CSEF

Centre for Studies in Economics and Finance

WORKING PAPER NO. 805

Evaluating the Quality of Hospital Services: A Matched, Precision-Weighted Comparison of Italian Academic Health Centres and Non-Academic Public Hospitals

Massimo Aria, Sergio Beraldo, Michela Collaro, and Agostino Gnasso

November 2026

University of Naples Federico II

University of Salerno

Bocconi University, Milan

CSEF - Centre for Studies in Economics and Finance
DEPARTMENT OF ECONOMICS AND STATISTICS – UNIVERSITY OF NAPLES FEDERICO II
80126 NAPLES - ITALY
Tel. and fax +39 081 675372 – e-mail: csef@unina.it
ISSN: 2240-9696

CSEF
Centre for Studies in Economics and Finance

WORKING PAPER NO. 805

Evaluating the Quality of Hospital Services: A Matched, Precision-Weighted Comparison of Italian Academic Health Centres and Non-Academic Public Hospitals

[*]**Massimo Aria,** [†]**Sergio Beraldo,** [‡]**Michela Collaro, and** [§]**Agostino Gnasso**

Abstract

Italian Academic Health Centres (AHCs), the university hospitals and research hospitals (IRCCS), are widely assumed to provide better care than other hospitals. The evidence comes mostly from comparisons of whole institutions, which differ in size and services. We compare each AHC with a non-academic public hospital of the same region and similar structure, using ten years (2015–2024) of risk-adjusted indicators from the National Outcomes Programme (Programma Nazionale Esiti), structural data and 214,848 articles published by the AHCs. Each pair is compared only on the indicators both hospitals report, and a multilevel model weights each rate by its precision, so that rates computed on few patients count less. We also ask whether AHCs do better in the clinical areas where their research is concentrated. Without structural adjustment, AHCs do better than the other public hospitals of their region (0.144standard deviations of the indicators). For the 22 of the 47 AHCs that have a comparable control, the difference is−0.033 (95% CI−0.121 to 0.054), which excludes an advantage larger than 0.054. The difference is stable over the decade and through COVID-19, and it is not larger in the areas where AHCs do more research (−0.029 per standard deviation of research specialisation, 95% CI−0.116 to 0.059). AHCs and their controls differ service by service, without a general quality gap. The 9 largest teaching hospitals have no comparable non-academic hospital in their region and remain outside the comparison. The comparison is descriptive and concerns quality, not efficiency.

JEL classification: I11, I12, I23, C21, C23.

Keywords: Academic Health Centres, Composite Indicators, Hospital Performance, Matched Comparison, Multilevel Model.

Acknowledgments: This research has been financed by the following research projects: PRIN-2022 “SCIK-HEALTH” (Project Code: 2022825Y5E; CUP: E53D2300611006); PRIN-2022 PNRR “The value of scientific production for patient care in Academic Health Science Centres” (Project Code: P2022RF38Y; CUP: E53D23016650001).

[*] University of Naples Federico II. E-mail: aria@unina.it

[†] University of Naples Federico II and CSEF. E-mail: s.beraldo@unina.it

[‡] University of Naples Federico II. E-mail: michela.collaro@unina.it

[§] University of Naples Federico II. E-mail: agostino.gnasso@unina.it

1 | INTRODUCTION

Italian Academic Health Centres (AHCs), comprising nine fully integrated university hospitals, seventeen affiliated university hospitals and twenty-three IRCCS (Istituti di Ricovero e Cura a Carattere Scientifico) research institutes recognised by the Ministry of Health, have long been assumed to deliver care of higher clinical quality because of their research activity and teaching mission. Whether the assumption holds matters to patients choosing where to be treated, to regions deciding where to concentrate complex care, and to the public bodies that fund research and teaching in hospitals. Much of the evidence for this assumption comes from comparisons of whole institutions, such as teaching with non-teaching hospitals or more with less research-active hospitals, which find small advantages for the former[1,2,3]. Whole institutions, however, differ in size and case mix, and their research spans many clinical areas. An academic hospital may therefore look better simply because it is a large, specialised hospital, and an advantage averaged over all its activity does not say whether it arises in the fields where the hospital does research. This paper asks two narrower questions with national outcome data. First, do AHCs deliver better care than non-academic public hospitals of the same region and similar structure? Second, is any difference larger in the clinical areas where the AHC's research is concentrated? The second question tests the mechanism usually invoked for the academic advantage: if research improves care, it should do so mainly in the fields where the hospital's clinicians do research. We answer both with ten years (2015–2024) of the risk-adjusted indicators that the National Outcomes Programme (Programma Nazionale Esiti, PNE) publishes for every Italian hospital, Ministry of Health data on hospital structure, and the 214,848 articles published by the AHCs. Answering these questions is harder than it seems. Unlike the quality of a manufactured product, the quality of a service cannot be inspected directly[4]: a hospital's quality is known only through a panel of indicators, each covering one condition or procedure, that are noisy, unequally applicable across providers, and of debatable relative importance[5,6]. Before any statistic is computed, a comparison between two groups of hospitals must therefore settle three questions, and each shapes the design.

On which indicators are hospitals compared? Hospitals are not evaluated on the same indicators: a cancer institute is assessed on oncological surgery, a general hospital on heart failure and childbirth. Averaging each hospital over its own indicators, as composite scores usually do, compares different baskets of services, so that differences in the mix of services are read as differences in quality. We compare each AHC with its control only on the indicators that both report in the same year. *With which hospitals is each AHC compared?* AHCs are larger than other hospitals, offer more specialised services and, more often, host the second-level emergency departments of their region. We match each AHC with a non-academic public hospital of the same region that is similar in size and service profile, check the result against designs that also hold the emergency role fixed, and exclude from the control pool every hospital with university or research status. The estimand is descriptive: the difference between AHCs and structurally similar non-academic hospitals, not the causal effect of academic status. *How are real differences separated from chance?* Many outcome rates rest on few patients, so that much of their variation is sampling noise. PNE publishes the standard error of every rate. The main estimate comes from a multilevel model of the paired differences that uses this known sampling variance, so that a rate computed on few patients counts less, and that separates differences persisting over the years from transient ones. Because the model covers 2015–2024, it also shows whether the difference changes over time. The second question requires knowing in which clinical areas each AHC does research. We assign each article published by the AHCs between 2005 and 2024 to the clinical areas covered by the indicators, such as stroke or cholecystectomy, using its subject headings and its title. For each AHC and area we then compare the share of the AHC's articles in that area with the same share for all AHCs together: an AHC with 6% of its articles on stroke, against 3% for all AHCs, is relatively specialised in stroke. We use this relative research specialisation rather than the number of articles, because the number of articles attributed to a hospital depends on how authors declare their affiliation, and we ask whether the AHC − control difference is larger in the areas where the AHC is more specialised (a moderator of the paired differences).

All differences are expressed in standard deviations of each indicator across Italian public hospitals, so that indicators measured on different scales can be combined; positive values mean that the AHC does better. With no structural adjustment, AHCs do better than the non-academic public hospitals of their region (0.144, 95% CI 0.056 to 0.232). Once each AHC is compared with a hospital of similar structure, the advantage disappears: the same estimator gives −0.082, and the precision-weighted model −0.033 (95% CI −0.121 to 0.054). The design would have detected a difference of 0.12 with 80% power (the minimum detectable difference), so it can exclude any but a small advantage for the AHCs. The difference is flat over ten years and through the COVID-19 pandemic. It is not more favourable to AHCs in the areas where they publish relatively more, and the design would have detected moderation of about a tenth of a standard deviation.

Two further findings qualify this result. The differences between an AHC and its control are real but specific to the service: an AHC may do persistently better than its control on one indicator and worse on another, but there is no general quality level that

lifts all its indicators together. And the comparison cannot be made for 9 large teaching hospitals, because their region contains no non-academic hospital of comparable size; for them the data cannot separate academic status from size.

The contribution is empirical. The paper links a decade of outcome indicators, structural data and an article-level bibliometric corpus, and it relies on established tools: optimal matching, multilevel models with known sampling variances, and moderator analysis. It does not propose a new method. It shows, however, how much the conclusion of a service-quality evaluation depends on design choices that are often left implicit. In our data the decisive choice is the comparator: holding structure fixed moves the estimate from a significant advantage for the AHCs to no difference. Modelling the precision of each rate, instead of averaging noisy rates, makes the estimate markedly more precise (minimum detectable difference 0.12 instead of 0.18). Comparing hospitals on a common set of indicators rather than on their own changes the estimate little here (-0.060 against -0.082), but it is what makes each comparison one of like with like. The article is organized as follows. Section 2 reviews the literatures on research hospitals, composite indicators and matched comparisons, and describes the National Outcomes Programme. Section 3 presents the data, Section 4 the comparison, the model and the measure of research, and Section 5 the results. Section 6 discusses them, including what the analysis cannot say about efficiency, and its limitations.

2 | BACKGROUND

2.1 | Research hospitals and the quality of care

Whether research activity in hospitals translates into better patient outcomes has been studied empirically since the early 2000s. Ayanian and Weissman[1] surveyed the US literature comparing teaching and non-teaching hospitals on structural and quality outcomes, finding a modest but persistent advantage for teaching institutions once case-mix was adjusted. Allison et al.[7] reported lower 30-day mortality for acute myocardial infarction at major teaching hospitals in the Medicare population. Khuri et al.[8] documented lower surgical complication rates in VA teaching hospitals across a broad set of procedures. Burke et al.[2] found lower risk-adjusted 30-day mortality at major teaching hospitals among Medicare beneficiaries across common medical and surgical conditions. Ozdemir et al.[3] found lower risk-adjusted mortality for acute admissions in the English hospital trusts more active in clinical research, also after adjustment for staffing and other structural factors. More recently, Burke et al.[9] found that community hospitals in the catchment area of academic medical centres exhibit lower risk-adjusted mortality than comparable hospitals located further away, suggesting a spillover mechanism in which research activity affects the wider care network rather than only the research institution itself. The literature converges on a qualitatively positive association, with effect sizes that are small and that depend heavily on case-mix adjustment.

The measurement of research activity itself has been reshaped by the expansion of open bibliometric infrastructure. `bibliometrix`[10] provides a common analytic layer over indexed corpora, and OpenAlex[11] offers a publicly accessible graph of scholarly works that supersedes proprietary products for many evaluation tasks. Citation-based impact indicators have been reviewed critically by Waltman[12]. Two governance documents frame the use of such indicators in evaluation. The Leiden Manifesto[13] sets ten principles for responsible metrics, including the requirement that quantitative indicators support rather than replace expert judgement. The Metric Tide review[14] reached compatible conclusions for the United Kingdom research-assessment system. Both warn against using aggregate counts as proxies for quality without an explicit mapping to the evaluation question.

The mapping used here follows the tradition of relative bibliometric indicators. The activity index compares the share of an institution's output in a field with the share of a reference set in the same field[15]; it is the bibliometric analogue of the revealed comparative advantage of trade economics[16]. Being a ratio of shares, it depends on how an institution's articles are distributed across fields and not on how many articles are attributed to it, a property that matters when affiliation practices differ across institutions (Section 3.3). In Italy, the research productivity and impact of AHCs differ with their organisational configuration: specialty-focused AHCs that are not mainly involved in teaching are more productive and have more research impact than fully integrated and affiliated university hospitals[17]. The PRIN 2022 SciK-Health project measures the research of the 49 AHCs on five dimensions (bibliometric impact, social engagement, industrial innovation, clinical research and competitive funding), combining publications, patents, clinical trials, grants, datasets and altmetric attention; Aria et al.[18] describe its open-data dashboard and document a marked heterogeneity across institutions.

2.2 | Composite indicators and comparisons of providers

The construction of composite indicators has been refined over two decades into a settled recipe. The OECD/JRC *Handbook on Constructing Composite Indicators*[19] codifies a six-step workflow: theoretical framework, variable selection, imputation of missing data, standardisation, weighting and aggregation, and robustness assessment. Each step embeds modelling choices whose consequences for the ranking of units are difficult to anticipate. Freudenberg[20] first catalogued the risks of mechanical composite construction for cross-country comparisons, and Saltelli[21] argued that an explicit audit of modelling assumptions is a precondition for policy legitimacy. The survey by Greco et al.[22] organises the subsequent literature around three open problems: which variables to combine, how to weight them, and how to aggregate them.

Weighting is the most debated step. Equal weights are transparent but silently impose a uniform prior over indicator importance. Expert elicitation is defensible in principle but opaque to external audiences and difficult to reproduce. Data-driven weights from principal components, data envelopment analysis or benefit-of-the-doubt models inherit the properties of the estimation sample and may shift when the unit set changes. Paruolo et al.[23] showed that nominal weights and revealed variable importance frequently diverge, using the Pearson correlation ratio to quantify how much variance in the composite is actually explained by each input. Becker et al.[24] extended this diagnostic into a Sobol-style variance-based sensitivity analysis. Luzzati and Gucciardi[25] proposed a Monte Carlo scheme that draws random weight vectors and reports the distribution of rankings rather than a single ordering. Non-compensatory alternatives, such as the generalised composite index of Mazziotta and Pareto[26], limit the substitution between indicators and make the treatment of low performers more demanding.

Saisana, Saltelli and Tarantola[27] formalised uncertainty and sensitivity analysis for composite indicators, propagating the uncertainty of the main modelling choices to the ranks of the units. In health care, Jacobs, Goddard and Smith[28] showed that hospital rankings based on composite performance measures are highly sensitive to such choices. A second line of work derives composite indicators from a statistical model of the indicators rather than from weights chosen outside the data. Cavicchia and Vichi[29] build hierarchical composite indicators on a latent dimensional structure and select it for the coherence of the policy conclusions drawn from the index; related work uses hierarchical disjoint non-negative factor analysis[30], hierarchical disjoint principal component analysis[31] and model-based ultrametric composite indicators[32]. In their assessment of a National Health Index for the United Kingdom, Freni-Sterrantino et al.[33] combine conceptual coherence, correlation analysis at several levels of the hierarchy, explicit comparison of weighting systems and full uncertainty and sensitivity analysis. Both lines raise a question that precedes aggregation: whether the indicators are reflective, that is, manifestations of a common latent quality of the unit, or formative, that is, components that jointly define the construct[34,35]. Section 4.3 shows that the correlations among the PNE indicators do not support a reflective model, which is one reason why this paper compares hospitals indicator by indicator and uses composites only as descriptive summaries.

A third line concerns the comparison of providers on noisy rates. Outcome rates of small hospitals or rare procedures have large sampling variance, so that league tables confound quality with chance[36,37]. Hierarchical models that shrink noisy rates were proposed for profiling providers[38], and surgical mortality was shown to be an unreliable measure of hospital quality for most procedures because of small samples[39]. Our model applies the same logic to paired differences: each difference carries its known sampling variance, which the model separates from the variance of real differences between hospitals.

2.3 | Matched comparisons of hospitals

The potential-outcomes framework formalises a causal contrast as the difference between two counterfactual states for the same unit, only one of which is observed[40]. In observational studies, identification rests on the unconfoundedness assumption that treatment assignment is independent of potential outcomes conditional on observed covariates. Rosenbaum and Rubin[41] established that, under unconfoundedness, the propensity score is a sufficient statistic for balancing observed covariates, reducing the multivariate matching problem to a univariate one. Imbens[42] reviewed the resulting family of estimators of the average treatment effect, including matching, weighting and subclassification on the propensity score. Abadie and Imbens[43] derived large-sample properties that justify inference under exact and nearest-neighbour matching.

Matching methods proliferated in the following two decades. Stuart[44] grouped them along three axes: matching structure (one-to-one, one-to-many, full), distance metric (Mahalanobis, propensity-score, or combinations), and assignment rule (greedy, optimal, with or without replacement). Ho et al.[45] reframed matching as nonparametric preprocessing that reduces the sensitivity of downstream parametric estimators to modelling choices. Iacus et al.[46] proposed coarsened exact matching, which bins continuous covariates and matches on the resulting strata, trading sample retention for exactness of balance. King and Nielsen[47]

warned that matching on estimated propensity scores can increase imbalance on individual covariates even while balancing the score, a phenomenon that motivates matching on the covariates themselves when their relative importance is known a priori.

We use matching as a design for a descriptive comparison, not as a device for causal identification. Unconfoundedness is implausible here: academic status is related to case mix, referral patterns and organisational features that the available data do not measure. Moreover, some structural covariates, such as the number of beds or the presence of high-specialty units, may themselves be consequences of academic status, and adjusting for a variable affected by the treatment changes the quantity being compared[48]. We therefore report a ladder of comparisons, from all hospitals of the region to designs that also hold fixed size, service profile and emergency role, and we read each estimate as a comparison between AHCs and non-academic hospitals that share the characteristics held fixed. Optimal matching[49], full matching, coarsened exact matching[46], entropy balancing[50] and genetic matching[51] provide alternative designs, and bias correction[52], Rosenbaum bounds[53] and E-values[54] probe the main one.

Two further considerations shape our choice of matching strategy. First, the comparison between research and non-research hospitals in Italy is particularly exposed to geographic confounding. Beraldo et al.[55] documented that patient mobility across Italian regions responds to budget regulation at the subnational level, and that flows concentrate on hospitals in wealthier catchment areas. Since health-care governance, financing and coding practices are regional, we match within region rather than within macro-area. Second, structural size is a highly right-skewed variable in Italian hospitals, ranging from small community facilities of fewer than one hundred beds to polyclinics above one thousand. The largest polyclinics have no non-academic hospital of comparable size in their region, a lack of common support that we report as a finding rather than conceal with distant matches.

2.4 | The Italian National Outcomes Programme as an evaluation instrument

The Italian National Health Service (Servizio Sanitario Nazionale, SSN) operates under a universal-coverage design in which most inpatient care is delivered by public or publicly accredited hospitals. Hospital discharge records (Scheda di Dimissione Ospedaliera, SDO) constitute a mandatory national dataset with patient-level demographics, diagnoses (ICD-9-CM), procedures, admission and discharge flags, and facility identifiers. The Italian outcomes-evaluation tradition, inaugurated by regional programmes such as P.Re.Val.E in Lazio[56], systematised the use of SDO linkage to produce risk-adjusted comparative indicators at facility level.

The Programma Nazionale Esiti (PNE), run by the Agenzia Nazionale per i Servizi Sanitari Regionali (AGENAS), under a Ministry of Health mandate, generalised this approach to the national level from 2010 onwards[57]. The 2025 edition assesses 1,117 public and accredited private hospitals using 218 indicators. These include outcome and process measures as well as indicators covering broader aspects of hospital and outpatient care. Outcome indicators generally capture adverse health outcomes following specific medical treatments or surgical procedures, including 30-day mortality, 30-day readmission, and post-operative complications for high-incidence conditions such as acute myocardial infarction, congestive heart failure, and stroke. Process indicators, instead, assess the timeliness and appropriateness of care delivery, for example by measuring the proportion of hip fracture patients treated within 48 hours. Risk adjustment is performed patient by patient through logistic regression models that control for age, sex, comorbidities and severity; Lenzi et al.[58] validated the adjustment for heart failure against clinical records. Fiore et al.[59] documented improvement trajectories among initially low-performing hospitals, suggesting that public disclosure of PNE results produces measurable quality gains.

PNE is therefore a mature evaluation instrument, methodologically comparable to the OECD Health Care Quality Indicators programme[60] and to the European strategies reviewed in Busse et al.[61]. For our purposes, it delivers the risk-adjusted substrate on which hospitals can be compared: the case-mix correction has already absorbed a substantial share of patient heterogeneity, and the hospital-year level of aggregation is the natural unit for combining with scientometric and structural data. What PNE does not absorb is selection at structure level; tertiary referral patterns, procedural specialisation and research intensity all operate above the patient-level adjustment, which is why we compare hospitals of similar structure and model research specialisation explicitly. PNE also publishes the standard error of every rate, which our model uses as its known sampling variance.

3 | DATA

3.1 | Outcome indicators

Table 1 shows how the indicators were selected. Of the 218 indicators of the 2025 edition of PNE, 51 belong to the clinical domains of the study. We then excluded indicators computed for geographical areas rather than hospitals, indicators without a hospital-level risk-adjusted rate, and near-duplicates. The remaining 41 indicators comprise 34 outcome indicators (mortality, readmission, complications and reoperation) and 7 process indicators (timeliness and appropriateness; codes 37, 202, 302, 514, 606, 621, 640), in 8 clinical domains and 24 clinical areas (Appendix A). PNE focuses on frequent conditions and procedures, but for rare outcomes a rate based on enough patients can still rest on few events (Section 5.1). We therefore apply no further volume threshold and weight each rate by its published standard error (Section 4.4). Moreover, each indicator is risk-adjusted at the patient level by AGENAS. The procedure applies a multivariate regression model to derive a population-specific severity measure, estimating the association between individual risk factors (such as age, sex, comorbidities, and clinical history) and observed outcomes. This severity measure is used to compute the predicted risk, defined as the risk that would be expected if the population treated by a given provider had the same distribution of risk factors as the overall reference population. Predicted and observed risks are compared through direct standardisation, supporting consistent hospital-level performance measurement.

For every hospital, year and indicator, PNE reports the adjusted rate, the number of eligible patients and the standard error of the logit of the rate. When a rate is 0 or 100%, its logit and the standard error are undefined, and PNE mostly reports a sentinel value (up to 43,593) instead of the standard error; for these 183 of the 85,404 hospital–year–indicator cells we compute the logit and its binomial standard error with a continuity correction of half an event. Rates are used as published, so for the outcome indicators and for the primary caesarean rate a higher rate means worse quality, whereas for the other process indicators it means better quality; the sign π_i of Section 4.1 orients every difference so that positive means the AHC does better. Four indicators (codes 620, 621, 630, 640) exist only from 2020; the other 37 cover all years. The temporal coverage of the extract used in this paper is 2015–2024, giving ten annual observations per hospital, an unusually large panel in the context of PNE data.

TABLE 1 Selection of the PNE indicators.

Step	Excluded (PNE codes)	Indicators
PNE indicators, 2025 edition		218
Selected in the clinical domains of the study		51
Area-level indicators (hospital not identifiable)	229, 65, 324, 325, 68	46
No hospital-level risk-adjusted rate	556, 998	44
Duplicates (2020–2024 versions, 120-day window)	652, 661, 605	41

3.2 | Hospitals and structural data

The study population consists of public hospitals. Of the 49 AHCs, 48 appear in PNE and 47 report at least one of the 41 indicators; the exception is a paediatric hospital. The control pool contains the non-academic public hospitals. Of the 505 public hospitals in PNE that are not AHCs, we excluded 22 with university, research-institute or research-body status (for example the sites of university hospital trusts), 20 reporting none of the indicators, 6 that could not be linked to the Ministry of Health bed data, 2 whose beds could not be attributed to the site, 1 belonging to the same trust as an AHC and 1 classified as private by the Ministry. The pool therefore contains 453 hospitals.

The data are annual, but the matching uses a single baseline: structural covariates are measured in 2015, the first year of the outcome window, so that pairs stay fixed over the decade. From the Ministry of Health open data on beds by facility and discipline we derived the number of beds and wards, the day-care share of beds, the number of high-specialty disciplines and the intensive-care share of beds; from PNE, the number of applicable indicators and the total volume in the first available year. The median AHC had 537 beds, the median control 177. The emergency-network level (no emergency department, emergency department, first- or second-level emergency department) comes from the Ministry's 2025 list of emergency facilities, the only one available, and is therefore a snapshot at the end of the period. A second-level emergency department is present in 27 of the 47 AHCs and in 55 of the 453 controls; 12 AHCs, all of them specialised research institutes (mostly cancer and neurological

institutes), have no emergency department. PNE identifies hospitals by their current code in every year, so hospitals recoded after 2015, for example after the regional reforms of 2017, take their covariates from the first year in which they can be linked.

3.3 | Publications

The PRIN 2022 SciK-Health project [18] assembles an integrated dataset of research output, innovation and engagement indicators for the 49 Italian AHCs, classified into nine fully integrated university hospitals (AOU-integrated), seventeen affiliated university hospitals and twenty-three IRCCS research institutes. Institutional disambiguation relies on the Research Organization Registry (ROR), which resolves the many variant names that individual hospitals use across publications and time. The bibliographic backbone of the dataset is OpenAlex [11], an open, comprehensive index of scholarly works that has become a standard research-evaluation source. From it we extracted the 214,848 journal articles with a PubMed identifier published by the 49 AHCs between 2005 and 2024. For each article we use the title, the Medical Subject Headings (MeSH) [62] with their descriptor codes and major-topic flags, and OpenAlex's normalised citation percentile and field-weighted citation impact.

Two features of the corpus shape the measure of Section 4.6. First, the number of articles attributed to an AHC depends on how authors declare their affiliation, to the hospital or to the university: in 2010–14 it ranges from 12 to 3,939 across the AHCs of the main comparison, so absolute amounts of research cannot be compared across AHCs. Second, MeSH indexing thins out over time: the share of articles with a major-topic descriptor falls from 73% in 2010 to 42% in 2020 and 0.2% in 2024, so a measure based on MeSH alone would drift. The SciK-Health CSV distributions and documentation are available at https://github.com/massimoaria/scikhealth and https://massimoaria.github.io/scikhealth/.

4 | METHODS

4.1 | Comparison and estimand

For a matched pair p of an AHC and a control, a year t and an indicator i reported by both hospitals, let y^A_{pti} and y^C_{pti} be the logits of the risk-adjusted rates of the AHC and of the control, and let s_{ti} be the standard deviation of the logit across the public hospitals, AHCs included, that report indicator i in year t. The paired difference

$$d_{pti} = \pi_i \, \frac{y^A_{pti} - y^C_{pti}}{s_{ti}} \tag{1}$$

is positive when the AHC does better; $\pi_i = \pm 1$ orients the indicator. Its sampling variance is $v_{pti} = (\sigma^2_{A,pti} + \sigma^2_{C,pti})/s^2_{ti}$, where $\sigma_{A,pti}$ and $\sigma_{C,pti}$ are the standard errors of the two logits published by PNE. The two hospitals of a pair are compared only on the indicators that both report in the same year, which removes differences in the mix of services, and only pair–years with at least five common indicators enter the main comparison.

The target is a weighted average of the indicator-specific differences in which each pair counts equally, each year counts equally within a pair and each common indicator counts equally within a pair–year. If A_{pt} is the set of common indicators of pair p in year t, $K_{pt} = |A_{pt}|$, T_p the number of years in which pair p enters the comparison and G the number of pairs, indicator i receives the weight

$$\omega_i = \frac{1}{G} \sum_p \frac{1}{T_p} \sum_t \frac{\mathbf{1}\{i \in A_{pt}\}}{K_{pt}}, \qquad \sum_i \omega_i = 1, \tag{2}$$

where the inner sum runs over the years of pair p. These are the weights implicit in the equal-weight composite of the common indicators (Section 4.3), so the composite and the model of Section 4.4 target the same weighted average. The estimand is descriptive: the average difference in PNE-measured quality between AHCs and non-academic public hospitals of the same region with similar structure.

4.2 | Matching

Each AHC is matched to one control, without replacement, by optimal pair matching [49] within region on the Mahalanobis distance of the covariates of specification M2: log beds, log wards, day-care share of beds, number of applicable PNE indicators and log

PNE volume. A caliper of 0.5 standard deviations of log beds (0.394 on the log scale, a maximum ratio of 1.48 between the beds of the two hospitals) excludes implausible pairs; the algorithm first maximises the number of pairs within the caliper and then minimises their total distance. AHCs with no control within the caliper in their region are excluded from the main comparison and described separately. Balance is assessed with standardised mean differences, variance ratios and Kolmogorov–Smirnov statistics on the covariates of M2 and on three covariates not used in the distance: high-specialty disciplines, intensive-care share and emergency level.

To show how the comparison depends on what is held fixed, the same quantity is estimated under 16 designs (Appendix C). They form a ladder of covariate specifications, from all controls in the region with no structural adjustment, through region and beds (M1) and M2, to M3, which adds high specialty, intensive care and emergency level, also with the emergency level matched exactly. They also include alternative methods (nearest neighbour with replacement, optimal full matching, coarsened exact matching, entropy balancing and genetic matching with a quantile–quantile balance criterion), alternative support rules (caliper of 0.25 standard deviations, no caliper, fallback to the macro-area) and alternative control pools. To compare designs with the same estimator, every design is summarised by the equal-weight composite of the common indicators, averaged over the controls of each AHC with the matching weights, with standard errors clustered on AHCs and controls. For the main design we also report the bias-corrected estimator of Abadie and Imbens[52], with the outcome–covariate slopes estimated on the control pool, Rosenbaum bounds for the paired Wilcoxon test[53] and E-values[54].

4.3 | Composite indicator and weights

As a descriptive summary we compute, for every pair–year, the equal-weight mean of the standardised differences on the common indicators. The submitted version of this paper drew random weights from a symmetric Dirichlet distribution instead. For a hospital–year with K standardised indicators $z_1, \dots, z_K$ and weights $w \sim \text{Dirichlet}(\alpha, \dots, \alpha)$, the composite $C = \sum_i w_i z_i$ satisfies

$$\mathrm{E}(C \mid z) = \bar{z}, \qquad \mathrm{Var}(C \mid z) = \frac{1}{K(K\alpha + 1)} \sum_{i=1}^{K} (z_i - \bar{z})^2 \tag{3}$$

(Appendix B). The mean of the Dirichlet composite is the equal-weight composite by construction, and its dispersion depends only on K, α and the spread of the hospital's own indicators, so the quantiles of C across draws describe the sensitivity of a score to the weights, not the uncertainty of an estimate; we call them weight-sensitivity intervals. To study weights in a way that matters for comparisons, we draw one weight vector at a time and apply it to all hospitals, for symmetric Dirichlet distributions with $\alpha = 0.5$, 1 and 5 and for a hierarchical scheme that weights clinical domains first and indicators within domains second; for each draw we recompute the paired comparison and the ranking of the public hospitals.

We treat the indicators as formative. Across the hospital–years of the public hospitals, AHCs included, the median Spearman correlation between two polarity-aligned indicators is 0.041 (0.072 within the same clinical domain), and only 1.8% of the 788 pairs of indicators observed together in at least 100 hospital–years exceed 0.3. The first principal component of the 9 indicators observed in at least half of the hospital–years explains 20% of their variance, against 11% if they were uncorrelated. The data do not support a common latent quality of the hospital, which reflective and latent-variable composites presuppose.

4.4 | Precision-weighted multilevel model

The main estimate comes from a multilevel model for the paired differences,

$$d_{pti} = \beta_i + u_p + \eta_{pi} + r_{pti} + a_{pt} + e_{pti} + \varepsilon_{pti}, \qquad \varepsilon_{pti} \sim N(0, v_{pti}), \tag{4}$$

where β_i is the mean difference on indicator i, a fixed effect; $u_p \sim N(0, \sigma_u^2)$ is a pair effect common to all indicators and years; $\eta_{pi} \sim N(0, \sigma_\eta^2)$ is a pair × indicator effect constant over time; r_{pti} is a pair × indicator effect that decays over time as a continuous-time first-order autoregressive process[63], with $\mathrm{Cov}(r_{pti}, r_{pt'i}) = \sigma_r^2 \, \phi^{|t-t'|}$ for $0 < \phi < 1$; $a_{pt} \sim N(0, \sigma_a^2)$ is a pair × year effect common to all indicators; $e_{pti} \sim N(0, \sigma_e^2)$ is a residual; and ε_{pti} is the sampling error, with the known variance v_{pti}. All terms are mutually independent. The known variances make noisy cells count less without volume thresholds: a rate computed on few patients has a large v_{pti} and receives a small weight. The indicator effects are parametrised by an intercept and sum-to-zero

contrasts, so that the intercept is their unweighted mean $\bar{\beta} = I^{-1} \sum_i \beta_i$ over the I indicators. The estimand of Section 4.1 is the linear combination $\theta = \sum_i \omega_i \beta_i$, with the weights of equation (2); we also report $\bar{\beta}$, in which every indicator counts equally.

All random effects are nested within pairs, so the covariance matrix of the differences is block-diagonal by pair. The variance components are estimated by restricted maximum likelihood, block by block, and the fixed effects by generalised least squares; the implementation reproduces the estimates of the `metafor` package[64] and of `lme4` on the specifications they can fit. Inference uses a t distribution with $G - 1$ degrees of freedom. It is checked against cluster-robust standard errors by pair (CR1) and against a restricted wild cluster bootstrap with Webb weights and 9,999 replications[65,66,67]. The calibration of the three tests is checked by simulating 400 data sets from the fitted model under the null hypothesis. We compare the temporal structure of equation (4) with a constant pair × indicator effect only and with a decaying effect only, and we report the minimum difference detectable with 80% power in a two-sided 5% test.

4.5 | Change over time

The year profile comes from the model with year-specific effects; for each year the difference is summarised with the weights of equation (2) restricted to that year's cells. Homogeneity across years is tested with an F test on nine and $G - 1$ degrees of freedom. We also estimate two period contrasts, 2020–21 and 2022–24 against 2015–19, and a linear trend, and we test whether pairs follow different trajectories by adding random pair slopes, with a likelihood-ratio test referred to a 50:50 mixture of χ^2_1 and χ^2_2. The family of tests was fixed before estimation: the primary test is homogeneity across years; the seven secondary tests are the trend, the two contrasts and the slope variance on all indicators and the trend and the two contrasts on outcome indicators, with Holm's correction[68]. Because four indicators start in 2020, the COVID-19 contrast is repeated on the 37 indicators present in all years.

4.6 | Research specialisation as a moderator

Measure. An article is pertinent to a clinical area when one of its major-topic MeSH descriptors belongs to the area, identified by descriptor code, or when its title contains one of the area's terms, matched as whole words and with British spelling variants. The 24 areas group the 41 indicators by condition or procedure; for example, the four cholecystectomy indicators form one area. The lists of descriptors and title terms (522 entries: 190 descriptors and 332 terms) were generated from the corpus, revised to exclude generic and ambiguous terms, and frozen before any outcome was examined. Both channels are needed, because the title rule is precise but finds only part of the articles that indexers assign to an area, and because MeSH indexing declines over time. Against the major-topic descriptors, which indexers of the National Library of Medicine assign independently of the title rule, the title rule has precision 0.95 and recall 0.73 in 2010–19, and precision between 0.94 and 0.95 in every five-year period from 2005–09 to 2020–24 (Appendix F).

Moderator. For AHC h and area g, let n_{hg} be the number of the AHC's pertinent articles, N_h the number of all its articles, and n_g and N the same counts summed over the 49 AHCs. The activity index $AI_{hg} = (n_{hg}/N_h)/(n_g/N) = n_{hg}/E_{hg}$, with $E_{hg} = N_h n_g/N$, compares the share of the AHC's articles in the area with the share of all AHCs[15]. It is computed in the baseline period 2010–14, before the outcome window. With few articles the index is unstable, so it is shrunk with an empirical Bayes Poisson–gamma model for each area[69]: $n_{hg} \sim \text{Poisson}(E_{hg}\theta_{hg})$ and $\theta_{hg} \sim \text{Gamma}(\psi_g, \rho_g)$, with ψ_g and ρ_g estimated by moments, so that the shrunken index is the posterior mean $\tilde{\theta}_{hg} = (n_{hg} + \psi_g)/(E_{hg} + \rho_g)$. The moderator S_{hg} is $\log \tilde{\theta}_{hg}$ centred on its mean over the 24 areas of the AHC, so that it measures the AHC's relative specialisation across areas and not its total output; it is divided by its standard deviation over the model cells. Shrinking the index acts as a regression calibration of the true specialisation[70]; the residual measurement error, shared by the indicators of an area, is absorbed by a pair × area random effect.

Model. The model of equation (4) is extended with the moderator of each cell, its mean over the cells of the pair and a pair × area random effect b_{pg}:

$$d_{pti} = \beta_i + \gamma_S S_{p\,g(i)} + \gamma_M \bar{S}_p + b_{p\,g(i)} + u_p + \eta_{pi} + r_{pti} + a_{pt} + e_{pti} + \varepsilon_{pti}, \tag{5}$$

where $g(i)$ is the area of indicator i, $S_{pg} = S_{hg}$ for the AHC h of pair p, and $b_{pg} \sim N(0, \sigma_b^2)$. With the pair mean $\bar{S}_p$[71] and the indicator effects, γ_S is a within-pair coefficient: it asks whether, within a pair, the AHC − control difference is larger in the areas where the AHC is more specialised. We report it per standard deviation of specialisation and as the predicted difference at percentiles of S.

Lags, sides and quality. To allow for lags, a second model replaces the baseline specialisation with the specialisation in the three years before each outcome year, $t-3$ to $t-1$, split into its mean over the years of the AHC–area unit and the deviation from that mean, with coefficient γ_D; the same model with the three following years, $t+1$ to $t+3$, serves as a placebo. If the AHC's research improved its own outcomes, the moderation would appear in the AHC's rates and not in the control's, so the model is also fitted separately to the standardised logits of the AHC and of the control. Research quality enters as the mean normalised citation percentile of the AHC's pertinent articles in the area, measured as a deviation from the mean of all the AHC's articles, shrunk towards zero and centred within the AHC (coefficient γ_Q), with log(1 + FWCI) as an alternative. The relative volume of the cell, $\log(n^A/n^C)$, enters as a covariate in a separate model, since volume may confound or mediate the association[72].

Pre-specification. The measure, the models, the sensitivity analyses and the family of tests were fixed before the outcomes were examined. The primary test is γ_S in equation (5), with the p value of the wild bootstrap; the secondary tests are γ_S on outcome indicators, γ_S net of relative volume, γ_Q and γ_D, with Holm's correction. The sensitivity analyses vary the measure, the period, the sample, the areas and the weights (Appendix F); one of them, which excludes the two areas where the title rule and MeSH agree least, was added after the main estimates.

5 | RESULTS

5.1 | Matched sample and reliability of the rates

Of the 47 AHCs, 9 have no non-academic public hospital within the caliper in their region (13 with a caliper of 0.25 standard deviations). They are the largest teaching hospitals of their regions, in Turin, Padua, Genoa, Parma, Bologna, Pisa, Florence, Ancona and Bari, whose regions contain no non-academic hospital of comparable size (Appendix Table A4). Against all the non-academic public hospitals of their region, 8 of them score higher, but that comparison is not like with like, and we describe these hospitals separately.

Optimal matching forms 38 pairs. In 22 of them the two hospitals share at least five indicators in at least one year. The other 16 involve specialised research institutes, with at most 7 applicable indicators each, which share few indicators with a general hospital. The main comparison therefore rests on 22 pairs, 218 pair–years and 4,323 pair–year–indicator cells. Matching removes most of the imbalance in size and service profile (Table 2; the standardised mean difference of log beds falls from 1.13 to 0.13); the day-care share of beds, which was nearly balanced, is slightly less so after matching (from 0.16 to 0.26). Of the covariates not in the distance, high-specialty disciplines remain imbalanced (0.48), because within their region the AHCs remain the most specialised hospitals. The emergency level is lower in the AHCs than in their controls (−0.31) only because of the specialised institutes, most of which have no emergency department and are matched with general hospitals; in the 22 pairs of the main comparison the standardised difference is 0.07. The designs that include these covariates are reported in Section 5.2.

TABLE 2 Covariate balance in the main design: standardised mean differences before (all AHCs and pool controls) and after matching (the 38 matched pairs), variance ratio and Kolmogorov–Smirnov statistic after matching. †Not in the matching distance.

Covariate (2015)	SMD before	SMD after	Variance ratio after	KS after
Beds (log)	1.13	0.13	1.26	0.26
Wards (log)	1.03	0.19	1.38	0.26
Day-care share of beds	0.16	0.26	3.09	0.21
Applicable PNE indicators	0.59	−0.06	1.61	0.18
PNE volume (log)	0.71	−0.10	1.41	0.16
High-specialty disciplines†	1.19	0.48	1.33	0.29
Intensive-care share of beds†	0.20	−0.02	0.87	0.18
Emergency-network level†	0.27	−0.31	2.58	0.29

The PNE rates are noisy. For each indicator and year we estimated by restricted maximum likelihood the variance of true differences between hospitals, τ^2, and computed the reliability $\lambda = \tau^2/(\tau^2 + \sigma^2)$ of each rate, where σ is the standard error of its logit. Reliability depends on the expected number of events (Spearman correlation 0.83) far more than on volume (0.40). The median reliability is below 0.3 for 22 of the 41 indicators (21 of the 34 outcome indicators) and above 0.7 for only 5, all process

indicators. Reliability is below 0.5 for 59% of the rates and for 59% of those based on at least 20 patients, so a volume threshold would not remove the noise (Appendix Table A1).

5.2 | AHCs and matched controls

Table 3 reports the main estimates. In the precision-weighted model the AHC − control difference is −0.033 standard deviations (95% CI −0.121 to 0.054; $p = 0.44$; wild bootstrap $p = 0.43$). The minimum detectable difference is 0.12, so the design would have detected differences of about a tenth of a standard deviation of the indicators. The difference is 0.012 on outcome indicators (−0.084 to 0.107) and −0.151 on process indicators (−0.376 to 0.074; $p = 0.18$), and none of the 41 indicator-specific differences is significant after Benjamini–Hochberg correction (Appendix Table A1). The estimate does not depend on how cells are weighted: it is −0.016 when every indicator counts equally, −0.045 without the known sampling variances and −0.031 on the log odds ratio scale (an odds ratio of 0.97); including the pair–years with fewer than five common indicators, which adds 8 pairs with a specialised institute (30 pairs in all), gives 0.003. Removing one pair at a time, the estimate ranges from −0.051 to −0.017. The model tests are well calibrated: under the null hypothesis the 5% tests reject in 3.5% (model), 4.5% (CR1) and 5.0% (wild bootstrap) of the simulations.

TABLE 3 AHC − control difference in the main design, in standard deviations of the indicators (positive: the AHC does better). Cells are pair–year–indicator combinations. *p*: *t* test of the model with $G - 1$ degrees of freedom; wild: restricted wild cluster bootstrap with Webb weights. ‡Unstandardised log odds ratio.

Specification	Pairs	Cells	Estimate	95% CI	*p*	*p* (wild)
Descriptive						
Composite: equal weights, standardised rates	22	4,323	−0.082	[−0.210, 0.047]	0.20	
Precision-weighted multilevel model						
Model: composite weights (primary)	22	4,323	−0.033	[−0.121, 0.054]	0.44	0.43
Model: every indicator weighted equally	22	4,323	−0.016	[−0.195, 0.163]	0.85	0.71
Model without the known sampling variances	22	4,323	−0.045	[−0.132, 0.043]	0.30	
Model on the log odds ratio scale‡	22	4,323	−0.031	[−0.096, 0.033]	0.32	
Subsets and extension						
Outcome indicators only	22	3,471	0.012	[−0.084, 0.107]	0.80	0.81
Process indicators only	22	852	−0.151	[−0.376, 0.074]	0.18	0.18
All common cells (extension)	30	4,475	0.003	[−0.088, 0.095]	0.95	

The descriptive composite on the same cells gives −0.082 (95% CI −0.210 to 0.047), with a larger minimum detectable difference (0.18); averaging each hospital over its own indicators instead, on the same pair–years, gives −0.060 ($p = 0.17$). The composite and the model use the same weights, those of equation (2), and differ in three steps: the logit scale (+0.020), modelling the indicator-specific differences instead of averaging raw ones (+0.017) and weighting them by precision (+0.011). One cell shows why precision matters. In 2015, one AHC's rate of readmission after vaginal delivery was computed on 2 deliveries, and its risk-adjusted rate of 46% produced a standardised difference of −7.99. In the composite that cell weighs 5.8 times the median cell; in the model 0.88 times, and removing it moves the model estimate only from −0.033 to −0.032.

Figure 1 shows how the comparison depends on the design. With no structural adjustment, the AHCs do better than the non-academic public hospitals of their region: 0.144 (95% CI 0.056 to 0.232; $p = 0.002$) on all indicators and 0.152 ($p = 0.009$) on outcome indicators. Matching on beds within region reduces the difference to 0.011, and the main design gives −0.082 with the same composite estimator. Across the remaining designs the estimate on all indicators ranges from −0.083 to 0.064, and none is significant at the 5% level. On outcome indicators the estimates range from −0.028 to 0.097 and are positive in 12 of 15 designs; they are significant only with optimal full matching (0.097, $p = 0.036$) and without a caliper (0.093, $p = 0.012$), two designs that admit less similar controls and balance the covariates worse than the main design (Appendix Table A3). With the multilevel model, the design that adds high specialty, intensive care and emergency level gives −0.022, and matching the emergency level exactly gives −0.020; only the design without a caliper is significant (0.079, $p = 0.035$). The bias-corrected estimator gives −0.071 (95% CI −0.172 to 0.029) with the M2 covariates and −0.101 (−0.217 to 0.016) with the M3 covariates.

Since no difference is significant, Rosenbaum bounds have no finding to probe (one-sided $p = 0.15$ already at $\Gamma = 1$); the E-value of the composite estimate of the main design is 1.37.

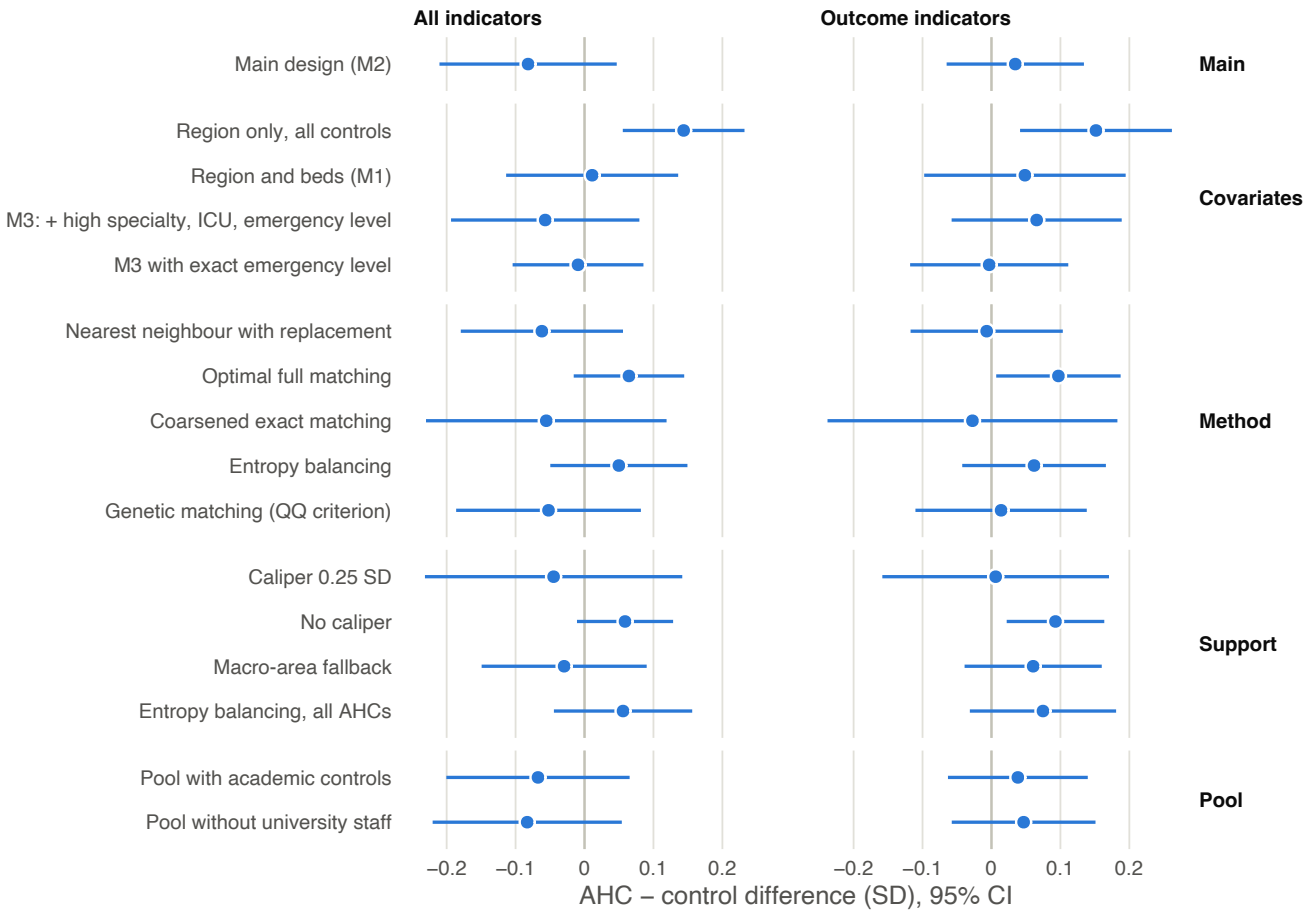


FIGURE 1 AHC − control difference, with 95% confidence interval, under the 16 designs of Appendix Table A3, with the equal-weight composite of the common indicators as common estimator. Positive values: the AHC does better.

5.3 | Where the variation lies

Table 4 decomposes the variance of the paired differences. Sampling noise accounts for 60% of it. Most of the remainder is specific to the service: a pair × indicator component that is constant over time (18%, standard deviation 0.57) and one that decays, with a correlation of 0.66 between consecutive years and a half-life of 1.7 years (21%, standard deviation 0.62). The component common to all the indicators of a pair is small (0.6%, standard deviation 0.11), and the pair × year component is estimated at zero. An AHC can do better than its control on one service and worse on another, and these differences persist for a few years; but there is no general quality level of the AHC that lifts all its indicators, and no shock in a given year that affects all of them. The structure with a constant and a decaying component fits better than a constant component only (by 243 points of the Akaike information criterion) and than a decaying component only (by 9.4). On outcome indicators sampling noise accounts for 76% of the variance; on process indicators for 17%, and the service-specific components reach 80%.

TABLE 4 Variance components of the paired differences in the main model: standard deviation and share of the total variance (the sampling share uses the mean known variance). The AR(1) component decays over time; the others are constant or independent across years.

Component	Nature	SD	Share of variance		
		(all)	All	Outcomes	Processes
Pair	persistent	0.11	0.6%	1.2%	2.6%
Pair × indicator, constant	persistent	0.57	18.0%	12.6%	33.8%
Pair × indicator, AR(1)	decaying	0.62	21.2%	10.3%	46.3%
Pair × year	transient	0.00	0.0%	0.1%	0.0%
Cell	transient	0.08	0.4%	0.0%	0.0%
Sampling (mean known variance)	sampling	1.04	59.8%	75.8%	17.4%

5.4 | Change over time

Figure 2 shows the year profile. The annual estimates range from −0.088 in 2023 to 0.002 in 2020, and the data are compatible with a constant difference ($F_{9,21}$ = 0.58, p = 0.80; Table 5). Relative to 2015–19, the difference changes by 0.014 in 2020–21 (95% CI −0.076 to 0.104; wild bootstrap p = 0.76) and by −0.015 in 2022–24; on the fixed set of 37 indicators the COVID-19 contrast is 0.010, and on outcome indicators 0.051. The linear trend is −0.0004 standard deviations per year (95% CI −0.014 to 0.013). The design would have detected period contrasts of 0.13 and a cumulated trend of 0.17 over nine years. The pairs follow nearly parallel trajectories: the standard deviation of the pair-specific slopes is 0.005 per year and is not distinguishable from zero (likelihood ratio 0.23, p = 0.76), so there is no evidence that pairs diverge or converge. The smallest Holm-adjusted p value among the secondary tests is 1.00. The conclusions hold on outcome indicators, on the pairs observed in all years, without 2020–21, in the designs that add the emergency level, on all common cells and without weights (Appendix Table A5).

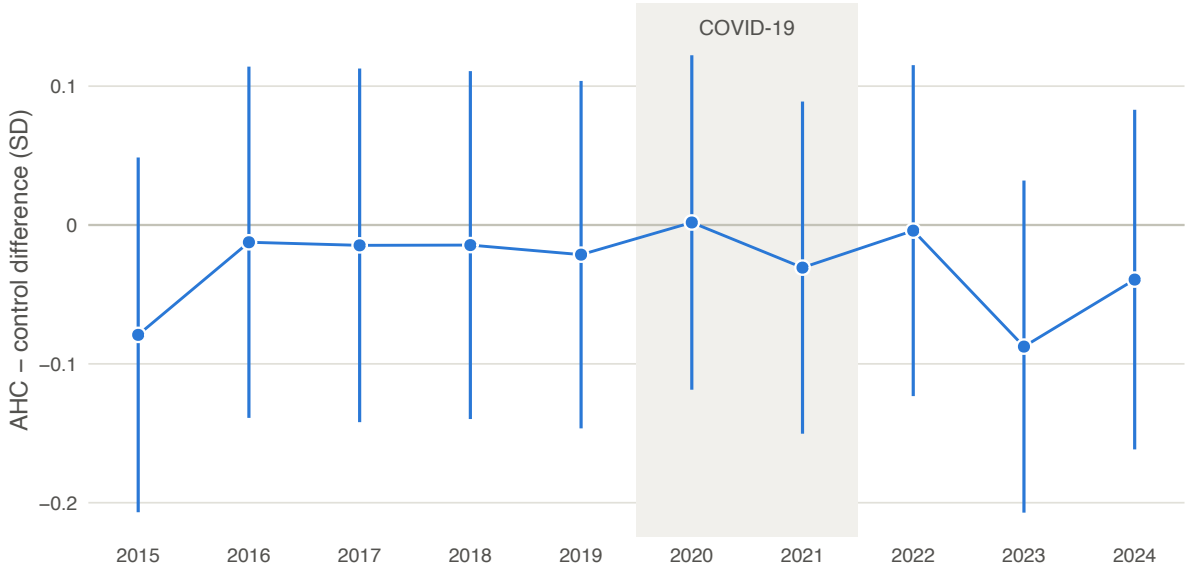


FIGURE 2 AHC − control difference by year in the main model, with 95% confidence intervals. The shaded band marks 2020–21.

5.5 | Weights and rankings

The comparison between groups is robust to the weights; the ranking of individual hospitals is not (Appendix Table A2). When one Dirichlet(1) weight vector is drawn and applied to all hospitals, the composite difference in the main design lies within

TABLE 5 Change over time in the main model. p (Holm): adjusted within the family fixed before estimation (primary test: homogeneity across years). MDD: minimum detectable difference with 80% power.

Test	Estimate	95% CI	p	p (wild)	p (Holm)	MDD
Homogeneity across years (F)	0.58		0.80		0.80	
2020–21 vs 2015–19	0.014	[−0.076, 0.104]	0.75	0.76	1.00	0.13
2022–24 vs 2015–19	−0.015	[−0.106, 0.076]	0.73	0.73	1.00	0.13
Linear trend (per year)	−0.0004	[−0.014, 0.013]	0.95	0.95	1.00	0.019
Variance of pair slopes (LR)	0.23		0.76		1.00	

−0.160 to −0.003 in 95% of the draws, and 2.8% of the draws produce a significant difference. Across the same draws, the median Spearman correlation between the ranking of the public hospital–years and the equal-weight ranking is 0.82, and 48% of the hospital–years change quintile; with more dispersed weights ($\alpha = 0.5$) the share rises to 55%. Since this paper estimates a group difference, and since reliable rankings would require their own model for the noise of each rate, we do not report hospital rankings.

5.6 | Research specialisation

Of the 214,848 articles, 45,515 are pertinent to at least one clinical area. The shares are small (Table 6): in a year of 2015–24, on average 1.44% of an AHC's articles concern the clinical area of the cerebrovascular indicators and 0.30% that of the respiratory ones, and the average for an oncological indicator is 1.82%. The title rule agrees with MeSH in every area except two, lower-limb arterial disease (precision 0.76) and childbirth (0.73), where the rule, the indexing or both are less reliable; elsewhere precision is at least 0.91. Of the 528 AHC–area units of the main comparison, 153 (29%) have no pertinent article in 2010–14 and are kept by the shrinkage. The moderator hardly depends on the channel (Spearman correlation 0.89 with a MeSH-only and 0.94 with a title-only version) and is moderately stable over time (correlation 0.49 between 2010–14 and 2015–19).

TABLE 6 Research pertinent to the indicators, by clinical domain, 2015–24 (49 AHCs). For each AHC, year and indicator of the domain: share of the AHC's articles of the year that are pertinent to the indicator's clinical area (mean) and presence of at least one such article (share of combinations); last column: articles pertinent to at least one area of the domain.

Clinical domain	Indicators	Pertinent share (%)	With any (%)	Pertinent articles
Oncology	11	1.82	68.5	23,216
Cerebrovascular	2	1.44	67.8	1,822
Urogenital	1	0.84	62.0	1,184
Cardiovascular	13	0.71	41.9	5,930
Musculoskeletal	5	0.33	31.8	874
Respiratory	2	0.30	39.6	410
Perinatal	3	0.22	35.5	360
Digestive	4	0.10	21.6	157

Within pairs, the AHC − control difference is not larger in the areas where the AHC is more specialised: $\gamma_S = -0.029$ per standard deviation of specialisation (95% CI −0.116 to 0.059; wild bootstrap $p = 0.27$; Table 7). The predicted difference is 0.004 at the 10th percentile of specialisation and −0.061 at the 90th, a change of −0.065 (95% CI −0.264 to 0.133; Figure 3). The minimum detectable γ_S is 0.12 per standard deviation, about 0.28 between the 10th and the 90th percentile. The moderator explains none of the service-specific variance: the pair × area and pair × indicator components sum to 0.714 without and 0.715 with it. No secondary test is significant: γ_S is 0.013 on outcome indicators and −0.048 net of relative volume (Holm 0.31), γ_Q is −0.037 (Holm 0.31) and γ_D is 0.011 (Holm 1.00). In 2015–21, the future-window placebo gives 0.006, against −0.012 for the past window.

If the AHC's research improved its own outcomes, the association would appear on the AHC's side. The two sides move in the same direction instead: −0.043 for the AHC's rates (model $p = 0.20$; the wild bootstrap gives $p = 0.042$, outside the test family and in the direction opposite to a benefit of research) and −0.020 for the control's rates, so the moderation coefficient is close to their difference and no association specific to the AHC's research emerges. All 14 sensitivity analyses give confidence intervals

TABLE 7 Moderation by research specialisation: the family of tests fixed before estimation. Coefficients per standard deviation of the moderator. The primary test uses the wild bootstrap p value; the secondary tests are adjusted with Holm's method.

Role	Test	Estimate	95% CI	p (model)	p (wild)	p (Holm)
Primary	γ_S: specialisation 2010–14, all indicators	−0.029	[−0.116, 0.059]	0.50	0.27	
Secondary	γ_S: outcome indicators only	0.013	[−0.069, 0.095]	0.75	0.65	1.00
Secondary	γ_S: net of relative volume	−0.048	[−0.135, 0.039]	0.26	0.078	0.31
Secondary	γ_Q: research quality (citation percentile)	−0.037	[−0.125, 0.052]	0.40	0.092	0.31
Secondary	γ_D: change in the $t-3..t-1$ window	0.011	[−0.048, 0.069]	0.71	0.66	1.00

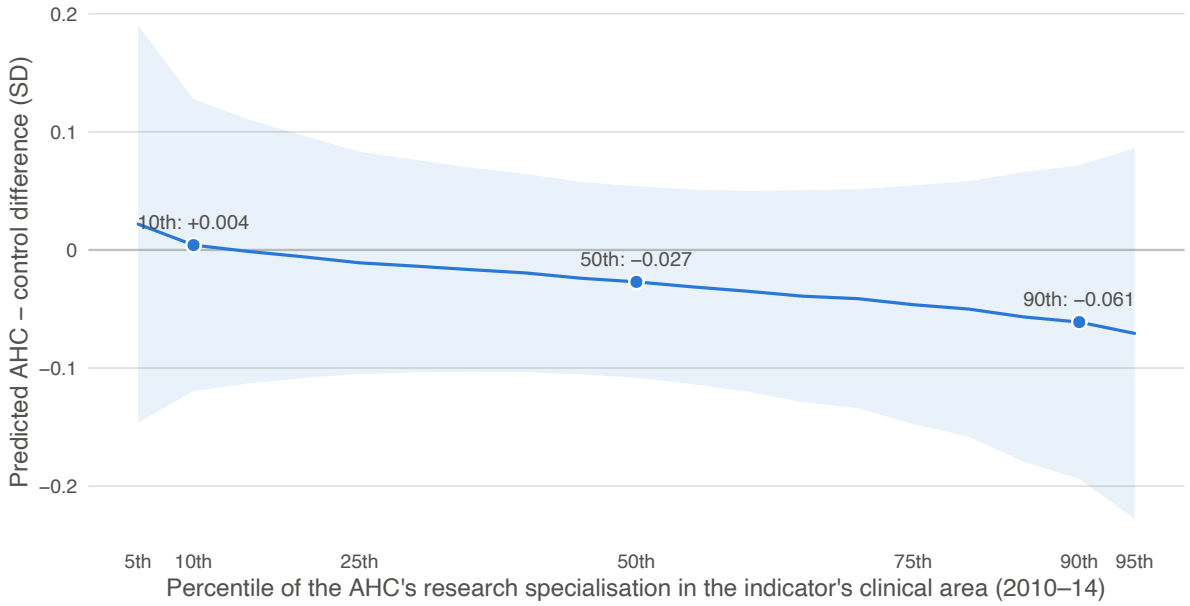


FIGURE 3 Predicted AHC − control difference by percentile of the AHC's research specialisation in the clinical area of the indicator, with 95% confidence band, for a pair with average specialisation mix and the weights of equation (2).

that include zero, with estimates between −0.047 and 0.080 (tertiles aside: top against bottom tertile −0.203, or −0.098 per standard deviation, $p = 0.057$; Appendix Table A7). The estimate is 0.001 with the MeSH channel only, −0.039 with the title channel only and −0.017 without the two areas where the channels agree least.

Independently of research, the AHC does better than its control on the indicators where it treats relatively more patients: 0.213 per unit of log volume ratio (95% CI 0.125 to 0.301; $p < 0.001$). This may reflect the volume–outcome relationship[72], but also the noise of rates computed on few patients, and we report it as a description.

6 | DISCUSSION

6.1 | Main findings

Compared with non-academic public hospitals of the same region and similar size and service profile, the AHCs show no difference in the quality measured by the National Outcomes Programme, and the same holds when the emergency role is also held fixed. The estimate is close to zero and precise enough to exclude an advantage for the AHCs larger than 0.054 standard deviations of the indicators, the upper limit of its 95% confidence interval, and it is stable from 2015 to 2024 and through the COVID-19 pandemic. The finding concerns the 22 of the 47 AHCs that have a structurally comparable non-academic hospital in their region and share enough indicators with it; the other AHCs are discussed below.

Three features qualify this conclusion. First, the crude advantage of the AHCs is real, but it belongs to their structure rather than to their academic status. The AHCs do better than the non-academic public hospitals of their region taken together (0.144), and the difference disappears once they are compared with hospitals of similar size and service mix. This matters for public reporting: a comparison of all hospitals of a region attributes to academic status differences that belong to structure. Second, the differences between an AHC and its control are specific to the service. They are persistent enough to be real, with a constant part and a part with a half-life of 1.7 years, but they do not add up to a general quality level of the AHC. Composite scores that average over services hide this structure, and rankings based on them are unstable to the choice of weights. Third, the conclusion does not cover two groups of AHCs. The 9 largest teaching hospitals have no comparable non-academic hospital in their region: 8 of them score higher than the non-academic hospitals of their region, but for them the data cannot separate academic status from size and role in the network. The specialised research institutes, in turn, share too few indicators with a general hospital to be compared with one (16 of the 38 matched pairs).

These results are consistent with the studies that find small advantages for teaching or research-active hospitals [1,2,3], once the comparison is read for what it holds fixed. Those studies compare whole institutions and, when they adjust for structure, do so in a regression across all hospitals; our crude comparison, which is closest to them, reproduces a small advantage for the AHCs. The advantage disappears when each AHC is compared, within its region and on the indicators both hospitals report, with a hospital of similar structure. Our data cannot say whether the same would happen in other health systems, where the regional organisation of care and the role of academic hospitals differ.

The contrast between the crude and the matched comparison also shows which design choices matter, as anticipated in the Introduction. The comparator matters most: matching on beds within region already reduces the difference from 0.144 to 0.011, and the main design, which also matches on service profile, gives -0.082 with the same estimator. Once the hospitals are matched, comparing them on their own indicators rather than on the common set changes the composite little (-0.060 against -0.082). Weighting the rates by their precision moves the estimate towards zero (-0.033) and makes it more precise, with a minimum detectable difference of 0.12 instead of 0.18.

6.2 | Research and quality

The study started from the idea that research improves the quality of care in the areas where it is concentrated. At the level of the PNE indicators the data do not support it. The difference between an AHC and its control is not more favourable in the areas where the AHC publishes relatively more, whether research is measured at baseline, in the years just before the outcomes or by its citation impact, and the design would have detected moderation of about a tenth of a standard deviation per standard deviation of specialisation. Nor does a favourable association appear in the AHC's own rates, where an effect of its research would show: the AHC's and the control's rates move in the same direction. Several explanations are compatible with this result. Research may improve care in the regional network rather than only in the research hospital [9]. The PNE indicators, designed to monitor common conditions and procedures, may not capture the complex cases in which research hospitals differ. And the measure of research, based on declared affiliations and on the clinical areas of the indicators, is imperfect (Section 6.4). The result concerns the association between research specialisation and PNE-measured quality, not the value of research.

6.3 | Quality, not efficiency

The study compares the quality of care, measured by risk-adjusted outcomes and processes. It does not compare the resources used to produce that care, so it cannot say whether AHCs are more or less efficient than non-academic hospitals, nor how any hospital could improve its efficiency. That question requires data on costs, staff and activity and a production or cost-frontier model [73]; since AHCs also produce teaching and research, it would also require an allocation of joint costs.

6.4 | Limitations

Design. The comparison is descriptive. Patient-level risk adjustment does not capture every difference in case mix, and tertiary referral patterns may still differ within pairs. The structural covariates refer to 2015 and may themselves reflect academic status, and the emergency level refers to 2025; we therefore report the full ladder of designs rather than a single adjusted estimate.

Coverage. The main comparison covers 22 of the 47 AHCs, those that have a comparable non-academic hospital in their region and share at least five indicators with it. Its results do not extend to the largest teaching hospitals or to the specialised research institutes, nor to private hospitals, which are not in the control pool.

Inference. With 22 pairs, inference relies on a t reference distribution with few degrees of freedom and on the wild cluster bootstrap, whose calibration we checked by simulation.

Measurement of quality. The known variances are those implied by the standard errors that PNE publishes, which coincide with the binomial standard errors of the logits of the adjusted rates (median ratio 1.02 in the cells of the main comparison); they do not reflect the uncertainty of the risk-adjustment models.

Measurement of research. The measure of research depends on the affiliations that authors declare, on MeSH indexing that thins out over time and almost disappears in 2024, and on clinical areas defined by the indicators. The moderator captures relative specialisation, not absolute amount, and is estimated with error, which the model absorbs in the pair $\times$ area component but which limits the power to detect moderation.

7 | CONCLUSIONS

Italian Academic Health Centres that can be compared with a non-academic public hospital of the same region and similar structure deliver quality of care, as measured by the National Outcomes Programme, that is indistinguishable from that of their controls, and the difference does not depend on where their research is concentrated. The advantage that the AHCs show over the hospitals of their region taken together reflects their structure. For the largest teaching hospitals and the specialised research institutes, which have no such comparator, the question remains open and needs other designs.

Beyond this application, the analysis illustrates three practices for comparisons of service providers on panels of indicators: comparing units on the indicators they share, matching on the structure that the evaluation holds fixed and reporting how the estimate changes along that ladder, and modelling the sampling noise of each indicator rather than averaging it away. In our data the second changes the conclusion and the third its precision; the first matters little here, but it is what makes each comparison one of like with like.

AUTHOR CONTRIBUTIONS

Massimo Aria: Methodology, Supervision, Writing - Review & Editing. **Sergio Beraldo**: Supervision, Validation, Writing - Review & Editing. **Michela Collaro**: Conceptualization, Validation, Formal analysis, Data Curation, Writing - Review & Editing. **Agostino Gnasso**: Conceptualization, Methodology, Software, Data Curation, Formal analysis, Visualization, Writing - Original Draft, Project administration.

FINANCIAL DISCLOSURE

See Acknowledgments.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY

The SciK-Health bibliometric indicators are publicly available at https://github.com/massimoaria/scikhealth and https://massimoaria.github.io/scikhealth/. The Programma Nazionale Esiti outcome indicators are published by AGENAS at https://pne.agenas.it/. The Ministry of Health data on beds by facility and discipline and the list of emergency facilities are distributed through the Ministry's open-data portal. OpenAlex data are available under a CC0 licence at https://openalex.org.

CODE AVAILABILITY

The R code that reproduces every number, table and figure of this paper from the public input data is available from the corresponding author on request.

How to cite this article: Aria M, Beraldo S, Collaro M, Gnasso A. Evaluating the quality of hospital services: a matched, precision-weighted comparison of Italian academic health centres and non-academic public hospitals. *Appl Stoch Models Bus Ind.* 2026;00(0):1–00.

APPENDIX

A INDICATORS

Table A1 lists the 41 indicators with their type, median reliability of the hospital rates and the indicator-specific AHC − control difference from the main model. Reliability is $\lambda = \tau^2/(\tau^2 + \sigma^2)$, where σ is the standard error of the logit of the rate and τ^2 is estimated by restricted maximum likelihood for each indicator and year across the public hospitals, AHCs included.

B THE DIRICHLET-WEIGHTED COMPOSITE

Let $w \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$ on K components. Then $\text{E}(w_i) = 1/K$, $\text{Var}(w_i) = (K-1)/\{K^2(K\alpha+1)\}$ and $\text{Cov}(w_i, w_j) = -1/\{K^2(K\alpha+1)\}$ for $i \neq j$. For fixed z, $\text{E}(C \mid z) = \sum_i z_i/K = \bar{z}$, and

$$\text{Var}(C \mid z) = \frac{(K-1)\sum_i z_i^2 - \sum_{i \neq j} z_i z_j}{K^2(K\alpha+1)} = \frac{K\sum_i z_i^2 - (\sum_i z_i)^2}{K^2(K\alpha+1)} = \frac{1}{K(K\alpha+1)} \sum_i (z_i - \bar{z})^2.$$

A Monte Carlo average of R draws therefore approximates the equal-weight composite, whatever the correlation among the indicators, and the spread of the draws measures the dispersion of the hospital's own indicators, not the uncertainty of its quality. Table A2 reports the analysis with common random weights described in Section 4.3.

C MATCHING DESIGNS

Table A3 lists the designs and Figure A1 shows the balance of the main design; Table A4 lists the AHCs without a comparable control in their region. Each design is summarised with the same estimator: for each AHC, the weighted mean of its differences from its controls on the common indicators, with standard errors clustered on AHCs and controls.

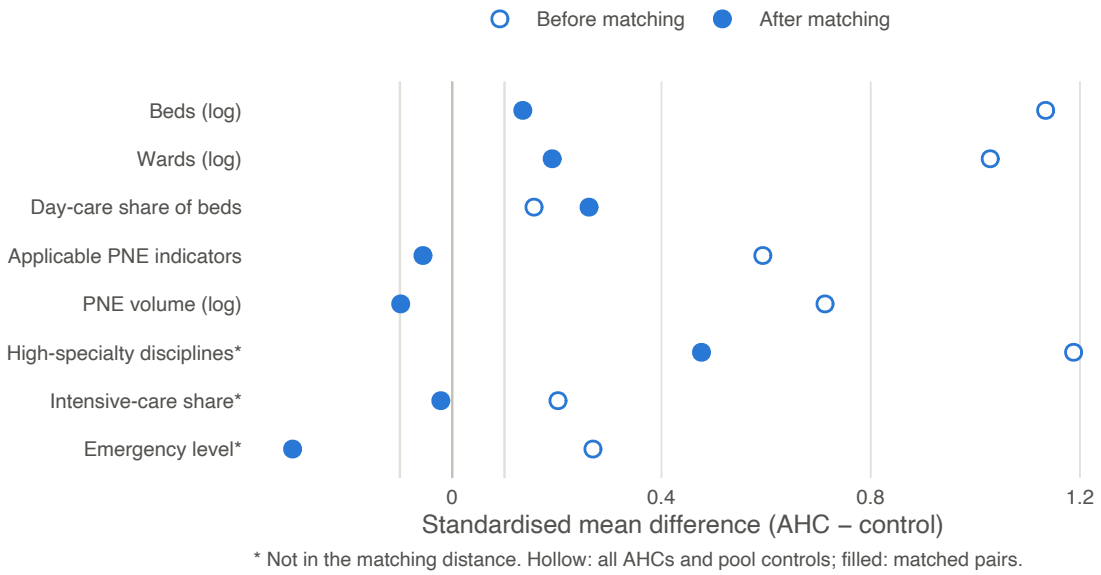


FIGURE A1 Standardised mean differences before matching (all AHCs and pool controls) and after matching (main design).

D ESTIMATION AND INFERENCE

The model of equation (4) is fitted by restricted maximum likelihood on the block-diagonal covariance matrix by pair, with the fixed effects obtained by generalised least squares at the estimated variance components. With a constant pair $\times$ indicator component only, the estimates coincide with those of `metafor::rma.mv`, and without known variances with those of `lme4`. The composite weights ω_i of equation (2) define the main estimand as a linear combination of the indicator effects. Inference uses the t distribution with $G - 1$ degrees of freedom; CR1 standard errors cluster by pair; the wild cluster bootstrap imposes the null hypothesis and uses Webb's six-point weights. In 400 data sets simulated from the fitted model under the null hypothesis, the 5% tests of the main estimand reject in 3.5%, 4.5% and 5.0% of the cases; for the average with equal indicator weights, CR1 rejects in 23.2%, because some indicators appear in only one or two pairs, and the model standard error is used instead. For the trend the rejection rates are 4.0%, 6.5% and 6.0%; for γ_S, 3.2%, 5.2% and 4.5%.

E CHANGE OVER TIME: SENSITIVITY

Removing one pair at a time, the trend ranges from -0.0031 to 0.0028 standard deviations per year.

F RESEARCH SPECIALISATION

Table A6 reports, for each clinical area, the indicators, the pertinent articles of the 49 AHCs in 2010–14, the number of AHCs of the main comparison with none, the distribution of the activity index across these AHCs, and the precision and recall of the title rule against major-topic MeSH descriptors in 2010–19. Table A7 and Figure A2 report the sensitivity analyses.

The coefficient of the pair mean of the moderator was not a pre-specified test. With the within-pair coefficient it implies a between-pair slope of -0.437 (95% CI -0.867 to -0.007; $p = 0.047$), that is -0.085 per standard deviation of the pair mean across the 22 pairs: pairs whose AHC is specialised in the areas of their common indicators have slightly lower differences. The comparison is between 22 pairs, depends on which indicators each pair shares and on unmeasured characteristics of the pairs, and changes sign in the extension to all common cells (0.062, $p = 0.42$); we do not interpret it as an effect of research.

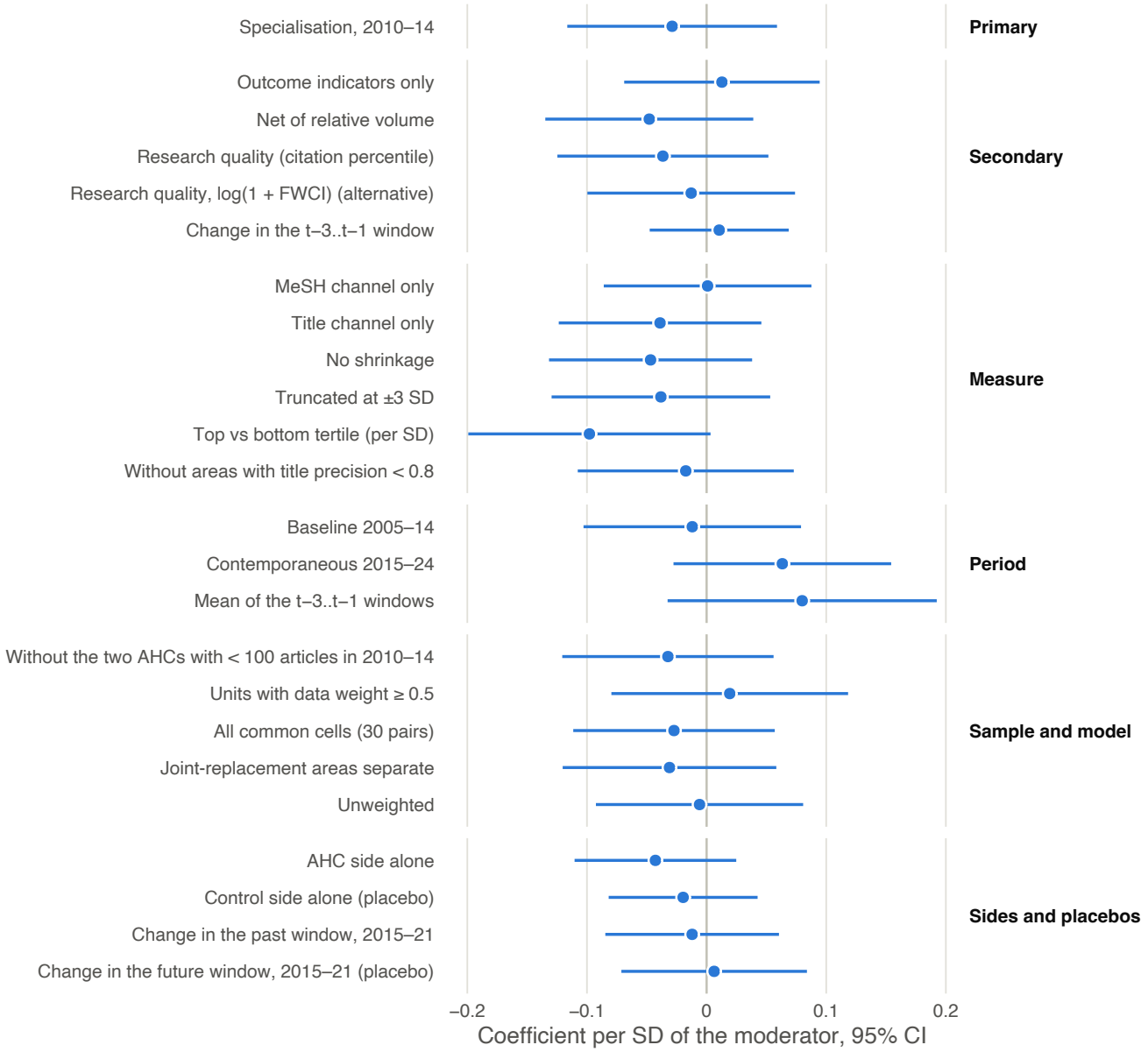


FIGURE A2 Moderation coefficients with 95% confidence intervals: family of tests (with research quality also measured as log(1 + FWCI), outside the family), sensitivity analyses, sides and placebos.

TABLE A1 The 41 indicators by clinical domain: type, median reliability of the rates, number of pairs in which the indicator is compared, and AHC − control difference with 95% confidence interval (main model). Table A6 groups them by clinical area.

Code	Indicator	Type	Reliability	Pairs	Estimate	95% CI
Cardiovascular						
15	Heart failure: 30-day mortality	Outcome	0.68	18	0.19	[−0.17, 0.54]
34	Heart valve repair or replacement: 30-day mortality	Outcome	0.42	8	−0.11	[−0.68, 0.47]
55	Unruptured abdominal aortic aneurysm repair: 30-day mortality	Outcome	0.00	13	0.41	[−0.28, 1.10]
95	STEMI: 30-day mortality	Outcome	0.42	17	0.07	[−0.31, 0.45]
201	Heart failure: 30-day readmissions	Outcome	0.34	18	0.02	[−0.35, 0.38]
409	Pulmonary embolism: 30-day mortality	Outcome	0.29	13	−0.11	[−0.54, 0.33]
410	Pulmonary embolism: 30-day readmissions	Outcome	0.06	12	−0.25	[−0.71, 0.21]
513	Lower-limb arterial disease, stage III–IV: amputation within 6 months	Outcome	0.18	6	0.39	[−0.38, 1.16]
514	Lower-limb arterial disease, stage III–IV: revascularisation within 6 months	Process	0.19	6	0.06	[−0.62, 0.74]
515	Lower-limb arterial disease, stage III–IV: 6-month mortality	Outcome	0.31	8	0.33	[−0.25, 0.90]
620	Acute myocardial infarction: 30-day mortality	Outcome	0.54	19	0.23	[−0.15, 0.62]
621	Coronary angioplasty (PTCA)	Process	0.93	18	−0.06	[−0.43, 0.32]
630	Isolated coronary artery bypass graft: 30-day mortality	Outcome	0.34	8	0.10	[−0.64, 0.84]
Cerebrovascular						
18	Ischaemic stroke: 30-day mortality	Outcome	0.65	17	0.13	[−0.24, 0.50]
19	Ischaemic stroke: 30-day readmissions	Outcome	0.28	16	0.02	[−0.38, 0.42]
Digestive						
28	Laparoscopic cholecystectomy, inpatient: 30-day complications	Outcome	0.25	18	−0.15	[−0.55, 0.25]
29	Laparoscopic cholecystectomy, inpatient: reoperation within 30 days	Outcome	0.06	18	−0.38	[−0.85, 0.10]
202	Laparoscopic cholecystectomy in day or one-day surgery	Process	0.98	17	−0.03	[−0.39, 0.32]
302	Laparoscopic cholecystectomy: post-operative stay under 3 days	Process	0.85	20	−0.20	[−0.53, 0.12]
Musculoskeletal						
38	Hip fracture: 30-day mortality	Outcome	0.38	17	−0.04	[−0.42, 0.35]
66	Hip replacement: 30-day readmissions	Outcome	0.20	18	−0.26	[−0.64, 0.12]
69	Knee replacement: 30-day readmissions	Outcome	0.03	9	−0.21	[−1.09, 0.66]
79	Shoulder replacement: 30-day readmissions	Outcome	0.00	3	−0.35	[−2.13, 1.42]
640	Hip fracture: surgery within 48 hours of arrival	Process	0.97	18	−0.31	[−0.68, 0.06]
Oncology						
63	Lung cancer surgery: 30-day mortality	Outcome	0.09	9	0.15	[−0.74, 1.03]
82	Stomach cancer surgery: 30-day mortality	Outcome	0.13	5	0.22	[−0.74, 1.17]
83	Colon cancer surgery: 30-day mortality	Outcome	0.13	22	0.10	[−0.27, 0.46]
84	Rectal cancer surgery: 30-day mortality	Outcome	0.00	1	−1.25	[−6.51, 4.01]
205	Prostate cancer surgery: 30-day readmissions	Outcome	0.22	8	−0.14	[−0.78, 0.50]
306	Kidney cancer surgery: 30-day mortality	Outcome	0.00	11	0.77	[−0.22, 1.77]
310	Liver cancer surgery: 30-day mortality	Outcome	0.03	4	−0.31	[−1.73, 1.11]
601	Brain tumour surgery (craniotomy): 30-day mortality	Outcome	0.18	11	0.10	[−0.43, 0.63]
606	Mastectomy for invasive breast cancer: reconstruction or expander at index admission	Process	0.66	11	−0.39	[−0.84, 0.07]
612	Breast-conserving surgery for breast cancer: re-excision within 90 days	Outcome	0.67	16	−0.11	[−0.49, 0.27]
727	Pancreatic cancer resection: 90-day mortality	Outcome	0.30	1	0.93	[−1.50, 3.37]
Perinatal						
37	Primary caesarean section	Process	0.96	17	−0.08	[−0.43, 0.27]
204	Vaginal delivery: readmission during the puerperium	Outcome	0.25	17	−0.25	[−0.66, 0.16]
318	Caesarean delivery: readmission during the puerperium	Outcome	0.13	17	−0.07	[−0.48, 0.34]
Respiratory						
21	COPD exacerbation: 30-day mortality	Outcome	0.67	18	0.01	[−0.34, 0.37]
22	COPD exacerbation: 30-day readmissions	Outcome	0.20	18	0.05	[−0.32, 0.42]
Urogenital						
602	Chronic kidney disease: 30-day mortality	Outcome	0.65	16	0.11	[−0.27, 0.50]

TABLE A2 Common random weights (2,000 draws per scheme, one weight vector per draw applied to all hospitals). Columns: composite difference with equal weights; 95% weight-sensitivity interval of the difference across draws; share of draws with $p < 0.05$; median Spearman correlation between the ranking of the 4,042 public hospital–years and the equal-weight ranking; share of hospital–years that change quintile.

Weighting scheme	Equal weights	Weight-sensitivity interval	$p < 0.05$	Rank correlation	Quintile change
Dirichlet(0.5): dispersed weights	−0.082	[−0.185, 0.024]	4.7%	0.73	55%
Dirichlet(1): uniform on the simplex	−0.082	[−0.160, −0.003]	2.8%	0.82	48%
Dirichlet(5): close to equal weights	−0.082	[−0.118, −0.046]	0.2%	0.95	29%
Hierarchical Dirichlet(1): domains, then indicators	−0.081	[−0.201, 0.043]	5.6%	0.77	52%

TABLE A3 The 16 designs: AHCs and controls in the estimate, mean absolute standardised mean difference of the eight covariates of Table 2 after matching, and AHC − control difference with 95% confidence interval on all indicators and on outcome indicators (equal-weight composite of the common indicators).

	Design	AHCs	Controls	Mean |SMD|	All indicators	Outcome indicators
Main design						
D01	Main: optimal 1:1, exact region, caliper 0.5 SD, covariates M2	22	22	0.19	−0.082 [−0.210, 0.047]	0.035 [−0.065, 0.134]
Covariate specification						
D02	Region only: all controls in the region, equal weights	29	304	0.52	0.144 [0.056, 0.232]	0.152 [0.041, 0.262]
D03	M1: region and beds	23	23	0.29	0.011 [−0.114, 0.136]	0.049 [−0.098, 0.195]
D04	M3: M2 plus high specialty, intensive care, emergency level	22	22	0.21	−0.057 [−0.194, 0.080]	0.066 [−0.058, 0.189]
D16	M3 with exact emergency level within region	20	20	0.13	−0.009 [−0.104, 0.085]	−0.003 [−0.118, 0.111]
Matching method						
D05	Nearest neighbour 1:1 with replacement	22	19	0.20	−0.062 [−0.179, 0.056]	−0.007 [−0.117, 0.103]
D06	Optimal full matching	26	65	0.27	0.064 [−0.016, 0.145]	0.097 [0.007, 0.187]
D07	Coarsened exact matching (region, beds, PNE indicators)	15	30	0.20	−0.055 [−0.230, 0.119]	−0.028 [−0.238, 0.183]
D08	Entropy balancing (region and M2 covariates)	30	371	0.12	0.050 [−0.050, 0.149]	0.062 [−0.042, 0.166]
D09	Genetic matching 1:1, QQ balance criterion	22	22	0.16	−0.052 [−0.186, 0.082]	0.014 [−0.110, 0.138]
Common support						
D10	Caliper 0.25 SD	17	17	0.25	−0.045 [−0.231, 0.142]	0.006 [−0.158, 0.170]
D11	No caliper: all AHCs	31	31	0.29	0.059 [−0.011, 0.129]	0.093 [0.022, 0.164]
D12	Macro-area fallback for AHCs without a regional control	28	28	0.20	−0.030 [−0.149, 0.090]	0.060 [−0.039, 0.160]
D13	Entropy balancing on all AHCs	39	371	0.11	0.056 [−0.044, 0.156]	0.075 [−0.031, 0.181]
Control pool						
D14	Pool including academic, IRCCS, same-company or private controls	22	22	0.15	−0.067 [−0.200, 0.065]	0.038 [−0.063, 0.140]
D15	Excluding controls of companies with $\geq$ 10 university physicians	21	21	0.22	−0.083 [−0.220, 0.054]	0.047 [−0.058, 0.151]

TABLE A4 AHCs with no non-academic public hospital within the caliper in their region: beds in 2015, beds of the largest control in the region, and difference from all the non-academic public hospitals of the region (equal-weight composite of the common indicators).

Hospital	Region	Beds (2015)	Largest control in region	vs regional controls
Sant'Orsola Polyclinic, Bologna	Emilia-Romagna	1,573	631	−0.036
University Hospital of Padua	Veneto	1,452	935	0.027
Careggi University Hospital, Florence	Tuscany	1,326	547	0.228
San Martino Polyclinic, Genoa	Liguria	1,318	467	0.097
University Hospital of Pisa	Tuscany	1,213	547	0.170
University Hospital of Parma	Emilia-Romagna	1,144	631	0.186
Molinette, Turin	Piedmont	1,142	711	0.188
Bari Polyclinic	Apulia	1,108	632	0.404
Umberto I–Lancisi, Ancona	Marche	583	365	0.303

TABLE A5 Sensitivity of the period contrasts and of the trend: estimates and wild bootstrap p values.

Analysis	Pairs	2020–21 vs 2015–19		2022–24 vs 2015–19		Trend per year	
		Est.	p	Est.	p	Est.	p
Main specification	22	0.014	0.76	−0.015	0.73	−0.0004	0.95
Outcome indicators only	22	0.051	0.27	−0.031	0.50	−0.0039	0.60
Fixed set of indicators present in all years	22	0.010	0.84	−0.011	0.78	0.0003	0.96
Pairs observed in all ten years	21	0.026	0.57	−0.005	0.90	0.0004	0.95
Excluding 2020 and 2021	22			−0.005	0.91	0.0004	0.94
Design M3	22	0.019	0.71	0.011	0.78	0.0031	0.59
Design M3 with exact emergency level	20	0.007	0.90	0.002	0.96	0.0040	0.53
All common cells (extension)	30	0.008	0.85	−0.021	0.63	−0.0012	0.84
Unweighted	22	0.079	0.072	0.026	0.56	0.0064	0.39

TABLE A6 The 24 clinical areas: indicators (PNE codes), pertinent articles in 2010–14, AHCs of the main comparison with no pertinent article, activity index of these AHCs (median, 10th–90th percentile; 1 is the average profile of all AHCs), and precision and recall of the title rule against major-topic MeSH descriptors, 2010–19.

Clinical area	Indicators	Articles 2010–14	AHCs with none	Activity index	Precision	Recall
Breast cancer	606, 612	1,115	0	0.76 (0.17–1.61)	0.93	0.80
Lung cancer	63	812	2	0.89 (0.34–1.91)	0.96	0.67
Rectal cancer	84	809	2	0.76 (0.31–1.15)	0.94	0.80
Colon cancer	83	785	1	0.73 (0.38–1.20)	0.93	0.76
Liver cancer	310	629	3	0.83 (0.21–1.48)	0.95	0.70
Acute coronary syndrome and PCI	620, 95, 621	621	3	0.75 (0.23–2.17)	0.97	0.69
Brain tumours	601	579	5	0.65 (0.18–1.27)	0.96	0.76
Prostate cancer	205	474	5	0.79 (0.27–2.02)	0.96	0.85
Kidney cancer	306	405	3	0.90 (0.50–2.30)	0.95	0.74
Chronic kidney disease	602	364	2	1.37 (0.58–2.52)	0.92	0.65
Pancreatic cancer	727	352	6	0.57 (0.28–1.43)	0.95	0.74
Ischaemic stroke	18, 19	298	4	0.99 (0.41–2.57)	0.93	0.68
Heart failure	15, 201	296	3	1.34 (0.51–2.54)	0.97	0.64
Heart valve surgery	34	260	8	0.87 (0.25–1.74)	1.00	0.82
Joint arthroplasty (hip, knee, shoulder)	66, 69, 79	232	9	0.41 (0.04–1.25)	0.91	0.67
Gastric cancer	82	155	10	0.74 (0.42–1.62)	0.97	0.64
COPD	21, 22	149	8	0.96 (0.64–1.87)	0.99	0.77
Pulmonary embolism	409, 410	147	6	0.88 (0.50–1.81)	0.95	0.77
Childbirth (caesarean and vaginal)	37, 204, 318	73	11	0.78 (0.40–1.83)	0.73	0.44
Abdominal aortic aneurysm	55	56	11	0.83 (0.43–2.43)	0.92	0.66
Peripheral artery disease	513, 514, 515	51	12	0.88 (0.41–2.76)	0.76	0.59
Cholecystectomy and gallstone disease	28, 29, 202, 302	48	11	0.91 (0.55–2.87)	1.00	0.59
Coronary artery bypass grafting	630	47	13	0.89 (0.61–2.44)	1.00	0.55
Hip fracture	38, 640	39	15	0.70 (0.36–1.74)	0.93	0.81

TABLE A7 Sensitivity of the moderation coefficient γ_S (per standard deviation of each version of the moderator). *Added after the main estimates.

Sensitivity analysis	Pairs	Cells	Estimate	95% CI	p
Measure: MeSH channel only	22	4,323	0.001	[−0.086, 0.088]	0.98
Measure: title channel only	22	4,323	−0.039	[−0.124, 0.046]	0.35
Measure: no shrinkage, $\log\{(n+0.5)/(E+0.5)\}$	22	4,323	−0.047	[−0.132, 0.038]	0.26
Measure: S truncated at ± 3 SD	22	4,323	−0.038	[−0.130, 0.053]	0.39
Measure: tertiles of S (top vs bottom, per SD)	22	4,323	−0.098	[−0.199, 0.003]	0.057
Measure: without areas with title precision < 0.8*	22	3,688	−0.017	[−0.108, 0.073]	0.69
Period: 2005–14	22	4,323	−0.012	[−0.103, 0.079]	0.79
Period: 2015–24 (contemporaneous)	22	4,323	0.063	[−0.028, 0.154]	0.16
Period: mean of the $t-3..t-1$ windows	22	4,323	0.080	[−0.033, 0.192]	0.15
Sample: without the two AHCs with < 100 articles in 2010–14	20	4,200	−0.032	[−0.121, 0.056]	0.45
Sample: units with data weight ≥ 0.5	21	3,083	0.019	[−0.080, 0.118]	0.69
Sample: all common cells (extension)	30	4,475	−0.027	[−0.111, 0.057]	0.51
Areas: hip, knee and shoulder replacement separate	22	4,323	−0.031	[−0.120, 0.058]	0.48
Weights: unweighted model	22	4,323	−0.006	[−0.093, 0.081]	0.89

REFERENCES

1. Ayanian JZ, Weissman JS. Teaching hospitals and quality of care: A review of the literature. *Milbank Quarterly.* 2002;80(3):569–593. doi: 10.1111/1468-0009.00023
2. Burke LG, Frakt AB, Khullar D, Orav EJ, Jha AK. Association Between Teaching Status and Mortality in US Hospitals. *JAMA.* 2017;317(20):2105. doi: 10.1001/jama.2017.5702
3. Ozdemir BA, Karthikesalingam A, Sinha S, et al. Research activity and the association with mortality. *PLOS ONE.* 2015;10(2):e0118253. doi: 10.1371/journal.pone.0118253
4. Parasuraman A, Zeithaml VA, Berry LL. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing.* 1988;64(1):12–40.
5. Donabedian A. Evaluating the quality of medical care. *The Milbank Quarterly.* 1966;44(3):166–203. doi: 10.2307/3348969
6. Goddard M, Jacobs R. *Using composite indicators to measure performance in health care*:339–368; Cambridge: Cambridge University Press . 2010.
7. Allison JJ, Kiefe CI, Weissman NW, et al. Relationship of hospital teaching status with quality of care and mortality for Medicare patients with acute MI. *Journal of the American Medical Association.* 2000;284(10):1256–1262. doi: 10.1001/jama.284.10.1256
8. Khuri SF, Najjar SF, Daley J, et al. Comparison of surgical outcomes between teaching and nonteaching hospitals in the Department of Veterans Affairs. *Annals of Surgery.* 2001;234(3):370–383. doi: 10.1097/00000658-200109000-00011
9. Burke LG, Burke RC, Orav EJ, Duggan CE, Figueroa JF, Jha AK. Association of academic medical center presence with clinical outcomes at neighboring community hospitals among Medicare beneficiaries. *JAMA Network Open.* 2023;6(2):e2254559.
10. Aria M, Cuccurullo C. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics.* 2017;11(4):959–975. doi: 10.1016/j.joi.2017.08.007
11. Priem J, Piwowar H, Orr R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. In: 26th International Conference on Science, Technology and Innovation Indicators (STI 2022), Granada. Zenodo; 2022. doi: 10.5281/zenodo.6936227.
12. Waltman L. A review of the literature on citation impact indicators. *Journal of Informetrics.* 2016;10(2):365–391. doi: 10.1016/j.joi.2016.02.007
13. Hicks D, Wouters P, Waltman L, Rijcke dS, Rafols I. Bibliometrics: The Leiden Manifesto for research metrics. *Nature.* 2015;520:429–431. doi: 10.1038/520429a
14. Wilsdon J. *The Metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management*. London: SAGE / HEFCE, 2015.
15. Schubert A, Braun T. Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics.* 1986;9(5-6):281–291. doi: 10.1007/bf02017249
16. Balassa B. Trade Liberalisation and "Revealed" Comparative Advantage. *The Manchester School.* 1965;33(2):99–123. doi: 10.1111/j.1467-9957.1965.tb00050.x
17. D'Aniello L, Spano M, Cuccurullo C, Aria M. Academic health centers' configurations, scientific productivity, and impact: insights from the Italian setting. *Health Policy.* 2022;126(12):1317–1323. doi: 10.1016/j.healthpol.2022.09.007
18. Aria M, Cuccurullo C, D'Aniello L, Spano M. SciK-Health: an open-data dashboard for the multidimensional evaluation of Italian academic health science centers. *Quality & Quantity.* 2026;60(5):16437–16466. doi: 10.1007/s11135-026-02827-6
19. OECD , European Commission, Joint Research Centre . *Handbook on Constructing Composite Indicators: Methodology and User Guide*. Paris: OECD Publishing, 2008
20. Freudenberg M. Composite indicators of country performance: A critical assessment. Tech. Rep. 2003/16, OECD Science, Technology and Industry Directorate; Paris: 2003.
21. Saltelli A. Composite indicators between analysis and advocacy. *Social Indicators Research.* 2007;81(1):65–77. doi: 10.1007/s11205-006-0024-9
22. Greco S, Ishizaka A, Tasiou M, Torrisi G. On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. *Social Indicators Research.* 2019;141(1):61–94. doi: 10.1007/s11205-017-1832-9
23. Paruolo P, Saisana M, Saltelli A. Ratings and rankings: Voodoo or science?. *Journal of the Royal Statistical Society: Series A.* 2013;176(3):609–634. doi: 10.1111/j.1467-985X.2012.01059.x
24. Becker W, Saisana M, Paruolo P, Vandecasteele I. Weights and importance in composite indicators: Closing the gap. *Ecological Indicators.* 2017;80:12–22. doi: 10.1016/j.ecolind.2017.03.056
25. Luzzati T, Gucciardi G. A non-simplistic approach to composite indicators and rankings: an illustration by comparing the sustainability of the EU Countries. *Ecological economics.* 2015;113:25–38.
26. Mazziotta M, Pareto A. On a generalized non-compensatory composite index for measuring socio-economic phenomena. *Social Indicators Research.* 2016;127(3):983–1003. doi: 10.1007/s11205-015-0998-2
27. Saisana M, Saltelli A, Tarantola S. Uncertainty and Sensitivity Analysis Techniques as Tools for the Quality Assessment of Composite Indicators. *Journal of the Royal Statistical Society Series A: Statistics in Society.* 2005;168(2):307–323. doi: 10.1111/j.1467-985x.2005.00350.x
28. Jacobs R, Goddard M, Smith PC. How Robust Are Hospital Ranks Based on Composite Performance Measures?. *Medical Care.* 2005;43(12):1177–1184. doi: 10.1097/01.mlr.0000185692.72905.4a
29. Cavicchia C, Vichi M. Statistical Model-Based Composite Indicators for Tracking Coherent Policy Conclusions. *Social Indicators Research.* 2021;156(2-3):449–479. doi: 10.1007/s11205-020-02318-7
30. Cavicchia C, Sarnacchiaro P, Vichi M. A composite indicator for the waste management in the EU via Hierarchical Disjoint Non-Negative Factor Analysis. *Socio-Economic Planning Sciences.* 2021;73:100832. doi: 10.1016/j.seps.2020.100832
31. Cavicchia C, Vichi M, Zaccaria G. Hierarchical disjoint principal component analysis. *AStA Advances in Statistical Analysis.* 2023;107(3):537–574. doi: 10.1007/s10182-022-00458-4
32. Cavicchia C, Sarnacchiaro P, Vichi M, Zaccaria G. A model-based ultrametric composite indicator for studying waste management in Italian municipalities. *Computational Statistics.* 2024;39(1):21–50. doi: 10.1007/s00180-023-01333-9
33. Freni-Sterrantino A, Prescott TP, Ceely G, Glickman M, Holmes C. Assessments and developments in constructing a National Health Index for policy-making, in the UK. *Journal of the Royal Statistical Society Series A: Statistics in Society.* 2025;188(2):491–514. doi: 10.1093/jrsssa/qnae060
34. Bollen K, Lennox R. Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin.* 1991;110(2):305–314. doi: 10.1037/0033-2909.110.2.305
35. Diamantopoulos A, Winklhofer HM. Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research.* 2001;38(2):269–277. doi: 10.1509/jmkr.38.2.269.18845

36. Goldstein H, Spiegelhalter DJ. League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society).* 1996;159(3):385. doi: 10.2307/2983325
37. Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Statistics in Medicine.* 2005;24(8):1185–1202. doi: 10.1002/sim.1970
38. Normand SLT, Glickman ME, Gatsonis CA. Statistical Methods for Profiling Providers of Medical Care: Issues and Applications. *Journal of the American Statistical Association.* 1997;92(439):803–814. doi: 10.1080/01621459.1997.10474036
39. Dimick JB, Welch HG, Birkmeyer JD. Surgical Mortality as an Indicator of Hospital Quality: The Problem With Small Sample Size. *JAMA.* 2004;292(7):847. doi: 10.1001/jama.292.7.847
40. Rubin DB. *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press, 2006.
41. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70(1):41–55.
42. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics.* 2004;86(1):4–29. doi: 10.1162/003465304323023651
43. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. *Econometrica.* 2006;74(1):235–267. doi: 10.1111/j.1468-0262.2006.00655.x
44. Stuart EA. Matching methods for causal inference: A review and a look forward. *Statistical Science.* 2010;25(1):1–21. doi: 10.1214/09-STS313
45. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis.* 2007;15(3):199–236. doi: 10.1093/pan/mpl013
46. Iacus SM, King G, Porro G. Causal inference without balance checking: Coarsened exact matching. *Political Analysis.* 2012;20(1):1–24. doi: 10.1093/pan/mpr013
47. King G, Nielsen R. Why propensity scores should not be used for matching. *Political Analysis.* 2019;27(4):435–454. doi: 10.1017/pan.2019.11
48. Rosenbaum PR. The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment. *Journal of the Royal Statistical Society. Series A (General).* 1984;147(5):656–666. doi: 10.2307/2981697
49. Hansen BB, Klopfer SO. Optimal Full Matching and Related Designs via Network Flows. *Journal of Computational and Graphical Statistics.* 2006;15(3):609–627. doi: 10.1198/106186006x137047
50. Hainmueller J. Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis.* 2012;20(1):25–46. doi: 10.1093/pan/mpr025
51. Diamond A, Sekhon JS. Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics.* 2013;95(3):932–945. doi: 10.1162/REST_a_00318
52. Abadie A, Imbens GW. Bias-Corrected Matching Estimators for Average Treatment Effects. *Journal of Business & Economic Statistics.* 2011;29(1):1–11. doi: 10.1198/jbes.2009.07333
53. Rosenbaum PR. *Observational Studies*. Springer New York, 2002
54. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Annals of Internal Medicine.* 2017;167(4):268–274. doi: 10.7326/m16-2607
55. Beraldo S, Collaro M, Marino I. Patient migration as a response to the regulation of subnational healthcare budgets. *Regional Studies.* 2023;57(11):2207–2219.
56. Fusco D, Barone AP, Sorge C, et al. P.Re.Val.E.: outcome research program for the evaluation of health care quality in Lazio, Italy. *BMC Health Services Research.* 2012;12:25. doi: 10.1186/1472-6963-12-25
57. Colais P, Pinnarelli L, Mataloni F, et al. The National Outcomes Evaluation Programme in Italy: The Impact of Publication of Health Indicators. *International Journal of Environmental Research and Public Health.* 2022;19(18):11685. doi: 10.3390/ijerph191811685
58. Lenzi J, Avaldi VM, Rucci P, Pieri G, Fantini MP. Risk-adjustment models for heart failure patients' 30-day mortality and readmission rates: the incremental value of clinical data abstracted from medical charts beyond hospital discharge record. *BMC Health Services Research.* 2016;16:473. doi: 10.1186/s12913-016-1731-9
59. Fiore M, Bianconi A, Acuti Martellucci C, et al. Impact of the Italian Healthcare Outcomes Program (PNE) on the Care Quality of the Poorest Performing Hospitals. *Healthcare.* 2024;12(4):431. doi: 10.3390/healthcare12040431
60. Kelley E, Hurst J. Health care quality indicators project: conceptual framework paper. OECD Health Working Papers No. 23, OECD Publishing; Paris: 2006.
61. Busse R, Klazinga N, Panteli D, Quentin W, eds. *Improving Healthcare Quality in Europe: Characteristics, Effectiveness and Implementation of Different Strategies*. Copenhagen: European Observatory on Health Systems and Policies, 2019. Health Policy Series No. 53.
62. Lipscomb CE. Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association.* 2000;88(3):265–266.
63. Pinheiro JC, Bates DM. *Mixed-Effects Models in S and S-PLUS*. New York: Springer, 2000
64. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software.* 2010;36(3). doi: 10.18637/jss.v036.i03
65. Cameron AC, Gelbach JB, Miller DL. Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics.* 2008;90(3):414–427. doi: 10.1162/rest.90.3.414
66. Webb MD. Reworking wild bootstrap-based inference for clustered errors. *Canadian Journal of Economics/Revue canadienne d'économique.* 2023;56(3):839–858. doi: 10.1111/caje.12661
67. MacKinnon JG, Nielsen MØ, Webb MD. Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics.* 2023;232(2):272–299. doi: 10.1016/j.jeconom.2022.04.001
68. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics.* 1979;6(2):65–70.
69. Clayton D, Kaldor J. Empirical Bayes Estimates of Age-Standardized Relative Risks for Use in Disease Mapping. *Biometrics.* 1987;43(3):671. doi: 10.2307/2532003
70. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models*. Chapman and Hall/CRC, 2006
71. Mundlak Y. On the Pooling of Time Series and Cross Section Data. *Econometrica.* 1978;46(1):69. doi: 10.2307/1913646
72. Halm EA, Lee C, Chassin MR. Is Volume Related to Outcome in Health Care? A Systematic Review and Methodologic Critique of the Literature. *Annals of Internal Medicine.* 2002;137(6):511–520. doi: 10.7326/0003-4819-137-6-200209170-00012
73. Hollingsworth B. The measurement of efficiency and productivity of health care delivery. *Health Economics.* 2008;17(10):1107–1128. doi: 10.1002/hec.1391