Annals of Operations Research
https://doi.org/10.1007/s10479-024-05968-9

ORIGINAL RESEARCH

Classification and estimation of case-mix adjusted performance indices for binary outcomes

Marco Doretti[1] · Giorgio E. Montanari[2]

Received: 5 April 2022 / Accepted: 22 March 2024

Abstract
In this paper, we propose a general class of indices that can be used for comparing the performances of organizations providing a given public service to citizens, such as universities, hospitals, nursing homes, employment agencies or other institutions. In particular, we handle the case where evaluation is performed by assessing the probability that a given event has happened as a result of the service provided to users requiring it. Indices are designed for settings where users can be divided into groups with similar characteristics in order to account for case-mix, that is, for the different composition of users within each organization with respect to personal features influencing the probability of the event at hand. For the proposed class, we build a taxonomy leading to nine index types. These different types constitute a useful toolbox to satisfy specific needs and/or criteria set by the evaluator in applied contexts. A general inferential framework is also discussed to deal with settings where, whatever the index chosen, its value has to be estimated from sample data. A simulation study based on a real-world dataset is presented to assess the behavior of indices' estimators.

Keywords Service quality evaluation · Generalized performance index · Probability-based indicator · Index taxonomy · Case-mix

1 Introduction

As is well known, the usage of quantitative methods is nowadays ubiquitous in the context of performance evaluation. In this paper, we deal with the broad framework of service quality measurement, where typically interest lies in comparing the performance of organizations (such as schools, universities, hospitals, employment agencies or nursing homes) delivering a given public service to categories of citizens, be they students, patients, the unemployed,

✉ Marco Doretti
marco.doretti@unifi.it

Giorgio E. Montanari
giorgio.montanari@unipg.it

[1] Department of Statistics, Computer Science, and Applications, University of Florence, Viale Morgagni, 59, 50134 Florence, Tuscany, Italy

[2] Department of Political Science, University of Perugia, Via Pascoli, 20, 06123 Perugia, Umbria, Italy

Published online: 15 April 2024

the elderly and so on. Our work follows other recent contributions in *Annals of Operations Research*, characterized by comparative performance evaluation analyses in the fields of sports (Chessa et al., 2022; Carpita et al., 2023), business and economics (Boubaker et al., 2023) and social well-being (Fortuna et al., 2022).

In the field of service quality measurement, three key determinants are usually defined: efficacy, efficiency and equity/accessibility (Aday et al., 1999; Kruk & Freedman, 2008). In particular, when the quantitative assessment of efficacy is undertaken, metrics adopted to compare organizations (henceforth termed *agencies*) should be influenced only by how the service is provided, and not by the characteristics of people availing of services (henceforth, *users*). In this setting, the fundamental concept of *case-mix* comes into play. In detail, the case-mix of an agency is the composition of its users with respect to those personal features influencing the outcome variable comparisons are based upon (Berlowitz et al., 1996). In observational settings, such a composition is typically different for every agency. Thus, adjusting for case-mix is crucial whenever data collected on a set of users are employed for the comparative assessment of the agencies serving them. This strategy is also supposed to prevent service providers from relying on unfair practices like adverse selection (Romano, 2000). For example, when hospital wards performing life-saving surgery are evaluated by means of mortality rates, it is fundamental to somehow account for patients' health condition at admission. Otherwise, comparisons would be affected by the different difficulty level handled by each ward, and hospital managers would be tempted to avoid taking the most serious cases whenever possible.

In the above context, case-mix adjusted performance indices are usually obtained via standardization techniques. The usage of these techniques spans several decades within numerous fields, including demography, epidemiology, education and health insurance (Wolfenden, 1962; Kitagawa, 1964; Inskip et al., 1983; Draper & Gittoes, 2004; Schokkaert & Van de Voorde, 2009). In essence, standardization techniques consist in comparing factual (i.e., observed) outcomes with counterfactual (i.e., hypothetical) ones within groups of users that are homogeneous with respect to characteristics influencing the outcome variables (Chu, 1994): examples include the definition of the Ten Group Classification System (TGCS) of patients in obstetric care (Maso et al., 2013) or of Resources Utilization Groups (RUGs) of the elderly for the evaluation of Nursing Homes (NHs) involved in eldercare (Fries et al., 1994). In particular, RUGs are good proxies of the clinical complexity of NH residents, since they collect patients sharing the type of impairment and requiring the same overall care load. Thus, they represent a useful tool to address case-mix in various evaluation exercises, including those targeting the overall ability of NHs to assure survival of their residents in a certain time span after admission. For such an exercise, a standardized performance index would be built by aggregating RUG-specific *contrasts* between the survival probability of residents of a given NH and the marginal survival probability, that is, the one obtained by averaging across residents of all NHs in the same RUG group. With this regard, a crucial matter concerns *how such contrasts should be built*, that is, which contrasting function should be adopted. Indeed, different choices emphasize different aspects of the problem. Incidentally, we remark that standardization techniques can be *de facto* thought of as causal inference methods, as recently noticed by Longford (2020). The parallel with the causal framework was highlighted also by Wray et al. (1997), who framed case-mix as a source of confounding for the treatment-outcome relationship, with the treatment being the supplying agency's conduct in responding to user needs.

In this paper, we focus on probability-based indices, that is, on a setting where the evaluation of agencies is based on the probability that a given event has happened to their users as a result of the provided service. In addition to the above-mentioned exercise concerning survival of patients in NHs, other examples may include university students obtaining a degree, or patients successfully recovering after a given hospital treatment. In such a framework, our ultimate goal is to provide guidance for a conscious choice of the contrasting function. To this end, we propose a general class of case-mix adjusted performance indices and develop a taxonomy for the elements in the class which is based on two key features of the contrast functions. Such a taxonomy leads to a nine-fold classification, that can be referred to for better characterizing the evaluative properties of any given index in the class. In this way, analysts can properly choose an index according to the specific evaluative needs and criteria they want to consider for the real-world application at hand. Besides, a general framework for the estimation of index values (as well as of their uncertainty measures) is illustrated for applications using sample data from observational studies.

The remainder of the paper is organized as follows. In Sect. 2, we briefly describe a motivating example in the context of NH performance evaluation. In Sect. 3.1, we lay out the general class of indices, while in Sect. 3.2 we introduce the proposed taxonomy. In Sect. 3.3, we show how our classification scheme depends on what we call the *departure paradigm*, that is, on the metric invoked (often implicitly) by analysts when evaluating the extent to which a probability varies with respect to another one. In Sect. 4, we illustrate some numerical examples with particular reference to the one discussed by Castro et al. (2015), which is re-analyzed in light of the introduced concepts (Sect. 4.1). In Sect. 5, we turn our attention to estimation and provide an approximate inferential framework for any element in the class, based on linearization techniques. A number of challenges related to the estimation of second-order moments are analyzed and addressed via further approximation (Sect. 5.1). The reliability of the whole estimation approach is assessed in Sect. 6 through a simulation study based on the real-world dataset introduced in Sect. 2. In Sect. 7, some concluding remarks are offered.

2 Motivating example

In Umbria, a region of Central Italy, the evaluation of the performance of NHs hosting residents in need of care is routinely conducted by the regional government, following international standards (Montanari & Doretti, 2019). Data on NH residents are gathered using an international health protocol named *Suite interRAI* (Carpenter & Hirdes, 2013); see e.g. Montanari et al. (2022) for details. In this context, patient survival is one of the major monitoring targets. With this respect, in order to deal with case-mix the Umbrian government has adopted the RUG classification introduced in Sect. 1, in line with the prevailing approaches within NH care (Punelli & Williams, 2013; Broussard & Reiter, 2020). As already mentioned, RUGs constitute the most common elderly classification scheme for case-mix adjustment, since patients in the same RUG have similar clinical complexity levels.

The dataset considered here refers to the years 2018–2019 and contains indicators of survival (after 1 year) and RUG membership for 1748 residents hosted in 47 Umbrian NHs. The number of residents in each NH ranges from 15 to 84 (with an average of 37.2). There are 30 RUGs in total, and the number of residents assigned to the same RUG varies between 7 and 193 (average 58.3). Survival is quite variable across RUGs, denoting a rather high relevance of case-mix with respect to this variable. Specifically, marginal (across NH) RUG-specific survival rates vary from 0.533 to 0.960, with a standard deviation of 0.094. Also, of the 1410 strata formed by NH-RUG crossing, only 653 (46%) contain at least one resident, which denotes a quite high degree of sparsity in the data. In this setting, case-mix adjusted NH performance indices could be built as described in Sect. 1, that is, via the aggregation of RUG-specific contrasts between agency-specific and across-agency-averaged survival probabilities.

3 A class of probability-based performance indices

3.1 Definition of the class

In order to introduce the general class of performance indices, some preliminaries are required. First, we assume that the evaluation process is carried out via assessing, for each agency, the probability that a given event has happened as a result of the service provided to users (for example, survival of an NH resident 1 year after baseline). As a consequence, the outcome variable into play is binary: either the event happened (success) or not (failure). Secondly, we assume that the homogeneous groups of users for identifying the case-mix of each agency are given (for example, the RUG classification for the elderly hosted in NHs). Thirdly, for an effective case-mix adjustment, it is also assumed that the probability of success depends on the variable used to describe case-mix, as in our motivating example where patients survival depends on the RUG group.

Starting from the above setting, some pieces of notation are needed. Specifically, we let $h \in \mathscr{H} = \{1, \dots, H\}$ and $j \in \mathscr{J} = \{1, \dots, J\}$ be two subscripts indexing supplying agencies and case-mix adjusting groups, respectively. Also, we let p_{hj} denote the success probability for one of the N_{hj} users of the jth group handled by the hth agency, whereas $p_{.j}$ indicates the group-specific success probability after marginalization across all agencies. The vector $\boldsymbol{p} = (p_{.1}, \dots, p_{.J})$ collects these marginal probabilities for all groups. The two marginal counts $N_{h.} = \sum_{j=1}^{J} N_{hj}$ and $N_{.j} = \sum_{h=1}^{H} N_{hj}$ and the corresponding proportions $W_{hj} = N_{hj}/N_{h.}$ and $Q_{hj} = N_{hj}/N_{.j}$ are also defined. In particular, W_{hj} denotes the distribution of users of the hth agency across the J groups, that is, the agency-specific case-mix distribution. Conversely, Q_{hj} denotes the distribution of users in the j-th group across the H agencies.

Armed with this notation, we introduce the generalized Adjusted Performance Index (API). For every agency $h \in \mathscr{H}$, the index value is given by

$$\mathrm{API}_h^{(G)} = 1 + \sum_{j \in \mathscr{J}_h} W_{hj} G(p_{hj}, \boldsymbol{p}), \tag{1}$$

where $\mathscr{J}_h = \{j \in \mathscr{J} : N_{hj} > 0\}$ is the subset of groups for which the hth agency handles at least one user, and $G(p_{hj}, \boldsymbol{p})$ is a generic contrast function such that:

(i) $G(p_{hj}, \boldsymbol{p}) = 0$ whenever $p_{hj} = p_{.j}$;
(ii) $G(p_{hj}, \boldsymbol{p})$ is an increasing function in its first argument p_{hj}, given its second argument $\boldsymbol{p}$.

Indeed, these two properties of the G function ensure that higher values of the generalized index in (1) are associated to better performances (ability to achieve success). Adding 1 to the summation assures that $\mathrm{API}_h^{(G)}$ takes values around 1. Such a level corresponds to an agency whose probability of success is equal to the marginal one for all its groups. Roughly speaking, index values lower than 1 denote a performance which is worse than the marginal one, whereas values greater than 1 denote a better performance with respect to the marginal one.

It is worth to underline that the use of the agency-specific distribution W_{hj} ascribes $\mathrm{API}_h^{(G)}$ in (1) to the class of *indirect* standardization methods. These are opposed to *direct* standardization, for which the marginal distribution $W_{.j} = N_{.j}/N$ (where $N = \sum_j N_{.j}$) would be used instead. While strengths and limitations of these two standardization methods have been for long analyzed in the literature (Inskip et al., 1983; Curtin, 1995; Julious et al., 2001; Pouw et al., 2013) and remain beyond the scope of this paper, we just remark that only indirect standardization is applicable when some agencies lack data for some case-mix groups. This is often the case when the sample size is small and/or the number of groups is large like in the motivating dataset of Sect. 2. In practice, lack of data is neutralized by the fact that a null weight $W_{hj} = 0$ is assigned to these groups. Consequently, the distinction between $\mathscr{J}_h$ and $\mathscr{J}$ (that would necessarily appear in (1) if the $W_{.j}$ distribution would be used) is formal but not substantial. This remark is rather important, since it highlights that our approach is independent of the kind of case-mix distribution in use. Thus, the key arguments illustrated in what follows would in principle hold in a direct standardization setting, too.

For every group $j \in \mathscr{J}_h$, the $G(p_{hj}, \boldsymbol{p})$ function within the summation in (1) characterizes the contrast between the agency-specific (p_{hj}) and marginal ($p_{.j}$) success probability. For example, two possible functions operating with differences are $G_d(p_{hj}, \boldsymbol{p}) = p_{hj} - p_{.j}$ and $G_{d_n}(p_{hj}, \boldsymbol{p}) = (p_{hj} - p_{.j})/\sum_{j \in \mathscr{J}_h} W_{hj} p_{.j}$, leading to

$$\mathrm{API}_h^{(G_d)} = 1 + \sum_{j \in \mathscr{J}_h} W_{hj}(p_{hj} - p_{.j}) \tag{2}$$

and

$$\mathrm{API}_h^{(G_{d_n})} = \frac{\sum_{j \in \mathscr{J}_h} W_{hj} p_{hj}}{\sum_{j \in \mathscr{J}_h} W_{hj} p_{.j}}. \tag{3}$$

Another function, operating with relative deviations, is $G_r(p_{hj}, \boldsymbol{p}) = (p_{hj} - p_{.j})/p_{.j} = p_{hj}/p_{.j} - 1$, resulting in

$$\mathrm{API}_h^{(G_r)} = \sum_{j \in \mathscr{J}_h} W_{hj} \frac{p_{hj}}{p_{.j}}. \tag{4}$$

In practice, in many cases contrasts are not a function of other arguments than p_{hj} and $p_{.j}$, that is, $G(p_{hj}, \boldsymbol{p}) \equiv G(p_{hj}, p_{.j})$.

3.2 Index taxonomy

We now introduce a taxonomy for the elements belonging to the API class, which generates a nine-fold classification system. To this purpose, we introduce the concepts of *balancing* and *level-invariance*. To better motivate the whole approach, we first offer a non-technical introduction to these concepts, framed in the context of the motivating example of Sect. 2. Then, we show how they can be expressed in terms of the contrast function G, the key object characterizing $\mathrm{API}_h^{(G)}$ in Eq. (1).

To capture the idea behind level-invariance, we recall that NH performance indices discussed in Sect. 2 were defined by aggregation of RUG-specific contrasts between NH-specific and marginal one-year-ahead survival probabilities. In many settings, these probabilities are likely to be relatively high, more than 0.5, say. Therefore, in order to compare and/or rank NHs, an analyst might want to choose one of the following alternatives:

(i) to raise the impact on $\mathrm{API}_h^{(G)}$ of RUGs with lower marginal probabilities, that could be deemed more informative than others (that is, the real "battlefield" where competition among NHs takes place);

(ii) to lower the impact on $\text{API}_h^{(G)}$ of RUGs with lower marginal probabilities, so that the evaluation exercise is built around the bulk of remaining RUGs (possibly deemed more representative);
(iii) not to distinguish among RUGs with respect to their marginal survival probability.

We term the indices satisfying the above requirements, respectively:

(i) negatively level-dependent: contrasts around lower marginal survival probabilities are emphasized;
(ii) positively level-dependent: contrasts around higher marginal survival probabilities are emphasized;
(iii) level-invariant: contrasts around lower and higher marginal survival probabilities are equally emphasized.

With respect to balancing, we can define negative (positive) imbalance as the overall tendency of an index to emphasize NH performance levels that are, *ceteris paribus*, worse (better) than the marginal one. Thus, an analyst may want to choose one of the following alternatives:

(i) to raise the impact on $\text{API}_h^{(G)}$ of contrasts where NH survival probabilities are lower than the marginal ones, so that a worse-than-average performance in a RUG is not counterbalanced by a better-than-average performance of the same intensity in another RUG;
(ii) to raise the impact on $\text{API}_h^{(G)}$ of contrasts where NH survival probabilities are higher than the marginal ones, so that a better-than-average performance in a RUG is not counterbalanced by a worse-than-average performance of the same intensity in another RUG;
(iii) not to distinguish among contrasts with worse-than-average and better than-average NH survival probabilities, so that compensation might in principle be possible when the intensity of the contrasts is the same.

The suited indices are termed:

(i) negatively unbalanced: contrasts with lower-than-average NH survival probabilities are emphasized;
(ii) positively unbalanced: contrasts with higher-than-average NH survival probabilities are emphasized;
(iii) balanced: contrasts with higher- and lower-than-average NH survival probabilities are equally emphasized.

In terms of the contrast function G, level-invariant indices are those where G assigns the same relevance to a given probability departure, whatever the level of the corresponding marginal probability. In other words, for level-invariant indices departures having the same intensity and direction lead to the same value of the group-specific contrast $G(p_{hj}, \boldsymbol{p})$, for any given value of the marginal probability $(p_{.j})$ in $\boldsymbol{p}$. In contrast, positively (negatively) level-dependent indices have a G function emphasizing the magnitude of departures from higher (lower) marginal probability levels.

With regard to balancing, balanced indices are those with a G function assigning the same relevance to positive and negative departures of p_{hj} from $p_{.j}$. The balance property assures that two departures from any given value of the marginal probability produce opposite values of G, whenever these have the same intensity but opposite directions. Conversely, the G function of positively (negatively) unbalanced indices increases the magnitude of departures towards higher (lower) levels from the marginal probabilities.

Table 1 The general taxonomy for case-mix adjusted performance indices

Type	Description
A	Negatively unbalanced and negatively level-dependent
B	Negatively unbalanced and level-invariant
C	Negatively unbalanced and positively level-dependent
D	Balanced and negatively level-dependent
E	Balanced and level-invariant (neutral)
F	Balanced and positively level-dependent
G	Positively unbalanced and negatively level-dependent
H	Positively unbalanced and level-invariant
I	Positively unbalanced and positively level-dependent

According to this twofold classification system, nine types of indices can be defined which we denote with letters from A to I. Their description is summarized in Table 1. An index which is both balanced and level-invariant is termed *neutral*. The main feature of neutral indices is that they are capable of *fully* capturing performance offsets across case-mix adjusting groups. Indeed—net of differences in case-mix distributions—for neutral indices the contribution of positive and negative probability departures of the same intensity *always* cancels out, even when these involve two groups with different marginal probability levels. In contrast, when an index is balanced but level-dependent, such a compensation occurs only when the marginal probabilities are equal. For this reason, we deem neutral indices a "natural" choice within the whole API class; see the numerical example in Sect. 4 for further details.

3.3 Dependence on the departure paradigm

A generic departure of a probability p_{hj} from its corresponding marginal $p_{.j}$ can be read according to various metrical paradigms. The three most common ones are: *i*) absolute deviations $p_{hj} - p_{.j}$, *ii*) variations $p_{hj}/p_{.j}$, and *iii*) relative deviations $(p_{hj} - p_{.j})/p_{.j}$. As a consequence, the meaning of the word *intensity* in Sect. 3.2 is not univocal: two departures towards opposite directions or from different starting levels might have the same intensity under a certain paradigm but not under another one. For instance, under the absolute deviation paradigm the departure of $p_{hj} = 0.4$ from $p_{.j} = 0.2$ has the same intensity of the departure of $p_{hj} = 0.6$ from $p_{.j} = 0.4$. However, this is not the case under the other two paradigms. Clearly, the concept of *performance offset* invoked in Sect. 3.2 to characterize neutrality is also influenced in turn.

In light of the considerations in the above paragraph, the mathematical formalization of the concepts of balancing and level-invariance (as well as their negations) must be paradigm-specific. In what follows, we report suitable definitions for the three major paradigms introduced, limiting our attention to *admissible* departures from $p_{.j}$, that is, departures resulting in a p_{hj} probability in the 0–1 range. These definitions are based on the coherence conditions in Sect. 3.1, i.e., that (i) $G(p_{hj}, \boldsymbol{p}) = 0$ if $p_{hj} = p_{.j}$ and (ii) $G(p_{hj}, \boldsymbol{p})$ is increasing in p_{hj} given $\boldsymbol{p}$.

For the absolute deviation paradigm, we define the balancing condition as

$$G(p_{.j}+\varepsilon,\boldsymbol{p})+G(p_{.j}-\varepsilon,\boldsymbol{p})\begin{cases}<0: & \text{Negative imbalance}\\ =0: & \text{Balance}\\ >0: & \text{Positive imbalance}\end{cases} \tag{5}$$

$\forall\ p_{.j}$ and $\forall\ \varepsilon > 0$, where, as mentioned before, $\boldsymbol{p}$ is the vector collecting group-specific marginal probabilities, its jth component being $p_{.j}$.

Similarly, the level-invariance condition is defined by

$$G(p_{.j}\pm\varepsilon,\boldsymbol{p})-G(p'_{.j}\pm\varepsilon,\boldsymbol{p}')\begin{cases}\lesseqgtr 0: & \text{Negative level-dependence}\\ =0: & \text{Level-invariance}\\ \gtreqless 0: & \text{Positive level-dependence}\end{cases} \tag{6}$$

$\forall\ (p_{.j}, p'_{.j}):\ p_{.j} > p'_{.j}$ and $\forall\ \varepsilon > 0$, where $\boldsymbol{p}'$ is the same as $\boldsymbol{p}$ except for $p'_{.j}$ replacing $p_{.j}$ in the jth element.

For the variation paradigm, the analogous definitions are

$$G(p_{.j}\varepsilon,\boldsymbol{p})+G(p_{.j}/\varepsilon,\boldsymbol{p})\begin{cases}<0: & \text{Negative imbalance}\\ =0: & \text{Balance}\\ >0: & \text{Positive imbalance}\end{cases} \tag{7}$$

$\forall\ p_{.j}$ and $\forall\ \varepsilon > 1$, and

$$G(p_{.j}\cdot\varepsilon^{\pm 1},\boldsymbol{p})-G(p'_{.j}\cdot\varepsilon^{\pm 1},\boldsymbol{p}')\begin{cases}\lesseqgtr 0: & \text{Negative level-dependence}\\ =0: & \text{Level-invariance}\\ \gtreqless 0: & \text{Positive level-dependence}\end{cases} \tag{8}$$

$\forall\ (p_{.j}, p'_{.j}):\ p_{.j} > p'_{.j}$ and $\forall\ \varepsilon > 1$.

Finally, for the relative deviation paradigm we set the balancing condition to

$$G(p_{.j}(1+\varepsilon),\boldsymbol{p})+G(p_{.j}(1-\varepsilon),\boldsymbol{p})\begin{cases}<0: & \text{Negative imbalance}\\ =0: & \text{Balance}\\ >0: & \text{Positive imbalance}\end{cases} \tag{9}$$

$\forall\ p_{.j}$ and $\forall\ \varepsilon > 0$, whereas the level-invariance condition is

$$G(p_{.j}(1\pm\varepsilon),\boldsymbol{p})-G(p'_{.j}(1\pm\varepsilon),\boldsymbol{p}')\begin{cases}\lesseqgtr 0: & \text{Negative level-dependence}\\ =0: & \text{Level-invariance}\\ \gtreqless 0: & \text{Positive level-dependence}\end{cases} \tag{10}$$

$\forall\ (p_{.j}, p'_{.j}):\ p_{.j} > p'_{.j}$ and $\forall\ \varepsilon > 0$.

4 Numerical examples

In Table 2, we report examples of G functions that define indices belonging to all the nine types quoted in Table 1 within each of the three major departure paradigms. Obviously, alternative sets could be considered in principle. All the reported functions are such that dependence on the second argument $\boldsymbol{p}$ operates through $p_{.j}$ only, that is, $G(p_{hj}, \boldsymbol{p}) \equiv G(p_{hj}, p_{.j})$. Their

Table 2 Some numerical examples of G functions generating all types of indices for the three major paradigms

		$j = 1$		$j = 2$		$\text{API}_h^{(G)} - 1$	
Type	$G(p_{hj}, \boldsymbol{p}) \equiv G(p_{hj}, p_{.j})$	$h = 1$	$h = 2$	$h = 1$	$h = 2$	$h = 1$	$h = 2$
Absolute deviations: $\varepsilon = 0.1$ ($\pm$ 10 percentage points)							
$p_{11} = 0.1$, $p_{.1} = 0.2$, $p_{21} = 0.3$; $p_{12} = 0.6$, $p_{.2} = 0.5$, $p_{22} = 0.4$							
A	$\log\{(1 + p_{hj})/(1 + p_{.j})\}$	−0.087	0.080	0.065	−0.069	−0.011	0.006
B	$\log\{p_{hj} - p_{.j} + 1\}$	−0.105	0.095	0.095	−0.105	−0.005	−0.005
C	$1 - \{(1 - p_{hj})/(1 - p_{.j})\}^2$	−0.266	0.234	0.360	−0.440	0.047	−0.103
D	$p_{hj}/p_{.j} - 1$	−0.500	0.500	0.200	−0.200	−0.150	0.150
E	$p_{hj} - p_{.j}$	−0.100	0.100	0.100	−0.100	0.000	0.000
F	$-(1 - p_{hj})/(1 - p_{.j}) + 1$	−0.125	0.125	0.200	−0.200	0.038	−0.038
G	$(p_{hj}/p_{.j})^2 - 1$	−0.750	1.250	0.440	−0.360	−0.155	0.445
H	$(p_{hj} - p_{.j} + 1)^2 - 1$	−0.190	0.210	0.210	−0.190	0.010	0.010
I	$p_{hj}^2 - p_{.j}^2$	−0.030	0.050	0.110	−0.090	0.040	−0.020
Variations: $\varepsilon = 2$ (doubling/halving)							
$p_{11} = 0.1$, $p_{.1} = 0.2$, $p_{21} = 0.4$; $p_{12} = 1$, $p_{.2} = 0.5$, $p_{22} = 0.25$							
A	$1/\sqrt{p_{.j}} - 1/\sqrt{p_{hj}}$	−0.926	0.655	0.414	−0.586	−0.256	0.035
B	$-\sqrt{p_{.j}/p_{hj}} + 1$	−0.414	0.293	0.293	−0.414	−0.061	−0.061
C	$\log\{2p_{hj}/(p_{hj} + p_{.j})\}p_{.j}$	−0.081	0.058	0.144	−0.203	0.031	−0.073
D	$\log\{p_{hj}/p_{.j}\}/p_{.j}$	−3.466	3.466	1.386	−1.386	−1.040	1.040
E	$\log\{p_{hj}/p_{.j}\}$	−0.693	0.693	0.693	−0.693	0.000	0.000
F	$\log\{p_{hj}/p_{.j}\}p_{.j}$	−0.139	0.139	0.347	−0.347	0.104	−0.104
G	$\log\{(p_{.j} + p_{hj})/2p_{.j}\}(1 - p_{.j})$	−0.230	0.324	0.203	−0.144	−0.014	0.090
H	$p_{hj}/p_{.j} - 1$	−0.500	1.000	1.000	−0.500	0.250	0.250
I	$p_{hj} - p_{.j}$	−0.100	0.200	0.500	−0.250	0.200	−0.025
Relative deviations: $\varepsilon = 0.2$ (20% increase/decrease)							
$p_{11} = 0.16$, $p_{.1} = 0.2$, $p_{21} = 0.24$; $p_{12} = 0.6$, $p_{.2} = 0.5$, $p_{22} = 0.4$							
A	$\log\{p_{hj}/p_{.j}\}/p_{.j}$	−1.116	0.912	0.365	−0.446	−0.376	0.233
B	$\log\{p_{hj}/p_{.j}\}$	−0.223	0.182	0.182	−0.223	−0.020	−0.020
C	$\log\{(1 + p_{hj})/(1 + p_{.j})\}$	−0.034	0.033	0.065	−0.069	0.015	−0.018
D	$(p_{hj}/p_{.j} - 1)/p_{.j}$	−1.000	1.000	0.400	−0.400	−0.300	0.300
E	$p_{hj}/p_{.j} - 1$	−0.200	0.200	0.200	−0.200	0.000	0.000
F	$p_{hj} - p_{.j}$	−0.040	0.040	0.100	−0.100	0.030	−0.030
G	$\{(p_{hj}/p_{.j})^2 - 1\}(1 - p_{.j})$	−0.288	0.352	0.220	−0.180	−0.034	0.086
H	$(p_{hj}/p_{.j})^2 - 1$	−0.360	0.440	0.440	−0.360	0.040	0.040
I	$p_{hj}^2 - p_{.j}^2$	−0.014	0.018	0.110	−0.090	0.048	−0.036

coherence with index types can be checked by verifying (via first-principle math) the set of equalities/inequalities in (5)–(10). For instance, it is easy to prove that the first index in the relative deviation group is an A-type index for this paradigm. Indeed, it emphasizes relative deviations *towards* lower-than-average values, as well as *from* lower average values. The respective inequalities are

$$G(p_{.j}(1+\varepsilon), \boldsymbol{p}) < -G(p_{.j}(1-\varepsilon), \boldsymbol{p})$$

for every feasible $(p_{.j}, \varepsilon)$ pair ($\varepsilon > 0$), and

$$G(p_{.j}(1+\varepsilon), \boldsymbol{p}) < G(p'_{.j}(1+\varepsilon), \boldsymbol{p}') \ \text{ and } \ G(p_{.j}(1-\varepsilon), \boldsymbol{p}) > G(p'_{.j}(1-\varepsilon), \boldsymbol{p}')$$

for every feasible $(p_{.j}, p'_{.j}, \varepsilon)$ triplet where $\varepsilon > 0$ and $p_{.j} > p'_{.j}$.

As mentioned before, indices can change their nature when the paradigm is modified. For example, the index based on $G_d(p_{hj}, \boldsymbol{p}) = p_{hj} - p_{.j}$ (i.e., $\text{API}_h^{(G_d)}$ in (2)) is an E-type index for absolute deviations, an I-type index for variations and an F-type index for relative deviations. Analogously, the index defined by $G_r(p_{hj}, \boldsymbol{p}) = p_{hj}/p_{.j} - 1$ ($\text{API}_h^{(Gr)}$ in (4)) is a D-type index for absolute deviations, an H-type index for variations and an E-type index for relative deviations.

In columns 3–8, Table 2 also offers the values of the $G(p_{hj}, p_{.j})$ functions in a setting where there are two agencies, each having users divided into two groups of the same size. This setting rules out the variability due to different case-mix compositions, thereby allowing to appreciate the most relevant features of indices of each type. For these two groups, we have $p_{.1} = 0.2$ and $p_{.2} = 0.5$, meaning that the first group includes "more difficult" users (that is, users associated to lower success probabilities on average), while the second one includes "easier" users. Moreover, the two agencies are such that the first one has a worse-than-average performance in the difficult group and a better-than-average performance in the easy group, with the opposite being true for the second agency. The intensities of these probability departures (anyhow measured) are equivalent. Specifically, for absolute deviations we set $\varepsilon = 0.1$, corresponding to an increase/decrease of ten percentage points in each success probability $p_{.j}$; for variations and relative deviations, we set $\varepsilon = 2$ (doubling/halving $p_{.j}$) and $\varepsilon = 0.2$ (20% increase/decrease with respect to $p_{.j}$), respectively. In this way, performance offsets for both agencies are created within each paradigm.

The last two columns in Table 2 contain the values of $\text{API}_h^{(G)}-1$ for the two agencies, with the subtractions of 1 made in order to make symmetries more evident. The analysis of these values provides many useful insights for every paradigm. The two key facts to notice are that: (i) level-invariant indices B, E and H take the same value for the two agencies, and (ii) balanced indices D, E and F sum to 0. As a consequence, only neutral indices E assign an equal-to-average performance level to the two agencies, thereby acknowledging—as expected—the induced performance offsets. In this sense, we argue that neutral indices can be thought of as a "natural" choice within each paradigm.

While E indices return null values in Table 2, B (H) indices result in negative (positive) values. They can be thought of as having an overall tendency to emphasizing departures towards low (high) values, in line with their nature. This tendency extends also to other Indices: it is reflected by the fact that negatively unbalanced indices A, B and C sum to negative values across the two agencies, while positively unbalanced indices G, H and I sum to positive values. Finally, a comment on D and F indices is in order. D indices favor the second agency and disadvantage the first one by putting more emphasis on departures in the first group, where the first agency performs worse than the average and the second performs better. For F indices, the reverse holds. In any case, the balance property ensures that these discrepancies from the null value have the same intensity.

Table 3 The example discussed in Castro et al. (2015)

	Surgeon A ($h = 1$)				Surgeon B ($h = 2$)			
	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 1$	$j = 2$	$j = 3$	$j = 4$
W_{hj}	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
$p_{.j}$	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
p_{hj}	0.2	0	0	0	0	0	0.2	0
$p_{hj} - p_{.j}$	0.1	−0.2	−0.3	−0.4	−0.1	−0.2	−0.1	−0.4
$p_{hj}/p_{.j}$	2	0	0	0	0	0	0.67	0

4.1 An example from the literature

The adoption of a specific departure paradigm is a subjective and often implicit choice. To better illustrate the relevance of this point—and how it might lead to controversial interpretations in some circumstances—we reanalyze the example discussed by Castro et al. (2015), where the usage of traditional risk-adjusted mortality rates is criticized in the spirit of other earlier contributions (Moreno & Apolone, 1997; Moreno et al., 1998; Metnitz et al., 2000). As is well-known, these rates are based on contrasting actual and expected mortality rates. The second ones are obtained by averaging marginal mortality rates with agency-specific case-mix distributions, and they represent the mortality levels that would occur if agencies had an "average behavior". Notice that in this example the outcome "mortality" instead of "survival" is considered, so better performances correspond to lower values of the indices.

The data of this example are summarized in Table 3. Specifically, Castro et al. (2015) consider a setting where there are $H = 2$ surgeons (corresponding to agencies) operating on two sets of patients (corresponding to users). These two sets share the same case-mix distribution W_{hj}, which is a uniform distribution across $J = 4$ risk groups (first row of the table). The vector of (group-specific) marginal mortality rates is $\boldsymbol{p} = (0.1, 0.2, 0.3, 0.4)$ (second row). For surgeon A, all mortality rates but the first ($j = 1$) are null, whereas for surgeon B the only non-null rate is the third ($j = 3$). In both cases, the non-null rates are equal to 0.2 (see third row).

According to the above data setting, the two surgeons show the same expected mortality rate ($\sum_{j=1}^{4} W_{hj} p_{.j} = 0.25$), as well as the same actual rate ($\sum_{j=1}^{4} W_{hj} p_{hj} = 0.05$). As a consequence, traditional risk-adjusted measures based on contrasting these two numbers mark the two surgeons with the same performance level, regardless of the chosen contrasting function. The critique moved by Castro et al. (2015) lies in the fact that "surgeon B's deaths are more likely to be justified than the ones that occurred to surgeon A" (since they were realized in a riskier group) and "a suitable performance metric should account for this and, clearly, traditional risk adjusted mortality rates are not such a metric".

We argue that the above point by Castro et al. (2015) implicitly invokes a relative deviation paradigm. For example, a neutral index for such a paradigm like the one in (4) (see Table 2) favors surgeon B at the expense of surgeon A. Indeed, the two averages of the $p_{hj}/p_{.j}$ contrasts (fifth row of Table 3) are 0.5 for surgeon A and 0.167 for surgeon B (we recall that, since we are dealing with mortality, higher index values correspond to worse performances).

Nevertheless, one might also argue that the null rate of surgeon A in group $j = 3$ should contribute to lower the index (and thus to increase the performance level) more than the null rate of surgeon B in group $j = 1$ since, again, it occurred in a riskier group. Such a dynamic, however, is not captured by the $p_{hj}/p_{.j}$ ratios, where we have

$$p_{13}/p_{.3} = 0/0.3 = 0 = 0/0.1 = p_{21}/p_{.1}.$$

It is captured, instead, by the $p_{hj} - p_{.j}$ differences, for which $p_{13} - p_{.3} = -0.3$ and $p_{21} - p_{.1} = -0.1$; see the fourth row in Table 3. It is easy to verify that the difference-based indices (2) and (3) - neutral for the absolute deviation paradigm—take the same value (0.8 and 0.2, respectively) for both surgeons, denoting the same performance level. As a matter of fact, these indices are indeed two traditional risk-adjusted mortality rates comparing, on different scales, actual ($\sum_{j\in\mathscr{J}_h} W_{hj}p_{hj}$) and expected ($\sum_{j\in\mathscr{J}_h} W_{hj}p_{.j}$) rates.

In our view, this example shows that there are no "right" or "wrong" indices in a general sense. Instead, we argue that the numerical specifics of a given problem might trigger, often unconsciously, the adoption of a certain metrical paradigm. Since every paradigm has some "naturally associated" indices (typically, the neutral ones), results from other indices might seem counter-intuitive in the first place. In summary, we believe that the analyst should explicit:

(i) the assumed paradigm, i.e, the way compensations (offsets) between positive and negative departures from marginal success probabilities are thought of;
(ii) whether negative departures from marginal probabilities should outweigh positive departures of the same magnitude (or viceversa);
(iii) whether departures from lower levels of the marginal probabilities should be more relevant than those from higher levels (or viceversa).

If the answer to (ii) and (iii) is "no", then a neutral index for the adopted paradigm is usually appropriate; otherwise, unbalanced and/or level dependent indices should be considered. Ultimately, every choice can be "right", as long as it is consciously motivated.

5 Estimation

As mentioned in the previous section, subject-matter considerations might lead to choose a paradigm, an index of a given type for that paradigm and, in turn, a coherent G function. Beyond these decisions, the problem arises of estimating the resulting $\text{API}_h^{(G)}$ index from sample data. We here take a probabilistic approach and introduce a random variable Y_{hji} representing the binary outcome of a user i among those served by the hth agency and belonging to the jth group ($i \in \{1, \ldots, N_{hj}\}$). We model the Y_{hji} variables as uncorrelated Bernoulli variables with success probability equal to p_{hj}, treating the observed outcomes as realizations from them. This framework allows to interpret p_{hj} as a group-specific quality measure, while also accounting for the uncertainty induced by accidental (i.e., external to case-mix) factors that possibly affect the outcome process. Notice, however, that we do not demand to observe data at the individual level: for each agency, group-specific observed success rates (for groups with at least one user) suffice. These observed rates can be considered as estimates from unbiased estimators $\hat{p}_{hj}$ of the associate p_{hj} probabilities. Clearly, such a rationale extends to estimators of marginal probabilities, that can be collected in the across-group vector $\hat{\boldsymbol{p}} = (\hat{p}_{.1}, \ldots, \hat{p}_{.J})$. In this setting, the natural estimator of the $\text{API}_h^{(G)}$ index (1) is

$$\widehat{\text{API}}_h^{(G)} = 1 + \sum_{j\in\mathscr{J}_h} W_{hj} G(\hat{p}_{hj}, \hat{\boldsymbol{p}}). \tag{11}$$

It is worth to remark that index estimates might not be mathematically defined in finite samples depending on the functional form of the G function, in particular when some probability estimates take one of the extreme values (0 or 1). For example, the estimate of (4) is not defined if $\hat{p}_{.j}$ is null for some groups in $\mathscr{J}_h$. In cases like this, a reasonable solution consists in postulating $G(\hat{p}_{hj}, \hat{\boldsymbol{p}}) \equiv 0$ for those groups, since all the involved $\hat{p}_{hj}$ equal the corresponding marginals $\hat{p}_{.j}$, be they 0 or 1. In other words, all the agencies will show the same observed performance, so it makes sense to nullify the contribution of the associate $G(\hat{p}_{hj}, \hat{\boldsymbol{p}})$ terms, in accordance with property *i*) in Sect. 3.1. In other cases, *ad hoc* adjustments need to be taken; see for example the problem discussed in Sect. 6.1.

In its most general formulation, the estimator in (11) involves non-linearities induced by the G function. While its exact distribution and moments are typically cumbersome, asymptotic properties can be derived via linearization methods. These are summarized by the following theorem:

Theorem 1 *Assume individual observations are realizations of uncorrelated binary variables* Y_{hji}, *with* $E(Y_{hji}) = p_{hj}$, *and G is a* C_2 *function in its arguments. Then, as* $N_{h.} \to \infty$ *and* $N_{.j} \to \infty$ *for all* $j \in \mathscr{J}$, *the* $\widehat{\mathrm{API}}_h^{(G)}$ *estimator is consistent and asymptotically unbiased for* $\mathrm{API}_h^{(G)}$; *moreover,*

$$\frac{\widehat{\mathrm{API}}_h^{(G)} - \mathrm{API}_h^{(G)}}{\hat{V}_0\big(\widehat{\mathrm{API}}_h^{(G)}\big)^{1/2}} \xrightarrow{d} N(0,1), \tag{12}$$

where $\hat{V}_0(\widehat{\mathrm{API}}_h^{(G)})$ *is a consistent estimator of the limiting variance*

$$V_0\big(\widehat{\mathrm{API}}_h^{(G)}\big) = \sum_{j \in \mathscr{J}_h} \{R_{hj}^2 + 2R_{hj}S_{h.j}Q_{hj}\}V(\hat{p}_{hj}) + \sum_{\ell \in \mathscr{J}} S_{h.\ell}^2 V(\hat{p}_{.\ell}), \tag{13}$$

with

$$V(\hat{p}_{hj}) = \frac{p_{hj}(1-p_{hj})}{N_{hj}}, \qquad V(\hat{p}_{.\ell}) = \sum_{m=1}^{H} Q_{m\ell}^2 \frac{p_{m\ell}(1-p_{m\ell})}{N_{m\ell}}, \tag{14}$$

$R_{hj} = W_{hj}A_{hj}$ *and* $S_{h.\ell} = \sum_{j \in \mathscr{J}_h} W_{hj}B_{hj.\ell}$, *being*

$$A_{hj} = \frac{\partial G(\hat{p}_{hj}, \hat{\boldsymbol{p}})}{\partial \hat{p}_{hj}}\bigg\|_{\hat{p}_{hj}=p_{hj},\, \hat{\boldsymbol{p}}=\boldsymbol{p}} \qquad B_{hj.\ell} = \frac{\partial G(\hat{p}_{hj}, \hat{\boldsymbol{p}})}{\partial \hat{p}_{.\ell}}\bigg\|_{\hat{p}_{hj}=p_{hj},\, \hat{\boldsymbol{p}}=\boldsymbol{p}}.$$

Proof is reported in "Appendix 1". According to the theorem, $\widehat{\mathrm{API}}_h^{(G)}$ is asymptotically normally distributed with limiting variance given by (13). Notice that only the marginal counts $N_{h.}$ and $N_{.j}$ (for all $j \in \mathscr{J}$) are required to tend to infinity for asymptotic theory to apply; the N_{hj} counts are not.

Equation (13) can be easily generalized to obtain the limiting covariance between the estimators of two different agencies, that is

$$\begin{aligned}\mathrm{Cov}_0\big(\widehat{\mathrm{API}}_h^{(G)}, \widehat{\mathrm{API}}_{h'}^{(G)}\big) = &\sum_{j \in \mathscr{J}_h} R_{hj}S_{h'.j}Q_{hj}V(\hat{p}_{hj}) + \sum_{j \in \mathscr{J}_{h'}} R_{h'j}S_{h.j}Q_{h'j}V(\hat{p}_{h'j}) \\ &+ \sum_{\ell \in \mathscr{J}} S_{h.\ell}S_{h'.\ell}V(\hat{p}_{.\ell}).\end{aligned} \tag{15}$$

5.1 Variance and covariance estimation

Consistent estimation of the right-hand side of (13) can be achieved by plugging-in estimated values $\hat{p}_{hj}$ and $\hat{p}_{.\ell}$ in place of the true probabilities p_{hj} and $p_{.\ell}$ appearing in the R_{hj} and $S_{h.\ell}$ terms as well as in the two variances in (14). Unbiased estimation of the latter variances is obtained with the usual sample size corrections, that is, replacing N_{hj} and $N_{m\ell}$ with $N_{hj}-1$ and $N_{m\ell}-1$, respectively. This setting extends to estimation of the covariance in (15).

When $N_{hj}=1$ for some groups j of the hth agency and/or $N_{m\ell}=1$ for some groups ℓ of any of the other agencies ($m \neq h$), the above corrections are not feasible. This is likely to occur when the overall sample size is small and the number of groups is relatively high. As a matter of fact, in these cases the limiting variance in (13) is systematically underestimated, since the estimate of $V(\hat{p}_{hj})$ and/or that of some components of $V(\hat{p}_{.\ell})$ is null.

To mitigate systematic underestimation of the index variance, we propose a strategy based on Strata Collapsing (SC), a well-known technique developed in the sample survey literature (Hansen et al., 1953; Rust & Kalton, 1987; Arnab, 2017). In practice, for each agency, pairs of similar case-mix adjusting groups are aggregated whenever at least one of the two groups contains just one unit. Two groups are similar when they have approximately the same marginal success probabilities. In this regard, we suggest to perform a preliminary ordering of groups according to the estimated $\hat{\boldsymbol{p}}=(\hat{p}_{.1},\ldots,\hat{p}_{.J})$ vector. In so doing, group ordering is constant and not influenced by agency-specific success rates, that are possibly unstable due to small sample size. Two aggregated groups form a collapsed group with weight given by the sum of their weights, and with a pooled success rate. The variance of the collapsed group can be estimated with the usual sample size correction, with a possible positive bias if the aggregated groups are not similar enough.

The approach outlined above yields a conservative solution for the estimation of the first summation in the right-hand side of (13). However, there still might be groups with a single user for agencies other than h, that might induce underestimation of some of the $V(\hat{p}_{.\ell})$ terms in the second summation. For this reason, we suggest to use the approximation

$$V(\hat{p}_{.\ell}) \approx \frac{p_{.\ell}(1-p_{.\ell})}{N_{.\ell}}, \tag{16}$$

which is prompted by recent studies on the behavior of the sum of a sequence of independent binomial random variables with distinct success probabilities (Butler & Stephens, 2017). The result is applicable since uncorrelated Bernoulli random variables, by standard results, are also independent. Clearly, an unbiased estimator of the right-hand side of (16) is $\hat{p}_{.\ell}(1-\hat{p}_{.\ell})/(N_{.\ell}-1)$, which is based on marginal sample sizes. Henceforth, we refer to this combined variance estimation approach, i.e., using SC to estimate the first summation in the right-hand side of (13) and Binomial Approximation (BA) to estimate the second summation, as to SC-BA approach. Such an approach can also be applied to estimation of the limiting covariance in (15), where similar sample size issues might occur.

6 Simulation study

The estimation framework presented in Sect. 5 hinges on a number of approximations involving both identification and estimation of the moments of API estimators. Thus, it is important to investigate the reliability of these approximations in finite samples. To this end, we here present a simulation study concerning the estimators of the neutral (E-type) indices for the three departure paradigms analyzed in Sect. 3. These are the $\text{API}_h^{(G_d)}$ and $\text{API}_h^{(G_r)}$ indices in (2) and (4) for the absolute deviation and relative deviation paradigms, respectively, as well as

$$\mathrm{API}_h^{(G_v)} = 1 + \sum_{j \in \mathscr{J}_h} W_{hj} \log \frac{p_{hj}}{p_{\cdot j}}, \tag{17}$$

which is a neutral index under the variation paradigm; see Table 2. Since index types change with the paradigm (Sect. 3.3), we remark that this simulation study is not limited to neutrality. In detail, $\mathrm{API}_h^{(G_d)}$ also acts like an I-type and F-type index, while $\mathrm{API}_h^{(G_r)}$ like a D-type and H-type one; see Sect. 4. As for $\mathrm{API}_h^{(G_v)}$, it is a B-type index for the relative deviation paradigm (Table 2). Moreover, it is easy to show that it becomes an A-type index when the absolute deviation paradigm is adopted.

6.1 Specifics of the study

As mentioned in Sect. 1, the simulation study is designed around the real-world dataset introduced in the context of the NH performance evaluation problem. While such a dataset has been already illustrated in Sect. 2, we just add that 312 (48%) of the 653 non-empty strata formed by NH-RUG crossing contain exactly one resident. Thus, the SC-BA approach outlined in Sect. 5.1 for (co)variance estimation appears to be necessary.

The statistical properties of the index estimators are achieved under the assumption of uncorrelated user-level random variables. Thus, it is first necessary to verify whether the absence of correlation between individual success indicators invoked in Sect. 5 is met in our data. To this purpose, a series of logistic mixed models (McCulloch & Searle, 2002; Faraway, 2016) can be used. In the statistical software `R` (R Core Team, 2021), these models can be fitted via the `glmer` command of the `lme4` package (Bates et al., 2015). In detail, individual indicators are regressed against RUG and NH membership, which can be alternatively included as fixed or random effects. A model including both membership effects as random effects can also be adopted. In any case, random effects are modeled via a normal distribution with null mean and unconstrained variance; we denote by σ_α^2 and σ_β^2 the variances of the NH and RUG random effects, respectively. Within this scheme, adjusted intraclass correlation coefficients indicate the degree of correlation between units in the same NH and/or RUG. These coefficients are given by $\rho_\alpha = \sigma_\alpha^2/(\sigma_\alpha^2 + \pi^2/3)$ and $\rho_\beta = \sigma_\beta^2/(\sigma_\beta^2 + \pi^2/3)$, where $\pi^2/3$ is the distribution-specific error variance for the logistic model in the absence of overdispersion (Nakagawa & Schielzeth, 2010; Nakagawa et al., 2017). For all the models fitted on the data at hand, the estimates of these intraclass coefficients are small (below 6%), indicating an almost negligible degree of correlation. However, it is worth to underline that the estimates of σ_α^2 and σ_β^2 always show a noticeable NH and RUG effect.

In order to implement our study, we draw simulated datasets that are generated from a set of strata-specific survival probabilities p_{hj} obtained from the corresponding $\hat{p}_{hj}$ rates observed in the original data. However, to avoid unrealistic values, extreme 0 and 1 rates (typically due to low sample sizes in some strata) are modified and set, respectively, to $\hat{p}_{\cdot j}/2$ and $(\hat{p}_{\cdot j}+1)/2$, where $\hat{p}_{\cdot j}$ $(j \in \mathscr{J})$ are the observed RUG-specific marginal rates. To sketch the asymptotic properties of the index and (co)variance estimators, increasing sample sizes are considered by multiplying the original stratum-specific cardinality N_{hj} by a Sample Size Factor (SSF) taking values 1, 2, 4, 8 and 16. In this way, the structure of the original dataset in terms of agency-specific case-mix distributions is not altered. Clearly, when SSF is greater than 1 there is no need of the collapsing method for variance estimation. For each sample size scenario, $N = 5000$ datasets are drawn and estimates of the indices as well as of their estimators' variances and covariances are computed.

Note that the $\widehat{\mathrm{API}}_h^{(G_v)}$ estimator of index (17) suffers from the indefiniteness issues mentioned in Sect. 5. In detail, when the estimate of p_{hj} is null for some strata, the whole estimator is not defined because of the presence of the logarithm function. Given the structure of the simulated data described above, this problem is likely to affect many strata, especially for the lower values of the SSF. To solve it, we adopt an heuristic approach based on replacing $\hat{p}_{hj} = 0$ with an arbitrary positive constant δ to be fine-tuned. Clearly, smaller values of δ correspond to a lesser extent of data manipulation, but when δ is too small $\log(\delta/\hat{p}_{.j})$ will be way lower than the bulk of the other $\log(\hat{p}_{hj}/\hat{p}_{.j})$ terms, which results in considerable downward bias for the overall estimator in finite samples. As a consequence, a compromise has to be found. In our study, we explore two alternatives by setting $\delta = 0.05$ and $\delta = 0.10$. This choice seems reasonable since RUG-specific marginal survival probabilities are always greater than 0.5. The two estimators deriving from this approach are denoted by $\widehat{\mathrm{API}}_h^{(G_v,5)}$ and $\widehat{\mathrm{API}}_h^{(G_v,10)}$.

6.2 Simulation results

The first slice of results is summarized by Figs. 1 and 2. These figures consist of a 2 × 2 panel where the four plots are split in two parts to include two estimators. Specifically, Fig. 1 refers to $\widehat{\mathrm{API}}_h^{(G_d)}$ (bottom part) and $\widehat{\mathrm{API}}_h^{(G_r)}$ (top part), while Fig. 2 to $\widehat{\mathrm{API}}_h^{(G_v,5)}$ (bottom part) and $\widehat{\mathrm{API}}_h^{(G_v,10)}$ (top part). All plots report the SSF on the x-axis and various estimator performance metrics, as a function thereof, on the y-axis. In detail, the top plots show the trends of Monte Carlo (MC) bias (left panel) and variance (right panel). The bottom-left plots focus on ratios between the average (across runs) of the variance estimates and the MC variance, while the bottom-right ones depict empirical coverage of 95% Confidence Intervals (CIs) based on the normal approximation. In each plot, the grey stripe comprehends the values of the y-axis quantity for all the $H = 47$ NHs: the border lines represent minimum and maximum values, while the inner line corresponds to median values.

The top plots of Fig. 1 provide empirical evidence about the consistency of the $\widehat{\mathrm{API}}_h^{(G_d)}$ and $\widehat{\mathrm{API}}_h^{(G_r)}$ estimators, with both bias and variance approaching zero as the SSF grows. As a matter of fact, these estimators exhibit very little bias and variance even at the lower values of the SSF. Moving to variance estimation, a certain amount of over-estimation is present for SSF=1, as shown by the bottom-left plot. In principle, this might be due to the approximations induced either by linearization or by the SC-BA approach; see Sect. 5. With this regard, it is important to mention that the $\widehat{\mathrm{API}}_h^{(G_d)}$ estimator does not require linearization since it does not involve any kind of non-linearity. Thus, it is an unbiased estimator and exact expressions for its variance and covariance are available. Since the performances of the estimators of $V(\widehat{\mathrm{API}}_h^{(G_r)})$ and $V(\widehat{\mathrm{API}}_h^{(G_d)})$ are similar, we argue that the effect of linearization for $\widehat{\mathrm{API}}_h^{(G_r)}$ is restrained. Despite variance over-estimation, a modest degree of under-coverage—possibly deriving from limited adherence to the normal distribution—is spotted for lower values of the SSF; see the bottom-right plot in the figure.

With respect to the $\widehat{\mathrm{API}}_h^{(G_v,5)}$ and $\widehat{\mathrm{API}}_h^{(G_v,10)}$ estimators (Fig. 2), for SSF=1 and SSF=2 it is possible to observe a non-negligible negative MC bias as well as an increase in MC variance (please acknowledge the difference in the scales between Figs. 1 and 2). Nevertheless, consistency is achieved essentially at the same rate as $\widehat{\mathrm{API}}_h^{(G_d)}$ and $\widehat{\mathrm{API}}_h^{(G_r)}$. These worse estimator performances are clearly due to the *ad hoc* replacement of the null $\hat{p}_{hj}$ estimates with δ; see the discussion at the end of Sect. 6.1. In this regard, it is immediate to see that $\widehat{\mathrm{API}}_h^{(G_v,10)}$ outperforms $\widehat{\mathrm{API}}_h^{(G_v,5)}$: as expected, a higher value of δ makes the replacing $\log(\delta/\hat{p}_{.j})$ terms less extreme in magnitude. It is also important to clarify that variance estimation ignores the

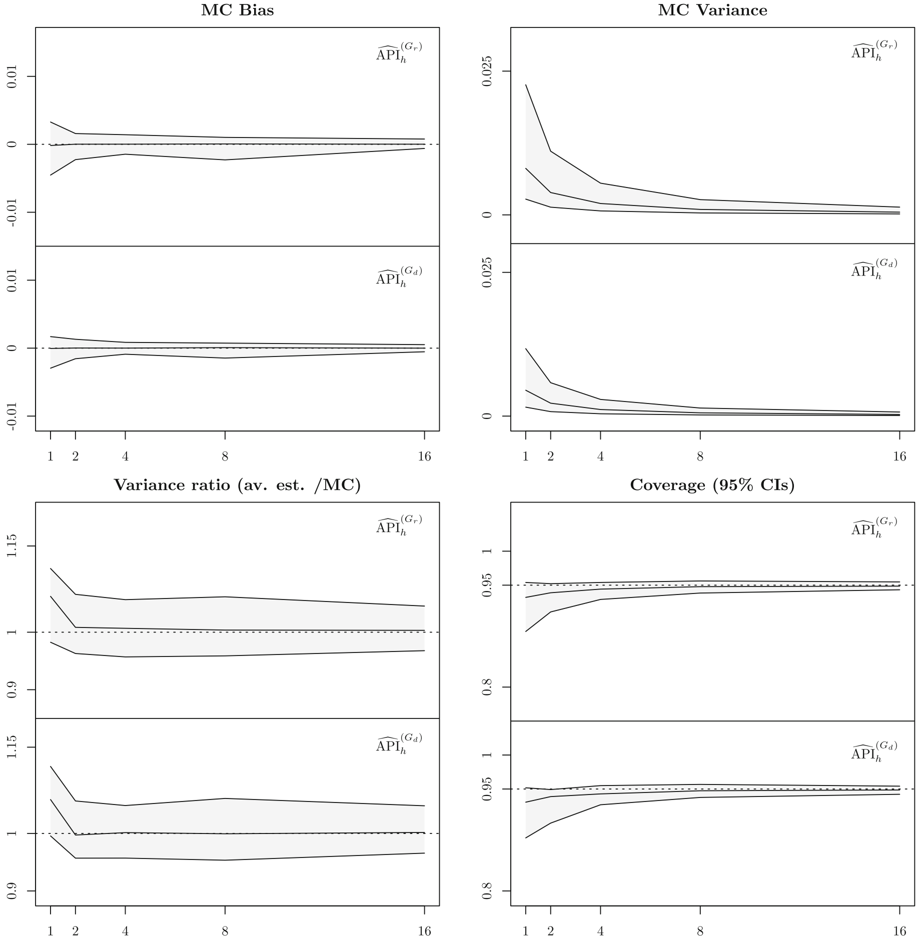

Fig. 1 Simulation results for $\widehat{\mathrm{API}}_h^{(G_d)}$ and $\widehat{\mathrm{API}}_h^{(G_r)}$

ad hoc replacements. In other words, a unique variance estimator is adopted, which explains why the variance ratios are higher for the less variable estimator, i.e., $\widehat{\mathrm{API}}_h^{(G_v,10)}$. Incidentally, this fact contributes to mitigate under-coverage due to downward bias, which instead is rather relevant for $\widehat{\mathrm{API}}_h^{(G_v,5)}$. Alternative strategies including δ adjustments also in variance estimation were tested. However, results from these strategies (not shown) highlight a considerably poorer performance of variance estimators for the first two SSFs.

Figure 3 shows further evidence from the simulation study. In particular, it focusses on the estimated pairwise comparisons between NHs, reporting summaries of the $H(H-1)/2 = 1081$ Pairwise Differences (PDs) between the estimated indices. The figure contains a 3 × 2 panel, with the left-hand column pertaining to $\widehat{\mathrm{API}}_h^{(G_d)}$ and $\widehat{\mathrm{API}}_h^{(G_r)}$ and the right-hand column devoted to $\widehat{\mathrm{API}}_h^{(G_v,5)}$ and $\widehat{\mathrm{API}}_h^{(G_v,10)}$. In each column, the included plots depict: (i) the percentage of estimated PDs concordant with the true ones, (ii) the percentage of estimated PDs that are concordant with the true ones and also Statistically Significant (SS), and (iii) the percentage of SS PDs that are discordant with the truth. Similarly to Figs. 1 and 2, these percentages are plotted as a function of the SSF, and grey stripes with black lines for minimum, median

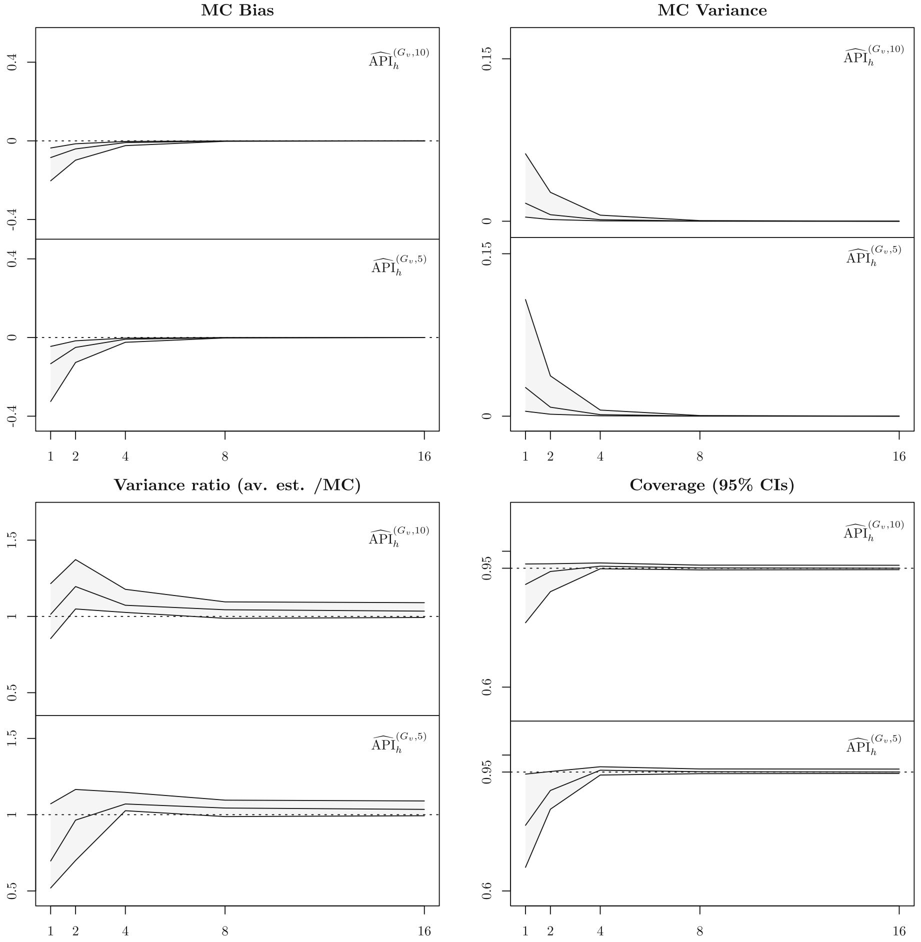

Fig. 2 Simulation results for $\widehat{\mathrm{API}}_h^{(G_v,5)}$ and $\widehat{\mathrm{API}}_h^{(G_v,10)}$

and maximum values are pictured. However, these now summarize the distributions across the $N = 5000$ simulation runs of the summary statistics for the 1081 PDs. Inference on PDs is performed by testing, through a standard z-test, the hypothesis that they are null at the 5% significance level. To this end, the covariance terms in (15) are estimated as described in Sect. 5.1 along with the variances, so that the estimated variance of each PD can be computed. Like for variance estimation, for $\widehat{\mathrm{API}}_h^{(G_v,5)}$ and $\widehat{\mathrm{API}}_h^{(G_v,10)}$ the covariance estimator ignores the δ adjustments.

Overall, the trends in the first row of Fig. 3 show a remarkable level of concordance with the truth. With this regard, we point out that there are no relevant differences among estimators. Indeed, the effect of the negative bias characterizing $\widehat{\mathrm{API}}_h^{(G_v,5)}$ and $\widehat{\mathrm{API}}_h^{(G_v,10)}$ vanishes when differences are taken. When statistical significance comes into play (second row), as expected, lower percentages are attained at smaller SSFs. Conversely, when SSF is greater than 8 the obtained values are considerably closer to those of the first row, especially for $\widehat{\mathrm{API}}_h^{(G_v,5)}$ and $\widehat{\mathrm{API}}_h^{(G_v,10)}$. Finally, the percentages of SS PDs discordant with the truth (third row) are always very small, which makes the likelihood of drawing wrong inferences on pairwise comparisons

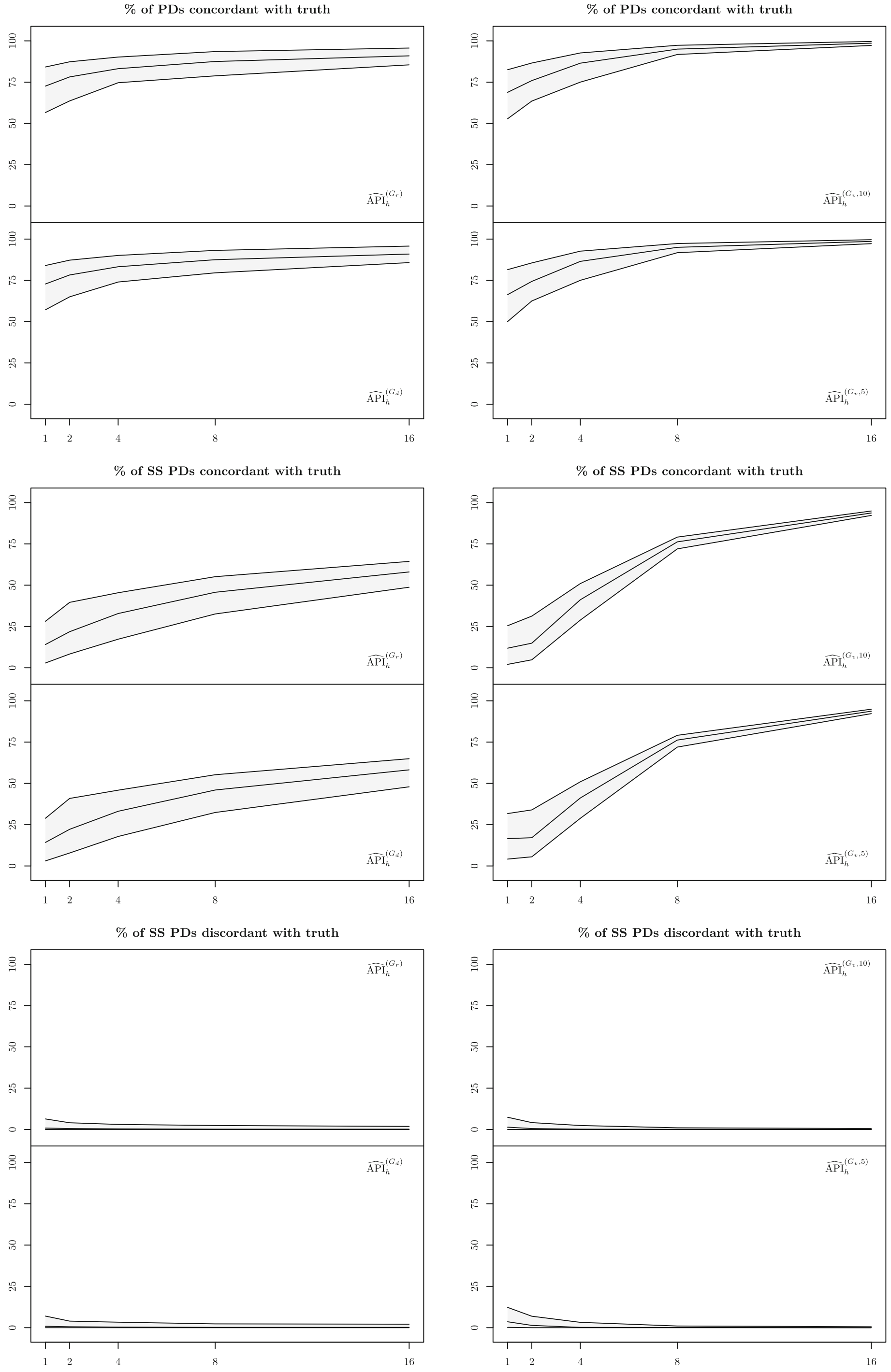

Fig. 3 Correctness of pairwise comparisons for the four estimators

desirably low. In other words, when a PD is SS, it almost always shows the correct ranking between the two NHs involved.

7 Conclusions

When the evaluation of the performance of a set of agencies operating on different users is undertaken based on binary outcomes, suitable indices exist for settings where users are divided into groups in order to account for case-mix, that is, for the different composition of users with respect to personal characteristics influencing the outcome variables. Motivated by a real-word example as well as by the recent discussion on the appropriateness of traditional performance measures like risk-adjusted rate ratios/differences, we introduce a class of case-mix adjusting probability-based performance indices and a taxonomy for the elements in the class. Our taxonomy is built around the concepts of balancing and level-invariance, leading to a nine-fold classification scheme. We show by means of an application how this range of index types can flexibly satisfy the diverse evaluative needs that might rise. We also illustrate how the mathematical formalization of these concepts—and thus our classification scheme—depends on the so-called *departure paradigm*, that is, on the metric adopted to quantify the distance between agency-specific and marginal probabilities of the outcome of interest. In our view, acknowledging and understanding such a dependence can help analysts address a number of interpretive issues, possibly including those related to the above mentioned critique to traditional rate contrasts. We focus on the three major departure paradigms: absolute deviations, variations and relative deviations; the proposed class of indices encompasses indices of all nine types for these three paradigms.

Whatever the index chosen for the evaluation exercise, the problem of estimating its value for each agency in real-world applications arises. The inferential framework presented in Sect. 5 relies on the absence of correlation between any pair of outcome variables at the user level. In principle, this assumption can be relaxed at different levels by introducing varying degrees of correlation, be they for units belonging to the same agency, group or stratum. However, such an approach would lead to substantial complication of the formulae for index estimators' variances and covariances, and possibly also to problems with the simultaneous identification of this set of correlation coefficients from sample data. While these issues might represent a future stream of research, we believe that the formulae provided here should be used once, like in our setting, correlation has been ruled out. Indeed, variance/covariance estimation is likely to benefit from the adoption of this simplified framework.

The behavior of the estimators of the proposed indices and of their variance and covariance estimators has been analyzed through an extensive simulation study. Simulation results confirm the consistency of the proposed estimators. Furthermore, empirical evidence shows that when the pairwise difference between two agencies is statistically significant, it almost always provides the correct ranking between them.

In principle, our framework could be easily adapted to continuous outcomes, with probabilities replaced by expectations/sample means. In that case, the SC-BA approach to variance estimation outlined in Sect. 5.1 would clearly need adjustments depending on the probabilistic distribution assumed for the underlying random variables. Such an extension would open the way to further application contexts, where the need to account for the departure paradigm has been sensed (though not explicitly formalized) also by other authors; see for example the discussion about well-being composite indicators in Fortuna et al. (2022). Conversely, extension to categorical data would posit a number of additional challenges. These are essentially linked to the identification of the category associated to the best performance level (if any) and to the quantification of dissimilarities among categories. With this regard, integration with recently proposed methods in this area (van de Velden et al., 2023) would represent an interesting development.

Acknowledgements We thank Regione Umbria for sharing the data and Fondazione Cassa di Risparmio di Perugia for providing financial support.

Funding Open access funding provided by Università degli Studi di Firenze within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

A Proof of Theorem 1

For a given agency set $\mathscr{H}$, consider a sequence of samples of users indexed in $t = 1, 2, \dots, \infty$ such that $N_{h.}^{(t)} = tN_{h.}$ for $h \in \mathscr{H}$ and $N_{.j}^{(t)} = tN_{.j}$ for all $j \in \mathscr{J}$. Extending dependence on t to all other pieces of notation, for a given agency h we can denote by

$$\hat{\boldsymbol{p}}_h^{(t)} = (\{\hat{p}_{hj}^{(t)}\}_{j \in \mathscr{J}_h^{(t)}}, \hat{p}_{.1}^{(t)}, \dots, \hat{p}_{.J}^{(t)})^\top$$

the vector of estimators for the tth sample, and by $\boldsymbol{p}_h^{(t)}$ the vector of the corresponding parameters (notice that the latter depends on t only through $\mathscr{J}_h^{(t)}$, whereas the probabilities are assumed to be fixed). Specifically, we have

$$\hat{p}_{hj}^{(t)} = \frac{\sum_{i=1}^{N_{hj}^{(t)}} Y_{hji}^{(t)}}{N_{hj}^{(t)}}, \qquad \hat{p}_{.j}^{(t)} = \frac{\sum_{h \in \mathscr{H}} \sum_{i=1}^{N_{hj}^{(t)}} Y_{hji}^{(t)}}{N_{.j}^{(t)}} = \sum_{h \in \mathscr{H}} Q_{hj}^{(t)} \hat{p}_{hj}^{(t)}.$$

It is immediate to see that these estimators are unbiased for the respective parameters, with variances—given the absence of correlation among the $Y_{hji}^{(t)}$ variables—equal to

$$V(\hat{p}_{hj}^{(t)}) = \frac{p_{hj}(1 - p_{hj})}{N_{hj}^{(t)}} \qquad V(\hat{p}_{.j}^{(t)}) = \sum_{h \in \mathscr{H}} (Q_{hj}^{(t)})^2 V(\hat{p}_{hj}^{(t)}).$$

Also, the only non-null covariances are $\mathrm{Cov}(\hat{p}_{hj}^{(t)}, \hat{p}_{.j}^{(t)}) = Q_{hj} V(\hat{p}_{hj}^{(t)})$ for $j \in \mathscr{J}_h^{(t)}$.

The second-order Taylor expansion of $\widehat{\mathrm{API}}_{th}^{(G)}$ returns

$$\begin{aligned}\widehat{\mathrm{API}}_{th}^{(G)} &= \mathrm{API}_{th}^{(G)} + \sum_{j\in\mathscr{J}_h^{(t)}} R_{hj}^{(t)}\big(\hat{p}_{hj}^{(t)} - p_{hj}\big) + \sum_{\ell\in\mathscr{J}} S_{h.l}^{(t)}\big(\hat{p}_{.\ell}^{(t)} - p_{.\ell}\big)\\ &\quad + \frac{1}{2}\sum_{j\in\mathscr{J}_h^{(t)}} W_{hj}^{(t)} A_{2,hj}^{(t)}\big(\hat{p}_{hj}^{(t)} - p_{hj}\big)^2\\ &\quad + \frac{1}{2}\sum_{j\in\mathscr{J}_h^{(t)}} W_{hj}^{(t)} \sum\sum_{\ell,\ell'\in\mathscr{J}} B_{2,hj.\ell\ell'}^{(t)}\big(\hat{p}_{.\ell}^{(t)} - p_{.\ell}\big)\big(\hat{p}_{.\ell'}^{(t)} - p_{.\ell'}\big)\\ &\quad + \sum_{j\in\mathscr{J}_h^{(t)}} W_{hj}^{(t)} \sum_{\ell\in\mathscr{J}} C_{2,hj.\ell}^{(t)}\big(\hat{p}_{hj}^{(t)} - p_{hj}\big)\big(\hat{p}_{.\ell}^{(t)} - p_{.\ell}\big)\\ &\quad + o\big[\big(\hat{\boldsymbol{p}}_h^{(t)} - \boldsymbol{p}_h^{(t)}\big)^\top \boldsymbol{\Omega}_h^{(t)}\big(\hat{\boldsymbol{p}}_h^{(t)} - \boldsymbol{p}_h^{(t)}\big)\big],\end{aligned} \tag{18}$$

where the coefficients for second-order derivates are

$$A_{2,hj}^{(t)} = \frac{\delta^2 G\big(\hat{p}_{hj}^{(t)}, \hat{\boldsymbol{p}}^{(t)}\big)}{\delta^2 \hat{p}_{hj}^{(t)}}\Bigg\|_{\hat{\boldsymbol{p}}_h^{(t)}=\boldsymbol{p}_h^{(t)}}, \qquad B_{2,hj.\ell\ell'}^{(t)} = \frac{\delta^2 G\big(\hat{p}_{hj}^{(t)}, \hat{\boldsymbol{p}}^{(t)}\big)}{\delta \hat{p}_{.\ell}^{(t)}\,\delta \hat{p}_{.\ell'}^{(t)}}\Bigg\|_{\hat{\boldsymbol{p}}_h^{(t)}=\boldsymbol{p}_h^{(t)}}$$

and

$$C_{2,hj.\ell}^{(t)} = \frac{\delta^2 G\big(\hat{p}_{hj}^{(t)}, \hat{\boldsymbol{p}}^{(t)}\big)}{\delta \hat{p}_{hj}^{(t)}\,\delta \hat{p}_{.\ell}^{(t)}}\Bigg\|_{\hat{\boldsymbol{p}}_h^{(t)}=\boldsymbol{p}_h^{(t)}}$$

and $\boldsymbol{\Omega}_h^{(t)} = \mathrm{diag}(\mathbf{W}_h^{(t)}, \mathbf{I}_J)$, with $\mathbf{I}_J$ being the identity matrix of order J and $\mathbf{W}_h^{(t)} = \mathrm{diag}(\{W_{hj}^{(t)}\}_{j\in\mathscr{J}_h^{(t)}})$.

Taking the expectation of both members of (18) and rearranging gives, after some algebra,

$$\begin{aligned}\mathrm{Bias}\big(\widehat{\mathrm{API}}_{th}^{(G)}\big) &= E\big(\widehat{\mathrm{API}}_{th}^{(G)}\big) - \mathrm{API}_{th}^{(G)}\\ &= \frac{1}{2}\sum_{j\in\mathscr{J}_h^{(t)}} W_{hj}^{(t)} A_{2,hj}^{(t)} V\big(\hat{p}_{hj}^{(t)}\big) + \frac{1}{2}\sum_{j\in\mathscr{J}_h^{(t)}} W_{hj}^{(t)} \sum\sum_{\ell,\ell'\in\mathscr{J}} B_{2,hj.\ell\ell'}^{(t)}\mathrm{Cov}\big(\hat{p}_{.\ell}^{(t)}, \hat{p}_{.\ell'}^{(t)}\big)\\ &\quad + \sum_{j\in\mathscr{J}_h^{(t)}} W_{hj}^{(t)} \sum_{\ell\in\mathscr{J}} C_{2,hj.\ell}^{(t)}\mathrm{Cov}\big(\hat{p}_{hj}^{(t)}, \hat{p}_{.\ell}^{(t)}\big) + o(t^{-1})\\ &= \frac{1}{2}\sum_{j\in\mathscr{J}_h^{(t)}} A_{2,hj}^{(t)} \frac{p_{hj}(1-p_{hj})}{N_{h.}^{(t)}}\\ &\quad + \frac{1}{2}\sum_{j\in\mathscr{J}_h^{(t)}} W_{hj}^{(t)} \frac{\sum_{m\in\mathscr{H}}\sum_{\ell\in\mathscr{J}} B_{2,hj.\ell\ell}^{(t)} Q_{m\ell}^{(t)} p_{m\ell}(1-p_{m\ell})}{N_{.j}^{(t)}}\\ &\quad + \sum_{j\in\mathscr{J}_h^{(t)}} W_{hj}^{(t)} \frac{C_{2,hj.j}^{(t)} p_{hj}(1-p_{hj})}{N_{.j}^{(t)}} + o(t^{-1})\\ &= O(t^{-1}).\end{aligned} \tag{19}$$

To derive the limiting variance, it is sufficient to consider the first-order expansion

$$\widehat{\mathrm{API}}_{th}^{(G)} = \mathrm{API}_{th}^{(G)} + \sum_{j \in \mathscr{J}_h^{(t)}} R_{hj}^{(t)} \big(\hat{p}_{hj}^{(t)} - p_{hj}^{(t)}\big) + \sum_{\ell \in \mathscr{J}} S_{h.l}^{(t)} \big(\hat{p}_{.\ell}^{(t)} - p_{.\ell}^{(t)}\big)$$
$$+\, o\Big[\Big\{\big(\hat{\boldsymbol{p}}_h^{(t)} - \boldsymbol{p}_h^{(t)}\big)^{\top} \boldsymbol{\Omega}_h^{(t)} \big(\hat{\boldsymbol{p}}_h^{(t)} - \boldsymbol{p}_h^{(t)}\big)\Big\}^{1/2}\Big].$$

Computing the variance of both sides and rearranging returns

$$V\big(\widehat{\mathrm{API}}_{th}^{(G)}\big) = \sum_{j \in \mathscr{J}_h^{(t)}} \Big\{\big(R_{hj}^{(t)}\big)^2 + 2R_{hj}^{(t)} S_{h.j}^{(t)} Q_{hj}^{(t)}\Big\} V\big(\hat{p}_{hj}^{(t)}\big) + \sum_{\ell \in \mathscr{J}} \big(S_{h.\ell}^{(t)}\big)^2 V\big(\hat{p}_{.\ell}^{(t)}\big) + o(t^{-1}),$$

which agrees with (13).

Recalling the definitions of $R_{hj}^{(t)}$ and $S_{h.j}^{(t)}$ from the theorem's statement, it is easy to show that the right-hand side in the above equation is $o(t^{-1})$ and, thus, the standard deviation of the $\widehat{\mathrm{API}}_{th}^{(G)}$ estimator is $o(t^{-1/2})$. This fact combined with (19) proves that $\widehat{\mathrm{API}}_{th}^{(G)}$ is consistent and asymptotically unbiased, since

$$\lim_{t \to \infty} \frac{\mathrm{Bias}\big(\widehat{\mathrm{API}}_{th}^{(G)}\big)}{\sqrt{V\big(\widehat{\mathrm{API}}_{th}^{(G)}\big)}} = 0$$

The convergence statement in (12) follows from the fact that, by standard results, uncorrelated binary variables are also independent. Thus, the linear approximation of $\widehat{\mathrm{API}}_{th}^{(G)}$ can be rewritten as a linear combination of the independent sample proportions $\hat{p}_{hj}^{(t)}$. For $t \to \infty$, the sampling distributions of these sample proportions either converge to the normal distribution by virtue of central limit theorem (in case the $w_{hj}^{(t)}$ weights converge to a non-null constant), or do not contribute to the limiting distribution of $\widehat{\mathrm{API}}_{th}^{(G)}$ (if the $w_{hj}^{(t)}$ weights converge to zero). Finally, for any consistent estimator $\hat{V}_{0}(\widehat{\mathrm{API}}_{h}^{(G)})$, Slutsky's Theorem can be invoked to prove (12).

References

Aday, L. A., Begley, C. E., Lairson, D. R., Slater, C. H., Richard, A. J., & Montoya, I. D. (1999). A framework for assessing the effectiveness, efficiency, and equity of behavioral healthcare. *American Journal of Managed Care, 5*(8), SP25–SP43.

Arnab, R. (2017). *Survey sampling theory and applications*. Academic Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Berlowitz, D. R., Ash, A. S., Brandeis, G. H., Brand, H. K., Halpern, J. L., Moskowitz, M. A., & Gwaltney, J. M., Jr. (1996). Rating long-term care facilities on pressure ulcer development: Importance of case-mix adjustment. *Annals of Internal Medicine, 124*(6), 557–563.

Boubaker, S., Le, T. D., Ngo, T., & Manita, R. (2023). Predicting the performance of MSMEs: A hybrid DEA-machine learning approach. *Annals of Operations Research*. https://doi.org/10.1007/s10479-023-05230-8

Broussard, D. M., & Reiter, K. L. (2020). *Estimated reduction in CAH profitability from loss of cost-based reimbursement for swing beds*. Technical report, North Carolina Rural Health Research Program.

Butler, K., & Stephens, M. A. (2017). The distribution of a sum of independent binomial random variables. *Methodology and Computing in Applied Probability, 19*(2), 557–571.

Carpenter, I., & Hirdes, J. P. (2013). Using interRAI assessment systems to measure and maintain quality of long-term care. In: A good life in old age? Monitoring and improving quality in long-term care. OECD Health Policy Studies, chaper 3 (pp. 93–139).

Carpita, M., Pasca, P., Arima, S., & Ciavolino, E. (2023). Clustering of variables methods and measurement models for soccer players' performances. *Annals of Operations Research*. https://doi.org/10.1007/s10479-023-05185-w

Castro, R. A., Oliveira, P. N., Silva Portela, C., Camanho, A. S., & Queiroz-e-Melo, J. (2015). Benchmarking clinical practice in surgery: Looking beyond traditional mortality rates. *Health Care Management Science, 18*(4), 431–443.

Chessa, A., D'Urso, P., De Giovanni, L., Vitale, V., & Gebbia, A. (2022). Complex networks for community detection of basketball players. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-04647-x

Chu, C. (1994). Resource intensity weighing and case mix grouping: Assumptions and implications for health service performance evaluation. *Healthcare Management Forum, 7*(1), 24–31.

Curtin, L. R. (1995). Direct standardization (age-adjusted death rates). 6, US Department of Health and Human Services, Public Health Service

Draper, D., & Gittoes, M. (2004). Statistical analysis of performance indicators in UK higher education. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 167*(3), 449–474.

Faraway, J. J. (2016). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Chapman and Hall/CRC.

Fortuna, F., Naccarato, A., & Terzi, S. (2022). Country rankings according to well-being evolution: Composite indicators from a functional data analysis perspective. *Annals of Operations Research*. https://doi.org/10.1007/s10479-022-05072-w

Fries, B. E., Schneider, D. P., Foley, W. J., Gavazzi, M., Burke, R., & Cornelius, E. (1994). Refining a case-mix measure for nursing homes: Resource Utilization Groups (RUG-III). *Medical Care, 32*(7), 668–685.

Hansen, M. H., Hurwitz, W. N., & Madow, W. G. (1953). *Sample survey methods and theory. Methods and applications* (Vol. I). Wiley.

Inskip, H., Beral, V., Fraser, P., & Haskey, J. (1983). Methods for age-adjustment of rates. *Statistics in Medicine, 2*(4), 455–466.

Julious, S. A., Nicholl, J., & George, S. (2001). Why do we continue to use standardized mortality ratios for small area comparisons? *Journal of Public Health, 23*(1), 40–46.

Kitagawa, E. M. (1964). Standardized comparisons in population research. *Demography, 1*(1), 296–315.

Kruk, M. E., & Freedman, L. P. (2008). Assessing health system performance in developing countries: A review of the literature. *Health Policy, 85*(3), 263–276.

Longford, N. T. (2020). Performance assessment as an application of causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 183*(4), 1363–1385.

Maso, G., Alberico, S., Monasta, L., Ronfani, L., Montico, M., Businelli, C., Soini, V., Piccoli, M., Gigli, C., Domini, D., & Fiscella, C. (2013). The application of the Ten Group classification system (TGCS) in caesarean delivery case mix adjustment. A multicenter prospective study. *PLoS ONE, 8*(6), e62364.

McCulloch, C. E., & Searle, S. R. (2002). *Generalized, linear and mixed models*. Wiley.

Metnitz, P. G., Lang, T., Vesely, H., Valentin, A., & Le Gall, J. R. (2000). Ratios of observed to expected mortality are affected by differences in case mix and quality of care. *Intensive Care Medicine, 26*(10), 1466–1472.

Montanari, G. E., & Doretti, M. (2019). Ranking nursing homes' performances through a latent Markov model with fixed and random effects. *Social Indicators Research, 146*(1–2), 307–326.

Montanari, G. E., Doretti, M., & Marino, M. F. (2022). Model-based two-way clustering of second-level units in ordinal multilevel latent Markov models. *Advances in Data Analysis and Classification, 16*(2), 457–485.

Moreno, R., & Apolone, G. (1997). Impact of different customization strategies in the performance of a general severity score. *Critical Care Medicine, 25*(12), 2001–2008.

Moreno, R., Apolone, G., & Reis Miranda, D. (1998). Evaluation of the uniformity of fit of general outcome prediction models. *Intensive Care Medicine, 24*(1), 40–47.

Nakagawa, S., & Schielzeth, H. (2010). Repeatability for Gaussian and non-Gaussian data: A practical guide for biologists. *Biological Reviews, 85*(4), 935–956.

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface, 14*(134), 1–11.

Pouw, M. E., Peelen, L. M., Lingsma, H. F., Pieter, D., Steyerberg, E., Kalkman, C. J., & Moons, K. G. (2013). Hospital standardized mortality ratio: Consequences of adjusting hospital mortality with indirect standardization. *PLoS ONE, 8*(4), e59160.

Punelli, D., & Williams, S. (2013). *Nursing facility reimbursement and regulation*. Technical report, Research Department, Minnesota House of Representatives.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Romano, P. S. (2000). Should health plan quality measures be adjusted for case mix? *Medical Care, 38*(10), 977–980.

Rust, K., & Kalton, G. (1987). Strategies for collapsing strata for variance estimation. *Journal of Official Statistics, 3*(1), 69–81.

Schokkaert, E., & Van de Voorde, C. (2009). Direct versus indirect standardization in risk adjustment. *Journal of Health Economics, 28*(2), 361–374.

van de Velden, M., D'Enza, A. I., Markos, A., & Cavicchia, C. (2023). A general framework for implementing distances for categorical variables. arXiv preprint arXiv:2301.02190

Wolfenden, H. H. (1962). On the theoretical and practical considerations underlying the direct and indirect standardization of death rates. *Population Studies, 16*(2), 188–190.

Wray, N. P., Hollingsworth, J. C., Petersen, N. J., et al. (1997). Case-mix adjustment using administrative databases: A paradigm to guide future research. *Medical Care Research and Review, 54*(3), 326–356.