other physician in terms of the necessary preference relation. The physicians satisfying these conditions may be considered a benchmark or "role models." A detailed analysis of their performances can facilitate developing an improvement plan and guidelines for the underperforming ones.

Third, to facilitate communicating the performance assessment results, we provide means for ranking the physicians. On the one hand, such ranking can be determined based on the expected efficiency scores or ranks. They offer an overview of the physicians' average performance (considering different weights), pointing out the overall good performers, niche performers, and lower-performing physicians. On the other hand, the rankings can be determined using a common vector of weights selected to represent the robust results attained for all feasible weight vectors. Such representative weights can also be interpreted as the priorities assigned to different inputs and outputs. They can be used in a practice-oriented model for a given complaint group.

Fourth, the results of robustness analysis are helpful in designing the corrective plans for underperforming physicians. In particular, the necessary preference relation can serve to construct the improvement paths based on the performance of other physicians, who outperform others. Hence, these outcomes may find application in a stepwise benchmarking process. Moreover, when referring to the robust results, the management may formulate detailed and diverse performance targets (e.g., improving inputs and outputs warranting a possible rank in the top three or the necessary preference over some other unit).

Fifth, the outcomes of multiscenario robustness analysis for different complaint groups are useful from individual physicians' and hospital managers' viewpoints. Specifically, we may identify physicians performing well given all complaint groups. They may be treated as universal benchmarks. Other physicians who performed well only for some complaint groups while underperforming for others may be considered "specialists," particularly efficient in managing patients of a given type. Overall, we observed a significant variability of results in the three complaint groups, indicating that medical practices and quality of care vary. From the managerial viewpoint, these outcomes help distinguish physicians into subsets treating patients with different complaints, which can positively affect the overall quality of care. They are also useful for identifying the most difficult complaint groups that are characterized by a low number of efficient physicians and a high number of inefficient physicians.

The main purpose of our research was to show the clinical management insights that can be gained from the robust analytical approach. These insights confirmed some hypotheses (e.g., on the differences between physicians in terms of their efficiencies and in the clinical judgments across the groups), supported common beliefs (e.g., that it is barely possible to excel at only all aspects of the clinical role), and provided answers to some performance-oriented questions (e.g., by identifying specialists or overall good performers). However, having such an approach actually applied in CHEO would require the Research Ethics Board approval and consent of the ED physicians, which was beyond the scope of this study.

Our model's main limitations come from the need to specify the marginal value functions and the lack of indicating precise performance improvements on the particular inputs or outputs that allow attaining efficiency. When it comes to the former, in MCDA, there exist some well-established techniques for eliciting such marginal functions. Moreover, such functions help differentiate between performances on a particular factor based on a given problem's specific features, a set of analyzed DMUs, and management preferences. If such a specification is not possible, one can use a default option of linear marginal value functions. As far as the required improvements are concerned, we instead opt for pointing out the peers from whom one should learn and improvement paths indicating the set of benchmarks.

Let us emphasize that when all components of the proposed methodology are employed simultaneously, the number of results to be considered by decision analysts can be prohibitively large. However, in the context of a real-world application, these components can be limited by accounting for the following three aspects. The first aspect refers to whether the performance of DMUs should be analyzed in single or multiple scenarios (in our paper, these scenarios corresponded to different complaint groups). The second aspect concerns the model exploitation by looking at the robustness of efficiency results or developing a univocal recommendation using a common set of weights emphasizing the robust outcomes. The last aspect refers to a type of output variability (extreme or stochastic) and a perspective on the efficiency analysis (scores, distances, ranks, or preference relations) that should be considered. Having answered such questions, one can limit the scope of the proposed methodological framework to one's own needs.

Several future research directions can be explored. From the application viewpoint, the most interesting one concerns extending the analysis to other complaint groups and more performance measures. In particular, the input- and output-oriented perspectives could be enriched by considering specialist consults and patient satisfaction, respectively. Such data were not available for our study. One could also analyze the impact of a trainee factor on physicians' performance by separately considering the visits when any trainee did not assist them, or junior or senior trainees supported them. From the methodological viewpoint, the proposed robustness analysis framework incorporating a value-based additive efficiency model can be extended to account for the imprecise (interval and ordinal) performances, the interactions between the considered factors, and a hierarchical structure of inputs and outputs.

Acknowledgments

The research of A. Labijak-Kowalska was supported by the Polish Ministry of Education and Science, grant no. 0311/SBAD/0709. M. Kadziński acknowledges financial support from the Polish National Science Center under the SONATA BIS project (grant no. DEC-2019/34/E/HS4/00045). L.C. Dias was supported by the Portuguese Foundation for Science and Technology (grant UIDB/05037/2020).

References

Akkan, C., Karadayi, M.A., Ekinci, Y., Ülengin, F., Uray, N., Karaosmanoğlu, E., 2020. Efficiency analysis of emergency departments in metropolitan areas. *Socio-Economic Planning Sciences* 69, 100679.

Amado, C., Santos, S., 2009. Challenges for performance assessment and improvement in primary health care: the case of the Portuguese health centres. *Health Policy* 91, 1, 43–56.

Andes, S., Metzger, L.M., Kralewski, J., Gans, D., 2002. Measuring efficiency of physician practices using data envelopment analysis. *Managed Care* 11, 11, 48.

Banker, R. D., Charnes, A., Cooper, W. W., 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science* 30(9), 1078-1092.

Basson, M.D., Butler, T., 2006. Evaluation of operating room suite efficiency in the veterans health administration system by using data envelopment analysis. *The American Journal of Surgery* 192, 5, 649–656.

Charnes, A., Cooper, W., Golany, B., Seiford, L., Stutz, J., 1985. Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. *Journal of Econometrics* 30, 1, 91–107.

Charnes, A., Cooper, W., Wei, Q., Huang, Z., 1989. Cone–ratio data envelopment analysis and multi-objective programming. *International Journal of Systems Science* 20, 1099–1118.

Charnes, A., Cooper, W.W., Rhodes, E., 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2, 6, 429–444.

Chen, Y., Wang, J., Zhu, J., Sherman, H.D., Chou, S.Y., 2019. How the great recession affects performance: a case of Pennsylvania hospitals using DEA. *Annals of Operations Research* 278, 1–2, 77–99.

Chilingerian, J.A., 1995. Evaluating physician efficiency in hospitals: a multivariate analysis of best practices. *European Journal of Operational Research* 80, 3, 548–574.

Chilingerian, J.A., Sherman, H.D., 1990. Managing physician efficiency and effectiveness in providing hospital services. *Health Services Management Research* 3, 1, 3–15.

Choi, S.O., 2005. Relative efficiency of fire and emergency services in Florida: an application and test of data envelopment analysis. *International Journal of Emergency Management* 2, 3, 218–230.

Ciomek, K., Kadziński, M., 2021. Polyrun: a Java library for sampling from the bounded convex polytopes. *SoftwareX* 13, 100659.

Contreras, I., 2020. A review of the literature on DEA models under common set of weights. *Journal of Modelling in Management* 15(4), 1277–1300.

Cook, W., Roll, Y., Kazakov, A., 1990. A DEA model for measuring the relative efficiency of highway maintenance patrols. *INFOR* 28, 113–124.

Färe, R., Grosskopf, S., 2000. Theory and application of directional distance functions. *Journal of Productivity Analysis* 13, 93–103.

Fiallos, J., Patrick, J., Michalowski, W., Farion, K., 2017. Using data envelopment analysis for assessing the performance of pediatric emergency department physicians. *Health Care Management Science* 20, 1, 129–140.

Flokou, A., Aletras, V., Niakas, D., 2017. A window-DEA based efficiency evaluation of the public hospital sector in Greece during the 5-year economic crisis. *PloS ONE* 12, 5, e0177946.

Gerami, J., Mavi, R.K., Saen, R.F., Mavi, N.K., 2020. A novel network DEA-R model for evaluating hospital services supply chain performance. *Annals of Operations Research* 1–26. https://doi.org/10.1007/s10479-020-03755-w.

Goddard, M., Jacobs, R., 2009. Using composite indicators to measure performance in health care. In Smith, P.C., Mossialos, E., Leatherman, S., Papanicolas, I. (eds) *Performance Measurement for Health System Improvement: Experiences, Challenges and Prospects*. Cambridge University Press, Cambridge, pp. 339–368.

Gouveia, M., Dias, L., Antunes, C., Mota, M., Duarte, E., Tenreiro, E., 2016. An application of value-based DEA to identify the best practices in primary health care. *OR Spectrum* 38, 3, 743–767.

Gouveia, M.C., Dias, L.C., Antunes, C.H., 2008. Additive DEA based on MCDA with imprecise information. *Journal of the Operational Research Society* 59, 1, 54–63.

Greco, S., Kadziński, M., Mousseau, V., Słowiński, R., 2012. Robust ordinal regression for multiple criteria group decision: UTA-GMS-GROUP and UTADIS-GMS-GROUP. *Decision Support Systems* 52, 3, 549–561.

Greco, S., Mousseau, V., Słowiński, R., 2008. Ordinal regression revisited: Multiple criteria ranking using a set of additive value functions. *European Journal of Operational Research* 191, 2, 416–436.

Hung, G., Chalut, D., 2008. A consensus-established set of important indicators of pediatric emergency department performance. 24, 9–15.

Jacobs, R., Goddard, M., Smith, P.C., 2005. How robust are hospital ranks based on composite performance measures? *Medical Care* pp. 1177–1184.

Jennings, N., Lee, G., Chao, K., Keating, S., 2009. A survey of patient satisfaction in a metropolitan emergency department: comparing nurse practitioners and emergency physicians. *International Journal of Nursing Practice* 15, 3, 213–218.

Johannessen, K.A., Kittelsen, S.A., Hagen, T.P., 2017. Assessing physician productivity following Norwegian hospital reform: a panel and data envelopment analysis. *Social Science & Medicine* 175, 117–126.

Kadziński, M., Greco, S., Słowiński, R., 2012a. Selection of a representative value function in robust multiple criteria ranking and choice. *European Journal of Operational Research* 217, 3, 541–553.

Kadziński, M., Greco, S., Słowiński, R., 2012b. Extreme ranking analysis in robust ordinal regression. *Omega* 40, 3, 488–501.

Kadziński, M., Labijak, A., Napieraj, M., 2017. Integrated framework for robustness analysis using ratio-based efficiency model with application to evaluation of Polish airports. *Omega* 67, 1–18.

Kang, H., Nembhard, H., DeFlitch, C., Pasupathy, K., 2017. Assessment of emergency department efficiency using data envelopment analysis. *IISE Transactions on Healthcare Systems Engineering* 7, 4, 236–246.

Keeney, R.L., Raiffa, H., 1993. *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press, Cambridge.

Ketabi, S., Teymouri, E., Ketabi, M., 2018. Efficiency measurement of emergency departments in Isfahan, Iran. *International Journal of Process Management and Benchmarking* 8, 2, 142–155.

Khushalani, J., Ozcan, Y.A., 2017. Are hospitals producing quality care efficiently? An analysis using dynamic network data envelopment analysis (DEA). *Socio-Economic Planning Sciences* 60, 15–23.

Kohl, S., Schoenfelder, J., Fügener, A., Brunner, J.O., 2019. The use of data envelopment analysis (DEA) in healthcare with a focus on hospitals. *Health Care Management Science* 22, 2, 245–286.

Kooreman, P., 1994. Nursing home care in The Netherlands: a nonparametric efficiency analysis. *Journal of Health Economics* 13, 3, 301–316.

Küçük, A., Özsoy, V.S., Balkan, D., 2020. Assessment of technical efficiency of public hospitals in Turkey. *European Journal of Public Health* 30, 2, 230–235.

Kuwahara, Y., Nagata, S., Taguchi, A., Naruse, T., Kawaguchi, H., Murashima, S., 2013. Measuring the efficiencies of visiting nurse service agencies using data envelopment analysis. *Health Care Management Science* 16, 3, 228–235.

Labijak-Kowalska, A., Kadziński, M., 2021. Experimental comparison of results provided by ranking methods in data envelopment analysis. *Expert Systems with Applications* 170, 114739.

Lahdelma, R., Salminen, P., 2006. Stochastic multicriteria acceptability analysis using the data envelopment model. *European Journal of Operational Research* 173, 1, 241–252.

Lee, R.H., Bott, M.J., Gajewski, B., Taunton, R.L., 2009. Modeling efficiency at the process level: an examination of the care planning process in nursing homes. *Health Services Research* 44, 1, 15–32.

Liu, J.S., Lu, L.Y., Lu, W.M., Lin, B.J., 2013. A survey of DEA applications. *Omega* 41, 5, 893–902.

Meyer, P., Bigaret, S., 2012. Diviz: a software for modeling, processing and sharing algorithmic workflows in MCDA. *Intelligent Decision Technologies* 6, 4, 283–296.

Ozcan, Y.A., Jiang, H., Pai, C.W., 2000. Do primary care physicians or specialists provide more efficient care? *Health Services Management Research* 13, 2, 90–96.

Rouyendegh, B.D., Oztekin, A., Ekong, J., Dag, A., 2019. Measuring the efficiency of hospitals: a fully-ranking DEA–FAHP approach. *Annals of Operations Research* 278, 1, 361–378.

Salo, A., Punkka, A., 2011. Ranking intervals and dominance relations for ratio-based efficiency analysis. *Management Science* 57, 1, 200–214.

Schang, L., Hynninen, Y., Morton, A., Salo, A., 2016. Developing robust composite measures of healthcare quality-ranking intervals and dominance relations for Scottish Health Boards. *Social Science & Medicine* 162, 59–67.

Shimshak, D.G., Lenard, M.L., Klimberg, R.K., 2009. Incorporating quality into data envelopment analysis of nursing home performance: a case study. *Omega* 37, 3, 672–685.

Siddharthan, K., Ahern, M., Rosenman, R., 2000. Data envelopment analysis to determine efficiencies of health maintenance organizations. *Health Care Management Science* 3, 1, 23–29.

Smith, C.A., Varkey, A.B., Evans, A.T., Reilly, B.M., 2004. Evaluating the performance of inpatient attending physicians. *Journal of General Internal Medicine* 19, 7, 766–771.

Testi, A., Fareed, N., Ozcan, Y.A., Tanfani, E., 2013. Assessment of physician performance for diabetes: a bias-corrected data envelopment analysis model. *Quality in Primary Care* 21, 6, 345–357.

Thau, M., Mikkelsen, M.F., Hjortskov, M., Pedersen, M.J., 2020. Question order bias revisited: a split-ballot experiment on satisfaction with public services among experienced and professional users. *Public Administration* 99, 189–204.

Thompson, R., Langemeier, L., Lee, C., Lee, E., Thrall, R., 1990. The role of multiplier bounds in efficiency analysis with application to kansas farming. *Journal of Econometrics* 46, 93–108.

Tosun, Ö., 2012. Using data envelopment analysis—neural network model to evaluate hospital efficiency. *International Journal of Productivity and Quality Management* 9, 2, 245–257.

Veloso, A.S., Vaz, C.B., Alves, J., 2018. determinants of nursing homes performance: the case of portuguese santas casas da misericórdia. In Vaz, A.I.F., Almeida, J.P., Oliveira, J.F., Pinto, A.A. (eds) *Operational Research*. Springer, Cham, pp. 393–409.

Wagner, J.M., Shimshak, D.G., 2000. Physician profiling using data envelopment analysis: a case study. *International Journal of Healthcare Technology and Management* 2, 1–4, 358–374.

Winkler, R.L. and Hay, W.L., 1985. Statistics: probability, inference, and decision. Rinehart & Winston, New York.

Zehra, Ö., Serpil, S., 2018. Evaluating healthcare system efficiency of OECD countries: a DEA-based study. In Kahraman, C., Ilker Topcu, Y. (eds) *Operations Research Applications in Health Care Management*. Springer, Cham, pp. 141–158.

Appendix

A.1. Computing the stochastic acceptability indices: an illustrative example

In this section, we discuss how to estimate the distribution of distances to the efficient unit and how to compute the ranks of physicians based on the expected efficiency, distance, or rank. We apply the hit-and-run algorithm to derive samples of weights for all inputs and outputs. Table A1 shows 10 examples of weight vectors used to compute the illustrative results in this section. Note that the outcomes reported in the main paper are derived from the analysis of 10,000 samples, which offers sufficient precision of the estimation.

Then, we compute a value-based efficiency score for each physician and each sample (see Table A2). When considering MD_i, its distance to the efficient unit is calculated as the difference between the maximal efficiency score of any physician obtained for a given sample and the efficiency score of MD_i. For example, for sample 1 and MD3, such a distance is equal to $d_3 = 0.275 - 0.066 = 0.209$. An efficiency rank of MD_i is computed based on the number of physicians with greater efficiencies than MD_i. For example, for sample 1, there are three physicians (MD6, MD11, and MD20) ranked better than MD5, and hence it is ranked fourth. The distances to the efficient unit and efficiency ranks for all physicians and samples are provided in Table A2.

Having computed the distances to the efficient unit for each decision-making unit (DMU) and each sample, we calculate $DAII$ as the ratio of the number of samples for which the distance lies within the analyzed interval to the number of all considered samples (see Table A3). For example, $DAII(MD1, (0.1, 0.2])$ is equal to 0.3 because for MD1, its distance to the efficient unit is in the (0.1,0.2] interval for 3 of 10 samples (samples 2, 5, and 9). The distributions of efficiency scores ($EAII$s), ranks ($ERAI$s), and preference relations ($PEOI$s) are computed analogously.

The results obtained for various samples can be averaged to estimate the expected measure values. The expected efficiencies EE, distances Ed, and ranks ER are presented in Table A2. To impose a

Table A1
Ten examples of input and output weight vectors obtained with the Monte Carlo simulation (for each vector, the weights sum up to 1)

	1	2	3	4	5	6	7	8	9	10
w_{i_1}	0.285	0.185	0.348	0.215	0.440	0.060	0.325	0.324	0.268	0.162
w_{i_2}	0.025	0.456	0.304	0.135	0.158	0.296	0.050	0.258	0.051	0.062
w_{i_3}	0.499	0.016	0.301	0.376	0.142	0.471	0.383	0.355	0.474	0.289
w_{o_1}	0.191	0.344	0.047	0.274	0.261	0.174	0.242	0.063	0.207	0.487

Table A2
Efficiency scores E, distances d, and ranks R for the considered physicians obtained and 10 examples of weight vectors

Sample		MD																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	E	0.216	0.083	0.066	0.116	0.252	0.275	0.134	0.146	0.172	0.064	0.262	0.109	0.218	0.126	0.203	0.177	0.126	0.175	0.199	0.261
	d	0.059	0.192	0.209	0.159	0.023	0.000	0.141	0.128	0.103	0.211	0.013	0.166	0.057	0.149	0.071	0.098	0.149	0.100	0.076	0.014
	R	6	18	19	16	4	1	13	12	11	20	2	17	5	15	7	9	14	10	8	3
2	E	0.435	0.281	0.243	0.417	0.527	0.611	0.419	0.438	0.464	0.401	0.595	0.411	0.582	0.378	0.484	0.480	0.406	0.536	0.471	0.547
	d	0.176	0.331	0.368	0.194	0.084	0.000	0.193	0.173	0.147	0.210	0.017	0.201	0.029	0.233	0.128	0.131	0.205	0.075	0.140	0.064
	R	12	19	20	14	6	1	13	11	10	17	2	15	3	18	7	8	16	5	9	4
3	E	0.122	0.188	0.188	0.310	0.478	0.462	0.315	0.359	0.359	0.281	0.411	0.283	0.394	0.305	0.411	0.322	0.309	0.346	0.418	0.476
	d	0.356	0.290	0.291	0.168	0.000	0.016	0.163	0.120	0.120	0.197	0.067	0.196	0.085	0.173	0.067	0.156	0.169	0.132	0.060	0.002
	R	20	18	19	13	1	3	12	9	8	17	5	16	7	15	6	11	14	10	4	2
4	E	0.314	0.137	0.112	0.192	0.309	0.358	0.207	0.213	0.243	0.150	0.354	0.192	0.316	0.183	0.265	0.267	0.199	0.276	0.255	0.321
	d	0.045	0.221	0.247	0.166	0.050	0.000	0.151	0.146	0.116	0.208	0.004	0.166	0.043	0.175	0.093	0.091	0.159	0.083	0.103	0.037
	R	5	19	20	15	6	1	13	12	11	18	2	16	4	17	9	8	14	7	10	3
5	E	0.312	0.155	0.109	0.223	0.384	0.460	0.226	0.278	0.302	0.175	0.412	0.191	0.395	0.243	0.339	0.242	0.202	0.323	0.335	0.419
	d	0.148	0.305	0.351	0.237	0.076	0.000	0.234	0.182	0.157	0.284	0.048	0.269	0.065	0.217	0.121	0.218	0.258	0.137	0.125	0.041
	R	9	19	20	15	5	1	14	11	10	18	3	17	4	12	6	13	16	8	7	2
6	E	0.236	0.187	0.193	0.290	0.390	0.384	0.308	0.292	0.309	0.268	0.394	0.304	0.352	0.246	0.334	0.381	0.315	0.338	0.328	0.374
	d	0.157	0.207	0.201	0.104	0.004	0.009	0.086	0.101	0.084	0.126	0.000	0.089	0.041	0.147	0.060	0.013	0.079	0.055	0.066	0.020
	R	18	20	19	15	2	3	12	14	11	16	1	13	6	17	8	4	10	7	9	5
7	E	0.272	0.102	0.072	0.139	0.278	0.327	0.153	0.175	0.203	0.087	0.307	0.126	0.270	0.152	0.233	0.192	0.139	0.218	0.227	0.298
	d	0.055	0.225	0.255	0.188	0.049	0.000	0.174	0.152	0.124	0.240	0.020	0.201	0.057	0.175	0.094	0.135	0.188	0.109	0.100	0.029
	R	5	18	20	16	4	1	13	12	10	19	2	17	6	14	7	11	15	9	8	3
8	E	0.129	0.170	0.169	0.277	0.439	0.424	0.285	0.321	0.325	0.245	0.381	0.255	0.358	0.273	0.374	0.300	0.280	0.313	0.379	0.436
	d	0.309	0.269	0.270	0.161	0.000	0.015	0.153	0.118	0.114	0.194	0.058	0.184	0.081	0.166	0.065	0.139	0.158	0.125	0.060	0.003
	R	20	18	19	14	1	3	12	9	8	17	4	16	7	15	6	11	13	10	5	2
9	E	0.235	0.096	0.077	0.134	0.265	0.293	0.151	0.162	0.188	0.084	0.282	0.129	0.239	0.139	0.218	0.198	0.144	0.197	0.212	0.275
	d	0.058	0.198	0.216	0.159	0.028	0.000	0.142	0.131	0.105	0.209	0.011	0.164	0.054	0.154	0.075	0.095	0.150	0.096	0.081	0.018
	R	6	18	20	16	4	1	13	12	11	19	2	17	5	15	7	9	14	10	8	3
10	E	0.509	0.128	0.072	0.145	0.229	0.353	0.163	0.152	0.206	0.096	0.374	0.158	0.329	0.136	0.206	0.256	0.147	0.284	0.183	0.264
	d	0.000	0.381	0.438	0.364	0.281	0.156	0.347	0.357	0.303	0.414	0.135	0.351	0.180	0.373	0.303	0.253	0.363	0.225	0.327	0.246
	R	1	18	20	16	8	3	12	14	9	19	2	13	4	17	10	7	15	5	11	6
	EE	0.278	0.153	0.130	0.224	0.355	0.395	0.236	0.254	0.277	0.185	0.377	0.216	0.345	0.218	0.307	0.282	0.227	0.301	0.301	0.367
	Ed	0.136	0.262	0.284	0.190	0.059	0.020	0.178	0.161	0.137	0.229	0.037	0.199	0.069	0.196	0.108	0.133	0.188	0.114	0.114	0.047
	ER	10.2	18.5	19.6	15.0	4.1	1.8	12.7	11.6	9.9	18.0	2.5	15.7	5.1	15.5	7.3	9.1	14.1	8.1	7.9	3.3

These results are used to estimate the expected efficiencies EE, distances Ed, and ranks ER.

Table A3
Distribution of the distances to the efficient unit (*DAII*s) based on 10 examples of weight vectors

MD	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
1	0.5	0.3	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.2	0.5	0.3	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.7	0.2	0.1	0.0	0.0	0.0	0.0	0.0
4	0.0	0.8	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
5	0.9	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	0.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.1	0.7	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0
8	0.0	0.9	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
9	0.1	0.8	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
10	0.0	0.3	0.6	0.0	0.1	0.0	0.0	0.0	0.0	0.0
11	0.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	0.1	0.5	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0
13	0.9	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	0.0	0.7	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
15	0.7	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
16	0.4	0.4	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
17	0.1	0.6	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0
18	0.4	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
19	0.5	0.4	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0
20	0.9	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0

complete order on the set of physicians, they need to be sorted accordingly (e.g., in the ascending order when accounting for Ed). In the considered example, the best (minimal) expected distance is associated with MD6 (0.020) and the worst (maximal) distance is attained by MD3 (0.284). These physicians are ranked at top and bottom, respectively. The rankings based on the expected ranks (*ER*s) or efficiencies can be constructed analogously. Note, however, that while lower distances and ranks are preferred, greater values are more favorable when considering the efficiency scores.

A.2. Descriptive statistics of input and output data for the three considered complaint groups

In Table A4, we report the descriptive statistics of input and output data for the three complaint groups considered in the main paper: $G1$—abdominal pain and constipation; $G2$—fever; and $G3$—lower or upper extremity injury, head injury, and laceration/puncture.

A.3. Distributions of the distances to the efficient unit and the efficiency scores for complaint group G1

In Tables A5 and A6, we report the distributions of the distances to the efficient unit and the efficiency scores for complaint group $G1$ estimated based on 10,000 weight vectors. They are captured

Table A4
Descriptive statistics of input and output data for the three considered complaint groups ($G1$, $G2$, and $G3$)

Group	Statistic	i_1 – AVG_MDTIME_PAT	i_2 – AVG_LAB_PAT	i_3 – AVG_RAD_PAT	o_1 – RATE_NR72
$G1$	Min	1.435	0.857	0.462	0.905
	Max	2.311	2.760	0.920	1.000
	Mean	1.811	1.739	0.656	0.958
	St. dev. SD	0.254	0.431	0.112	0.022
$G2$	Min	1.017	0.357	0.207	0.907
	Max	1.752	1.101	0.419	1.000
	Mean	1.367	0.668	0.322	0.963
	SD	0.227	0.206	0.061	0.020
$G2$	Min	0.836	0.000	0.478	0.957
	Max	1.293	0.176	0.847	1.000
	Mean	1.058	0.071	0.684	0.985
	SD	0.132	0.055	0.090	0.010

Table A5
Distribution of the distances to the efficient unit ($DAII$s)

MD	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
1	0.405	0.237	0.187	0.120	0.051	0.000	0.000	0.000	0.000	0.000
2	0.000	0.017	0.617	0.365	0.001	0.000	0.000	0.000	0.000	0.000
3	0.000	0.005	0.423	0.518	0.054	0.000	0.000	0.000	0.000	0.000
4	0.004	0.598	0.344	0.054	0.000	0.000	0.000	0.000	0.000	0.000
5	0.750	0.196	0.050	0.004	0.000	0.000	0.000	0.000	0.000	0.000
6	0.955	0.043	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	0.027	0.618	0.309	0.046	0.000	0.000	0.000	0.000	0.000	0.000
8	0.008	0.781	0.170	0.041	0.000	0.000	0.000	0.000	0.000	0.000
9	0.072	0.834	0.085	0.009	0.000	0.000	0.000	0.000	0.000	0.000
10	0.000	0.365	0.493	0.122	0.020	0.000	0.000	0.000	0.000	0.000
11	0.965	0.035	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
12	0.022	0.470	0.434	0.074	0.000	0.000	0.000	0.000	0.000	0.000
13	0.891	0.103	0.006	0.000	0.000	0.000	0.000	0.000	0.000	0.000
14	0.000	0.515	0.420	0.063	0.002	0.000	0.000	0.000	0.000	0.000
15	0.495	0.420	0.076	0.009	0.000	0.000	0.000	0.000	0.000	0.000
16	0.295	0.512	0.191	0.002	0.000	0.000	0.000	0.000	0.000	0.000
17	0.044	0.511	0.373	0.071	0.001	0.000	0.000	0.000	0.000	0.000
18	0.418	0.553	0.029	0.000	0.000	0.000	0.000	0.000	0.000	0.000
19	0.457	0.432	0.090	0.021	0.000	0.000	0.000	0.000	0.000	0.000
20	0.853	0.119	0.028	0.000	0.000	0.000	0.000	0.000	0.000	0.000

by distance acceptability interval indices $DAII$s, and efficiency acceptability interval indices, $EAII$s, respectively. These results are referred to in Section 3.1 of the main paper.

The analysis of such distributions allows identifying the DMUs for which the results vary much in the set of feasible weights. High dispersion of scores and distances should prompt investigation as to whether the guidelines for standard practice can be used to reduce variance in management.

Table A6
Distribution of the efficiency scores ($EAII$s)

MD	[0.0, 0.1]	(0.1, 0.2]	(0.2, 0.3]	(0.3, 0.4]	(0.4, 0.5]	(0.5, 0.6]	(0.6, 0.7]	(0.7, 0.8]	(0.8, 0.9]	(0.9, 1.0]
1	0.030	0.183	0.230	0.265	0.206	0.086	0.000	0.000	0.000	0.000
2	0.040	0.533	0.427	0.000	0.000	0.000	0.000	0.000	0.000	0.000
3	0.199	0.485	0.316	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.000	0.233	0.332	0.325	0.110	0.000	0.000	0.000	0.000	0.000
5	0.000	0.012	0.152	0.318	0.320	0.187	0.011	0.000	0.000	0.000
6	0.000	0.000	0.030	0.250	0.399	0.291	0.030	0.000	0.000	0.000
7	0.000	0.197	0.348	0.335	0.120	0.000	0.000	0.000	0.000	0.000
8	0.001	0.150	0.335	0.321	0.188	0.005	0.000	0.000	0.000	0.000
9	0.000	0.063	0.324	0.371	0.239	0.003	0.000	0.000	0.000	0.000
10	0.098	0.274	0.282	0.277	0.069	0.000	0.000	0.000	0.000	0.000
11	0.000	0.000	0.032	0.313	0.402	0.244	0.009	0.000	0.000	0.000
12	0.019	0.243	0.341	0.312	0.085	0.000	0.000	0.000	0.000	0.000
13	0.000	0.001	0.118	0.348	0.355	0.175	0.003	0.000	0.000	0.000
14	0.009	0.216	0.423	0.332	0.020	0.000	0.000	0.000	0.000	0.000
15	0.000	0.029	0.253	0.369	0.309	0.040	0.000	0.000	0.000	0.000
16	0.000	0.075	0.304	0.331	0.267	0.023	0.000	0.000	0.000	0.000
17	0.004	0.236	0.332	0.325	0.103	0.000	0.000	0.000	0.000	0.000
18	0.000	0.018	0.231	0.380	0.307	0.064	0.000	0.000	0.000	0.000
19	0.000	0.056	0.262	0.355	0.285	0.042	0.000	0.000	0.000	0.000
20	0.000	0.006	0.099	0.322	0.364	0.196	0.013	0.000	0.000	0.000

In our study, the example units for which such verification should be carried out are MD1, MD8, MD12, MD17, and MD19.

*A.4. Analysis of efficiency ranks for complaint group G*1

In this section, we discuss the robustness of efficiency ranks for complaint group $G1$. The distances to the efficient DMU and efficiency scores are derived from the cardinal-oriented comparison of physicians. In turn, efficiency ranks build on the ordinal comparisons between the physicians. In Table A7, we report the extreme (R_* and R^*) and expected (ER) ranks. The physicians identified as efficient have the best ranks equal to 1. Based on R_*, MD13 is the best among the inefficient units. (S)he is ranked second in the best case $R_* = 2$), which means that in the most favorable scenario, it is less efficient only than a single efficient MD, while attaining better scores than the remaining 18 physicians. MD2 and MD3 have the least positive results in terms of R_*. For these physicians, there are at least 13 and 17 other physicians in a group who are more efficient for any feasible weight vector.

The analysis of the worst efficiency ranks (R^*) indicates that four efficient physicians (MD5, MD6, MD11, and MD20) never fall out of the top eight. Thus, the stability of derived ranks is the highest for MD6 because even in the least favorable scenario, only four other physicians attain better efficiencies. The performance of the other two efficient physicians is less stable. In particular,

Table A7
The extreme and expected ranks, and efficiency rank acceptability indices ($ERAI$s) for the considered physicians

MD	R_*	R^*	ER	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	20	11.041	0.166	0.040	0.055	0.039	0.028	0.039	0.023	0.045	0.031	0.017	0.037	0.028	0.005	0.014	0.011	0.024	0.028	0.138	0.023	0.209
2	14	20	18.669	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.013	0.017	0.261	0.691	0.015
3	18	20	19.640	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.121	0.118	0.761
4	11	18	14.409	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.026	0.069	0.081	0.270	0.405	0.146	0.003	0.000	0.000
5	1	8	4.541	0.145	0.065	0.179	0.077	0.147	0.136	0.138	0.113	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
6	1	5	1.860	0.466	0.228	0.290	0.012	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	9	15	12.798	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.001	0.024	0.335	0.459	0.166	0.013	0.000	0.000	0.000	0.000	0.000
8	8	17	11.693	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.056	0.094	0.138	0.209	0.253	0.045	0.066	0.055	0.078	0.006	0.000	0.000	0.000
9	8	16	9.722	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.105	0.275	0.465	0.120	0.021	0.012	0.001	0.001	0.000	0.000	0.000	0.000	0.000
10	13	20	17.347	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.037	0.231	0.262	0.283	0.168	0.015
11	1	8	2.914	0.148	0.344	0.153	0.237	0.046	0.063	0.009	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
12	9	18	15.150	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.071	0.041	0.073	0.126	0.170	0.195	0.314	0.007	0.000	0.000
13	2	14	4.763	0.000	0.000	0.181	0.328	0.256	0.058	0.142	0.031	0.002	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
14	10	19	15.339	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.088	0.057	0.133	0.080	0.089	0.064	0.307	0.180	0.000	0.000
15	5	11	7.412	0.000	0.000	0.000	0.000	0.069	0.212	0.244	0.227	0.210	0.037	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
16	1	16	9.110	0.000	0.015	0.027	0.034	0.065	0.060	0.112	0.119	0.079	0.077	0.194	0.099	0.054	0.044	0.015	0.006	0.000	0.000	0.000	0.000
17	6	18	14.347	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.008	0.031	0.035	0.077	0.122	0.206	0.201	0.243	0.063	0.010	0.000	0.000
18	4	14	7.385	0.000	0.000	0.000	0.034	0.117	0.238	0.138	0.198	0.111	0.126	0.031	0.005	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
19	4	15	8.178	0.000	0.000	0.000	0.099	0.086	0.107	0.104	0.099	0.188	0.101	0.164	0.015	0.014	0.023	0.000	0.000	0.000	0.000	0.000	0.000
20	1	8	3.682	0.075	0.308	0.115	0.140	0.182	0.087	0.090	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

MD16 is ranked 16th in the worst case, whereas MD1 is ranked at the very bottom. There are only three other physicians (MD2, MD3, and MD10) ranked 20th for at least one feasible weight vector.

The analysis of extreme efficiency ranks can be enriched with consideration of the $ERAI$s (see Table A7), indicating for each physician the distribution of ranks over the feasible weight vectors. For some physicians, the derived ranks are relatively stable. For example, MD3 is ranked at the very bottom for 76.1% weights and MD2 is ranked 18th or 19th for 95.2% samples. MD6 is ranked at the top for 46.6% weight, making him/her the most efficient physician in the group. In general, such a high value for the first rank acceptability index may indicate the outlier DMU. It may motivate the management to investigate the results without considering such an overall good performer who influences the distances of many other DMUs.

As far as MD13 is concerned, its possible rank interval is relatively wide [2, 14]. However, for 96.5% feasible weights, it is ranked in the top seven. For some other physicians, the ranks are more distributed. In particular, the $ERAI$s for MD1 are positive for all ranks with $ERAI(MD1, 1)$ (16.6%) being close to $ERAI(MD1, 20)$ (20.9%). This means that, depending on the chosen input/output weights, it is almost equally likely for MD1 to be ranked at the top or at the bottom. A similar distribution of ranks can be observed for MD16. For this physician, $ERAI$s are nonzero for ranks between 2 and 16, with the greatest one being lower than 0.2.

The ERs (see Table A7) can also be used to order all physicians. The top-ranked physicians are MD6 ($ER = 1.860$) and MD11 ($ER = 2.914$), whereas the bottom-ranked physicians are MD2 ($ER = 18.669$) and MD3 ($ER = 19.640$). The ranking determined by ERs is very similar to the orders imposed by Eds and EEs. The swaps occur only for two pairs, (MD5, MD13) and (MD17, MD4), which confirms the stability of conclusions derived from the multiperspective robustness analysis.

In general, the expected results exhibit which units perform good or bad for different priorities assigned to inputs and outputs. In some situations, the expected efficiencies or ranks of inefficient units can be, on average, better than for some efficient units (see, e.g., the average ranks of inefficient MD13 and MD15 compared to the expected positions for the efficient MD1 and MD16). Such results may indicate the need to implement the corrective actions for the average bad performers who prove to be efficient only under specific scenarios.

*A.5. Analysis of pairwise preference relations for complaint group G*1

Another aspect considered in the robustness analysis concerns pairwise comparisons between physicians. The Hasse diagram of the necessary preference relation is presented in Fig. A1. No physician is necessarily preferred over the six efficient physicians. However, there is also one inefficient physician (MD13) who is not necessarily worse than any other physician (depending on the weights, the physicians performing better than MD13 are not the same). Overall, MD5, MD6, and MD20 are necessarily preferred to the largest number of other physicians (12), which confirms their superior performance. On the other hand, MD1, MD2, MD3, and MD10 are not necessarily preferred to any other physician. MD1 can be seen as a potential outlier because it is neither necessarily better nor worse than any other physician.

The graph of the necessary preference relation can be used for constructing the corrective actions and improvement paths for inefficient physicians. From a short-term perspective, one can focus on

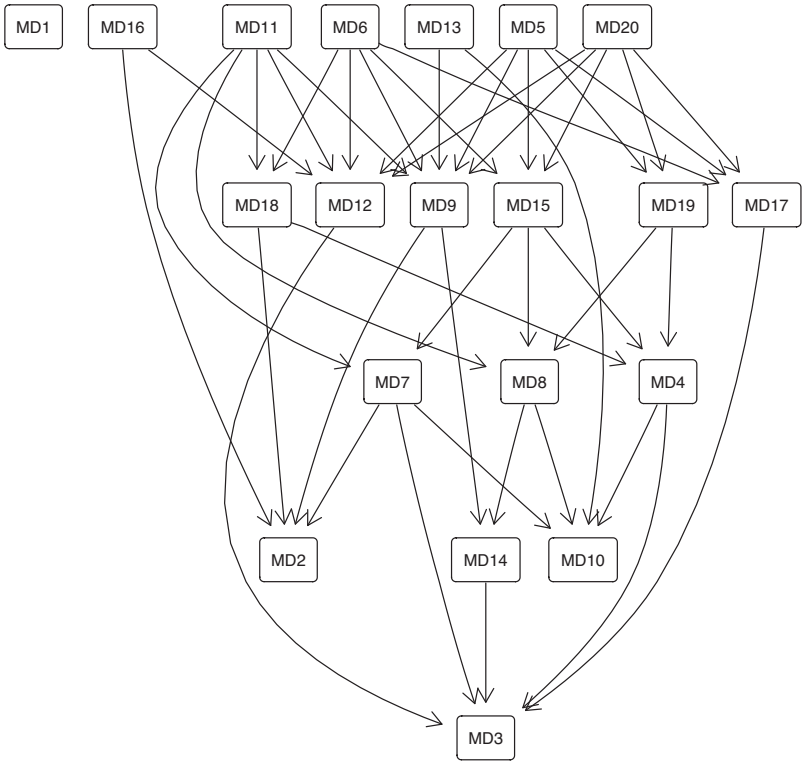

Fig. A1. The Hasse diagram of the necessary efficiency preference relation $\succsim_{\sim E}^{N}$.

the units that are necessarily preferred over some inefficient DMUs. For example, for MD8, these can be MD11, MD15, or MD19. The differences in inputs and outputs for such units indicate the improvement potential. From a long-term perspective, one can apply the stepwise benchmarking based on the paths observed in the Hasse diagram of $\succsim_{\sim E}^{N}$. For example, MD3—ranked at the bottom—can improve by following some improvement paths, for example, (MD14, MD8, MD19, MD5) or (MD7, MD15, MD20).

For pairs of physicians who are incomparable in terms of $\succsim_{\sim E}^{N}$, the efficiency comparison results are not univocal, given all feasible weights. Such pairs are not connected by an arc in Fig. A1. The shares of feasible weights confirming one physician's better performance over another are captured by $PEOI$s (see Table A8). For some other pairs, one physician performs clearly better, for example, $PEOI(MD16, MD17)$—0.980 indicates that for 98% of feasible weights, MD16 is at least as efficient as MD17. Thus, even if the preference relation is not fully robust for this pair, it is close to being so. Similar conclusions can be drawn for (MD18, MD12), (MD13, MD7), and (MD8, MD2). For some pairs of physicians these shares are more balanced, for example, for (MD13, MD5)—$PEOI(MD13, MD5)$—0.513 and $PEOI(MD5, MD13)$—0.487. Similar observations apply to (MD17, MD4) or (MD18, MD15).

The remaining DMUs do not influence such pairwise comparisons. The analyst may be interested in such a one-on-one perspective if (s)he knows some units better than others. Then, they can be

Table A8
Pairwise efficiency outranking indices ($PEOIs$) for all pairs of physicians

MD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1.000	0.779	0.794	0.563	0.357	0.209	0.538	0.503	0.463	0.624	0.215	0.579	0.272	0.580	0.424	0.440	0.560	0.377	0.437	0.324
2	0.221	1.000	0.883	0.000	0.000	0.000	0.000	0.002	0.000	0.182	0.000	0.013	0.000	0.013	0.000	0.000	0.013	0.000	0.000	0.000
3	0.206	0.117	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.012	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
4	0.437	1.000	1.000	1.000	0.000	0.000	0.104	0.104	0.000	1.000	0.000	0.664	0.000	0.679	0.000	0.072	0.550	0.000	0.000	0.000
5	0.643	1.000	1.000	1.000	1.000	0.191	1.000	1.000	1.000	1.000	0.377	1.000	0.487	1.000	1.000	0.817	1.000	0.701	1.000	0.212
6	0.791	1.000	1.000	1.000	0.809	1.000	1.000	1.000	1.000	1.000	0.727	1.000	1.000	1.000	1.000	0.982	1.000	1.000	1.000	0.806
7	0.462	1.000	1.000	0.896	0.000	0.000	1.000	0.262	0.019	1.000	0.000	0.879	0.001	0.753	0.000	0.088	0.873	0.004	0.038	0.000
8	0.497	0.998	1.000	0.896	0.000	0.000	0.738	1.000	0.089	1.000	0.000	0.792	0.000	1.000	0.000	0.349	0.791	0.150	0.000	0.000
9	0.537	1.000	1.000	1.000	0.000	0.000	0.981	0.911	1.000	1.000	0.000	0.979	0.000	1.000	0.034	0.485	0.955	0.179	0.265	0.000
10	0.376	0.818	0.988	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.140	0.000	0.313	0.000	0.006	0.007	0.000	0.000	0.000
11	0.785	1.000	1.000	1.000	0.623	0.273	1.000	1.000	1.000	1.000	1.000	1.000	0.986	1.000	0.916	0.988	1.000	1.000	0.888	0.571
12	0.421	0.987	1.000	0.336	0.000	0.000	0.121	0.208	0.021	0.860	0.000	1.000	0.000	0.527	0.000	0.000	0.326	0.001	0.055	0.000
13	0.728	1.000	1.000	1.000	0.513	0.000	0.999	1.000	1.000	1.000	0.014	1.000	1.000	1.000	0.817	0.918	0.996	1.000	0.810	0.441
14	0.420	0.987	1.000	0.321	0.000	0.000	0.247	0.000	0.000	0.687	0.000	0.473	0.000	1.000	0.000	0.134	0.392	0.007	0.000	0.000
15	0.576	1.000	1.000	1.000	0.000	0.000	1.000	1.000	0.966	1.000	0.084	1.000	0.183	1.000	1.000	0.636	1.000	0.458	0.706	0.000
16	0.560	1.000	1.000	0.928	0.183	0.018	0.912	0.651	0.515	0.994	0.012	1.000	0.082	0.866	0.364	1.000	0.980	0.221	0.409	0.143
17	0.440	0.987	1.000	0.450	0.000	0.000	0.127	0.209	0.045	0.993	0.000	0.674	0.004	0.608	0.000	0.020	1.000	0.016	0.030	0.000
18	0.623	1.000	1.000	1.000	0.299	0.000	0.996	0.850	0.821	1.000	0.000	0.999	0.000	0.993	0.542	0.779	0.984	1.000	0.571	0.176
19	0.563	1.000	1.000	1.000	0.000	0.000	0.962	1.000	0.735	1.000	0.112	0.945	0.190	1.000	0.294	0.591	0.970	0.429	1.000	0.000
20	0.676	1.000	1.000	1.000	0.788	0.194	1.000	1.000	1.000	1.000	0.429	1.000	0.559	1.000	1.000	0.857	1.000	0.824	1.000	1.000

Table A9
Input and output values for the complaint groups $G2$ (fever) and $G3$ (lower or upper extremity injury, head injury, and laceration/puncture) by physician

Group	$G2$				$G3$			
MD	i_1	i_2	i_3	o_1	i_1	i_2	i_3	o_1
MD1	1.639	0.604	0.333	1.000	1.293	0 .000	0.699	0.957
MD2	1.682	1.031	0.374	0.969	1.287	0.166	0.847	0.983
MD3	1.386	0.551	0.318	0.907	1.123	0.030	0.723	0.970
MD4	1.482	0.600	0.419	0.943	1.122	0.115	0.803	1.000
MD5	1.362	0.561	0.305	0.952	1.050	0.021	0.609	0.979
MD6	1.017	0.496	0.207	0.953	0.914	0.021	0.689	0.992
MD7	1.457	0.934	0.316	0.969	1.056	0.108	0.652	0.990
MD8	1.084	0.632	0.212	0.964	0.95	0.000	0.728	0.981
MD9	1.223	0.751	0.279	0.959	1.027	0.090	0.754	0.983
MD10	1.140	0.357	0.260	0.959	1.173	0.024	0.778	0.986
MD11	1.538	0.384	0.299	0.943	1.046	0.020	0.654	0.986
MD12	1.061	0.407	0.407	0.966	0.943	0.074	0.595	0.992
MD13	1.255	0.730	0.340	0.977	0.995	0.052	0.617	0.991
MD14	1.473	0.659	0.388	0.976	1.139	0.176	0.617	0.991
MD15	1.265	0.581	0.372	0.977	0.852	0.090	0.639	0.976
MD16	1.752	0.912	0.412	0.985	0.988	0.000	0.478	1.000
MD17	1.571	1.101	0.314	0.977	1.092	0.127	0.756	0.991
MD18	1.597	0.772	0.308	0.965	1.264	0.110	0.793	0.984
MD19	1.306	0.743	0.273	0.97	1.010	0.109	0.592	0.990
MD20	1.044	0.549	0.302	0.941	0.836	0.085	0.667	0.977

employed as fixed benchmarks for the inefficient DMUs. For example, if an expert knows MD16 quite well, (s)he may use it to formulate guidelines for MD2 and MD12, which are worse than MD16 for all possible weights assigned to inputs and outputs.

A.6. Input and output values for the complaint groups G2 and G3

In Table A9, we present the input and output values for the complaint groups $G2$ (fever) and $G3$ (lower or upper extremity injury, head injury, and laceration/puncture). Together with group $G1$, they form the basis for conducting a multiscenario robustness analysis, whose results are discussed in Section 3.3 of the main paper and Section A7.

A.7. The analysis of pairwise preference relations for a multiscenario setting

This section presents the pairwise comparisons of physicians for three complaint groups. Table A10 reports the truth of the necessary-necessary $\succsim_{E,S}^{N,N}$ and necessary-possible $\succsim_{E,S}^{N,P}$ preference relations for all pairs of physicians. Since $\succsim_{E,S}^{N,N}$ is transitive, it can be presented graphically by its Hasse diagram (see Fig. A2). For 10 pairs of physicians, the necessary preference relation holds for all complaint groups. In particular, five physicians (MD5, MD6, MD8, MD19, and MD20) are always

Table A10
The truth of the necessary-necessary $\succsim_{E,S}^{N,N}$ (NN) and necessary-possible $\succsim_{E,S}^{N,P}$ (NP) efficiency preference relations for all pairs of physicians based on the analysis of three complaint groups

MD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	NN	NP	NP																	
2		NN																		
3			NN																	
4		NP	NP	NN						NP								NP		
5	NP	NP	NN	NP	NN		NP	NP	NP	NP		NP		NP	NP		NP	NP	NP	
6	NP	NP	NN	NP		NN	NP	NP	NN	NP		NP		NP	NP		NP	NP		
7	NP	NN	NP				NN			NP				NP				NP		
8		NP	NN					NN		NP				NP						
9		NP	NP						NN					NP				NP		
10		NP	NP			NP		NP		NN								NP		
11	NP	NP	NP	NP			NP	NP	NP	NP	NN	NP		NP				NP		
12	NP	NP	NP		NP		NP		NP	NP	NP	NN	NP	NP		NP	NP	NP	NP	
13	NP	NN	NP				NP	NP	NP	NP	NP		NN	NN		NP	NP	NP		
14	NP	NP	NP							NP				NN				NP		
15	NP	NP	NP	NP		NP	NP	NP		NP	NP			NP	NN			NP	NP	
16	NP	NP	NP	NP	NP		NP		NP	NP	NP	NP	NP	NP		NN	NP	NP	NP	
17		NP	NP											NP			NN	NP		
18		NN	NP	NP						NP								NN		
19	NP	NP	NN	NP				NP	NP	NP	NP			NP				NP	NN	
20	NP	NP	NN	NP			NP	NP	NP	NP		NP		NP	NP		NP	NP	NP	NN

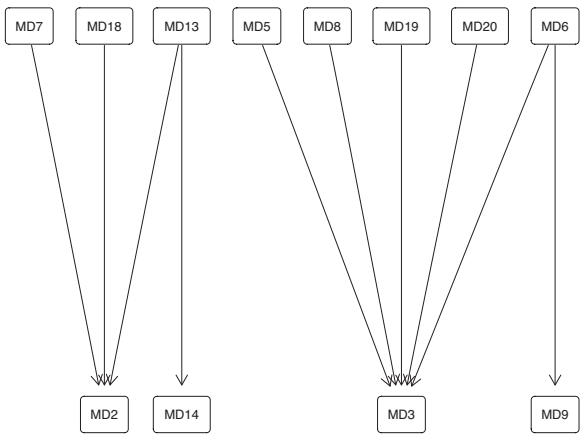

Fig. A2. The Hasse diagram of the necessary-necessary efficiency preference relation $\succsim_{E,S}^{N,N}$ based on the analysis of three complaint groups (for clarity of presentation, physicians not related by $\succsim_{E,S}^{N,N}$ with any other physician have been omitted).

Table A11
Ranks attained by physicians in the orders imposed by different measures derived from robustness analysis for complaint group $G1$ and differences between extreme distances, efficiencies, and ranks

	Ranks according to different measures											Widths of intervals		
MD	d_*	d^*	Ed	E^*	E_*	EE	R_*	R^*	ER	$\lvert \succsim^N \rvert$	$\lvert \precsim^N \rvert$	$d^* - d_*$	$E^* - E_*$	$R^* - R_*$
1	1	20	11	8	19	11	1	17	11	17	1	0.558	0.560	19
2	19	11	19	19	17	19	19	17	19	17	18	0.227	0.250	6
3	20	19	20	20	20	20	20	17	20	17	20	0.306	0.265	2
4	15	13	14	15	12	14	17	13	15	13	16	0.323	0.365	7
5	1	7	5	2	5	5	1	2	4	1	1	0.366	0.522	7
6	1	2	1	3	1	1	1	1	1	1	1	0.260	0.408	4
7	13	10	13	17	9	13	14	8	13	9	12	0.319	0.330	6
8	16	15	12	11	14	12	12	12	12	9	15	0.340	0.461	9
9	14	8	10	12	6	10	12	10	10	9	12	0.293	0.387	8
10	17	18	18	18	18	18	18	17	18	17	19	0.368	0.397	7
11	1	1	2	4	2	2	1	2	2	4	1	0.204	0.429	7
12	12	14	16	13	15	16	14	13	16	14	12	0.354	0.393	9
13	7	3	4	5	3	4	7	6	5	6	1	0.254	0.438	12
14	18	17	17	16	16	17	16	16	17	14	17	0.304	0.391	9
15	9	9	7	9	8	7	10	5	7	5	10	0.332	0.467	6
16	1	6	9	10	10	9	1	10	9	9	1	0.361	0.455	15
17	11	16	15	14	13	15	11	13	14	14	10	0.372	0.368	12
18	8	4	6	7	7	6	8	6	6	8	8	0.252	0.469	10
19	10	12	8	6	11	8	8	8	8	6	8	0.358	0.509	11
20	1	5	3	1	4	3	1	2	3	1	1	0.346	0.511	7

at least as efficient as MD3, and three physicians (MD7, MD18, and MD13) are more efficient than MD2. MD13 and MD6 can serve as the benchmark to follow for two other pairs (MD2 and MD14 or MD3 and MD9, respectively).

The necessary-possible preference relation $\succsim_{E,S}^{N,P}$ is more dense (see Table A10; note that the truth of $\succsim_{E,S}^{N,N}$ implies $\succsim_{E,S}^{N,P}$). There are 153 ordered pairs of physicians for whom the necessary relation holds for at least one complaint group. Interestingly, for some pairs (e.g., MD10, MD18), this relation is instantiated in both directions. Such observations, along with a high density of $\succsim_{E,S}^{N,P}$ and a scarcity of $\succsim_{E,S}^{N,N}$, suggest that the performance of physicians is strongly related to the complaint group and therefore some of them are better in treating specific groups of patients.

A.8. Summary of results derived from the robustness analysis

In this section, we summarize the results derived for each physician with the proposed robustness analysis framework for a single scenario referring to complaint group $G1$ and multiple scenarios concerning groups $G1$–$G3$.

In Table A11, we present the ranks of all physicians in the orders imposed by different measures following the application of robust efficiency analysis framework to group $G1$. These measures include extreme and expected distances, efficiency score, and ranks as well as the numbers of other

physicians which are less ($|\succsim^N|$) or more ($|\precsim^N|$) preferred than a given physician according to the necessary relation. The rankings are enriched with the differences between extreme distances, efficiencies, and ranks that indicate the stability of results for each physician.

These results confirm that the efficiency results are stable for some physicians irrespective of the accounted perspective and considered weight vectors. For example, MD6 is ranked at the top for 9 of 11 considered measures while attaining the second and third positions in the rankings determined by d^* and E^*, respectively. Such favorable results are justified by the relatively good performance of MD6 on all inputs and outputs. Furthermore, MD11 and MD20 also attain the ranks among the top five MDs according to all measures. On the other extreme, MD2, MD3, MD10, and MD14 are ranked relatively low. For example, MD3 is never ranked better than 17th. Its scores, efficiencies, and ranks are stable irrespective of the considered weights with the interval widths equal to 0.306, 0.264, and 2, respectively. This is understandable given its unfavorable performances on all accounted factors.

Even though the ranks attained by the vast majority of physicians are relatively stable irrespective of the accounted measure, one can indicate a few examples for which these indications are inconsistent. This is because of their unbalanced input/output profiles, making their performance strongly dependent on the considered weights and their ranks more prone to fluctuations with the change in the accounted measure. For example, the widest distance, efficiency, and rank intervals can be observed for MD1. Its ranks range from the most favorable (see, e.g., d_* and R_*) through medium (see, e.g., Ed, EE, and ER) to the least favorable (see, e.g., d^* and E_*). The great variability of results can also be noted for MD16. Its rank ranges from first (see, e.g., d_* and R_*) to tenth (see, e.g., E^* and E_*) depending on the selected measure, whereas a difference between extreme efficiency ranks ($R^* - R_*$) is 15.

Analogous results derived from the analysis of three complaint groups are presented in Table A12. The considered measures are extreme possible-possible distances to the efficient physician, efficiency scores, and ranks as well as the numbers of physicians which proved to be worse ($|\succsim^{N,P}|$) or better ($|\precsim^{N,P}|$) than a given physician according to the necessary-possible relation.

The ranks attained by different physicians according to the measures quantifying the results for multiple scenarios are, in general, less stable than for a single complaint group only. This confirms that the considered physicians attain more favorable results for complaint groups for which they have specialized skills while performing worse for other groups. Nevertheless, the conclusions on the best and worse performing physicians are similar. For example, MD15 attains ranks between first (see d_* and R_*) and eighth (see E_*) in the orders imposed by different measures. Furthermore, when considering the numbers of other physicians who proved to be necessarily-possibly worse or better than MD15, it is ranked sixth. Also, MD3 attains relatively stable ranks. It reaches the 14th position (i.e., the worst rank shared with six other physicians) in the order imposed by $R^{*,P}$, while being ranked in the bottom four according to all remaining measures. When compared to the results for group $G1$, significant changes in the outcomes attained for multiple scenarios considered jointly can be noted for MD12. For $G1$, MD12 was ranked outside the top 10 according to all measures. When considering all groups jointly, this happens for only two measures (see $d^{*,P}$ and E_*^P). Moreover, for some indicators, MD12 is ranked at the very top (see d_*^P and R_*^P). Such differences are implied by the relatively poor performance of MD12 for $G1$ and its favorable evaluation for other complaint groups.

Table A12
Ranks attained by physicians in the orders imposed by different measures derived from robustness analysis for complaint groups $G1$, $G2$, and $G3$

MD	d_*^P	$d^{*,P}$	$E^{*,P}$	E_*^P	R_*^P	$R^{*,P}$	$\lvert \succsim^{N,P} \rvert$	$\lvert \precsim^{N,P} \rvert$
1	1	20	16	19	1	14	18	15
2	20	6	20	17	20	14	19	19
3	19	19	17	20	19	14	19	19
4	10	18	2	12	9	14	13	11
5	1	13	13	5	1	4	2	2
6	1	9	6	1	1	1	5	2
7	15	5	14	9	18	6	10	11
8	9	7	7	14	9	4	13	11
9	16	10	11	6	17	6	13	11
10	12	17	18	18	12	14	11	17
11	1	12	12	2	1	14	8	7
12	1	14	8	15	1	6	2	7
13	11	1	9	3	9	1	6	2
14	17	11	16	16	15	12	11	16
15	1	2	4	8	1	6	6	6
16	1	16	1	10	1	14	1	2
17	18	8	15	13	12	6	13	10
18	14	4	19	7	15	12	13	18
19	13	3	10	11	12	3	9	7
20	1	15	3	4	1	6	2	1