Scientometrics (2026) 131:3411–3441
https://doi.org/10.1007/s11192-026-05638-6

Data source effects in research performance assessment of individuals and institutions: comparing OpenAlex with established bibliographic databases

Giovanni Abramo[1] · **Tindaro Cicero**[1] · **Ciriaco Andrea D'Angelo**[2]

Received: 22 December 2025 / Accepted: 19 April 2026 / Published online: 8 May 2026
© The Author(s) 2026

Abstract

Performance metrics inform hiring, promotion, and funding decisions; however, the choice of bibliographic database can substantially influence these outcomes. OpenAlex has emerged as a significant open alternative to proprietary sources; however, its suitability for research evaluation remains uncertain. Existing comparative studies across databases have largely focused on coverage and product-level assessments. In contrast, our work directly compares performance scores and rankings of scholars and institutions across data sources to identify the origins of any discrepancies. We rank the productivity and impact of Italian universities and professors in management engineering using OpenAlex, Web of Science (WoS), and Scopus. Productivity is measured using harmonized document types and author name disambiguation, while impact is derived from year- and field-normalized citations weighted by fractional authorship. We find strong agreement across databases for productivity, but only moderate alignment for impact. OpenAlex displays more heavily right-skewed distributions and greater dispersion, producing higher volatility in scholar and institutional rankings. Median absolute percentile differences are approximately 5–6 points across database pairs, with extremes exceeding 70 points for impact; around 30% of scholars shift at least one quartile when OpenAlex replaces WoS or Scopus. These discrepancies stem primarily from differences in field taxonomies, normalization procedures, and indexing-year conventions. We conclude that although OpenAlex enables open and reproducible analyses, high-stakes evaluations should triangulate multiple data sources and harmonize methodological assumptions in advance to ensure validity and reliability.

Keywords Research evaluation · Bibliometrics · Total fractional impact · OpenAlex · Scopus · Web of Science · Management Engineering · Italy

✉ Ciriaco Andrea D'Angelo
dangelo@dii.uniroma2.it

Giovanni Abramo
giovanni.abramo@unimercatorum.it

Tindaro Cicero
tindaro.cicero@unimercatorum.it

[1] Laboratory for Studies in Research Evaluation, Department of Engineering and Science, Universitas Mercatorum, Rome, Italy

[2] Department of Business Engineering, University of Rome "Tor Vergata", Rome, Italy

Introduction

In research-intensive organizations, bibliometric indicators are increasingly used to inform decisions about hiring, promotion, and funding. University administrators and policymakers rely on citation-based metrics and publication counts to compare individuals and institutions, allocate resources, and shape strategic priorities. Bibliometrics, however, has important limitations that are central to its responsible application. These limitations underpin concerns reflected in contemporary governance frameworks, which emphasise transparency, pluralism, and caution in the use of indicators.

Three frameworks are particularly influential in current debates. The Leiden Manifesto sets out principles for aligning metrics with institutional missions and for guarding against their misuse (Hicks et al., 2015). The UK's Metric Tide advances a framework of "responsible metrics," highlighting robustness, humility, transparency, and reflexivity in the governance of indicators (Wilsdon et al., 2015). The Declaration on Research Assessment, DORA (2012), and the Agreement on Reforming Research Assessment (CoARA, 2022) further reinforce concerns about reductive metric practices and call for broader, more context-sensitive evidence bases in research assessment. Bibliometric indicators depend critically on the underlying bibliographic databases from which they are derived. Variations in database design—such as coverage policies, classification systems, indexing conventions, and metadata completeness—can substantially impact the resulting measurements of research output and impact. This study aims to understand how these variations influence the performance scores and rankings of researchers and institutions.

Among the available data sources, Web of Science (WoS) and Scopus have long served as the dominant commercial platforms for bibliometric analysis. Their established infrastructures provide stable, curated datasets; however, their restricted accessibility and coverage biases, particularly against regional and non-English outputs, limit inclusivity and transparency (Céspedes et al., 2025). The emergence of OpenAlex (OA), a fully open and rapidly expanding bibliographic database, represents a significant shift toward open infrastructures for research assessment.

The growth of open infrastructures aligns with a broader demand for transparency and accountability in research evaluation. National assessment exercises, university rankings, and performance-based funding systems are increasingly requiring traceable and auditable metrics. In this context, the choice between open and commercial bibliographic databases is not merely technical, as it influences fairness, reproducibility, and even strategic positioning in academic systems (Waltman et al., 2012). Understanding how these data sources differ, and what implications their differences have for research assessment, is therefore of central concern to engineering managers, evaluators, and policymakers.

Despite a growing number of comparative studies, empirical evidence remains fragmented, and the extent to which open and proprietary databases yield consistent evaluations is still debated. To date, most work has focused on comparing coverage—publication and document types (Haupka et al., 2024), linguistic dimensions (Céspedes et al., 2025), geographical and disciplinary distributions (Maddi et al., 2025), and references (Culbert et al., 2025); metadata accuracy (Zhang et al., 2024); and publication-level citation impact (Scheidsteger et al., 2025; Thelwall & Jiang, 2025). Departing from these publication-level perspectives, our study examines how OA and commercial databases compare when used to evaluate the performance of individual researchers and institutions.

We investigate differences in research assessment scores and rankings by comparing performance metrics derived from OA, Scopus, and WoS for a well-defined academic

population. By focusing on a single national and disciplinary context, we enable a detailed and controlled comparison of database behavior and its implications for evaluative use. Our analysis contributes to a clearer understanding of how database selection shapes perceived research performance and provides evidence to support managers and policymakers in making more informed, transparent, and equitable decisions.

By explicitly comparing researcher- and institution-level rankings across OpenAlex, WoS, and Scopus, our study complements recent publication-level analyses and provides evidence on the systemic effects of database choice in performance-based evaluation systems.

The paper is organized as follows. "Background" section reviews the recent literature on comparative analyses of bibliographic databases. "Methods" section details our methodology. "Results" section presents our findings. "Discussion" section discusses the results, their implications, and our recommendations. "Conclusions" section concludes the study.

Background

The Open Science movement, as presented in activist and institutional discourses at the beginning of the twenty-first century, aims to make scientific research (including publications, data, physical samples, and software) and its dissemination accessible to all levels of society, amateur or professional (Parsons et al., 2022; Vicente-Saez & Martinez-Fuentes, 2018).

One of the key Open Science elements is Open Access, which is a set of principles and a range of practices through which nominally copyrightable research products are delivered to users free of access charges or other barriers (Suber, 2004).

OpenAlex is one of the infrastructures that operationalizes Open Access, explicitly integrating open-access information. OA is an open-access, free bibliographic database of the global research system, indexing scholarly works, authors, institutions, publishers, and research topics. Launched in January 2022 by OurResearch as the successor to the discontinued Microsoft Academic Graph (MAG), its name is inspired by the Library of Alexandria, which created the first bibliographic catalogue in human history (Jason, 2021).

OA incorporates MAG data and data from additional sources, i.e., DOAJ, Unpaywall, Crossref, and ORCID (Priem et al., 2022). Offering bibliometric tools and a free API, it supports research assessment, monitoring of open-access trends, and alignment with UN Sustainable Development Goals (Dalmeet, 2022).

As of March 2024, OA indexed approximately 209 million works, 13 million authors, 124,000 sources, 109,000 institutions, and 65,000 Wikidata-linked concepts. In the same year, the platform processed around 115 million API queries per month and received a $7.5 million grant from the Arcadia philanthropic initiative to reinforce its position as an open alternative (Price, 2024). The French Ministry of Research also committed financial support to OA, recognizing it as a key component of open-science infrastructure.

In recent years, several institutions have adopted OA, including Sorbonne University, which replaced Scopus, and Leiden University, which uses it for its rankings (Brooks, 2023; Clavey, 2024).

The renowned Leiden Rankings**,** curated by the Centre for Science and Technology Studies (CWTS) at Leiden University, have relied on data from the WoS for years**,** with all the limitations associated with accessibility and coverage (Waltman et al., 2012). Recently, in response to the increasing demand for transparency, along with the rise of

open infrastructures, CWTS launched the Leiden Ranking Open Edition, a parallel and experimental version of the traditional ranking, built exclusively on OA data (CWTS, 2023). This initiative pursues a dual aim: on the one hand, to increase reproducibility and transparency in the Leiden Ranking; on the other hand, to stimulate methodological comparisons between indicators constructed on closed sources (such as WoS) and those based on open sources.

The landscape of bibliographic databases is rapidly evolving, with OA emerging as a prominent open-access alternative to established, subscription-based platforms such as Scopus and WoS. Recent comparative studies have examined the strengths and limitations of OA in relation to the above two commercial platforms, focusing on their coverage, data quality, accessibility, and implications for research evaluation. In the following, we present recent empirical findings and meta-analytical insights to inform researchers, institutions, and policymakers about the practical implications of selecting a bibliographic database.

Along the coverage dimension, OA indexes a substantially larger number of active open-access journals (34,217) compared to Scopus (7351) and WoS (6157), with 24,976 OA journals unique to OA (Maddi et al., 2025). Scopus generally has broader coverage than WoS in social sciences, humanities, and engineering, while both have more comprehensive coverage in medicine and natural sciences. WoS is more selective, favoring base research disciplines and core journals, often at the expense of applied research and regional diversity (Maddi et al., 2025).

All three databases exhibit structural biases against non-Western, non-English, and arts/humanities research, limiting their global representativeness, particularly in some emerging countries and Africa (Asubiaro & Onaolapo, 2023). This imbalance is also reflected in European publishing systems, where English strongly dominates academic communication, though national languages remain crucial for knowledge dissemination and public engagement (Pölönen & Kulczycki, 2025). Language bias is prevalent in Scopus and WoS, which disproportionately index English-language publications, further marginalizing non-English and regional journals (Céspedes et al., 2025).

OA achieves reference coverage on par with both WoS and Scopus, while offering broader integration of ORCID identifiers. However, it offers fewer abstracts, and its metadata quality is still in development (Culbert et al., 2025). In June 2024, issues emerged when fabricated citations entered via Crossref were also reflected in OA (Besançon et al., 2024).

All three databases exhibit stable growth in publications and citations over time; however, coverage and metrics vary across disciplines, which impacts the evaluation of longitudinal research (Khanna et al., 2022). OA is rapidly evolving, with improvements in coverage but persistent metadata quality issues, particularly in earlier years and in the social sciences/humanities (Aria et al., 2024).

Along the metadata quality dimension, OA suffers from severe metadata completeness issues, with over 60% of journal articles missing institutional information, especially in the social sciences and humanities. Scopus and WoS have more complete institutional metadata but face their own challenges, such as duplicate author profiles and inconsistent document type classification (Zhang et al., 2024).

WoS is plagued by missing and incorrect references, particularly for non-English sources and recent publications, resulting in an underestimation of citation impact. Scopus faces issues with duplicate publications and author profiles, which can inflate citation counts and distort researcher evaluation (Scheidsteger et al., 2025).

According to Thelwall and Jiang (2025), OA provides better citation counts than Scopus; however, its metadata inconsistencies (e.g., missing abstracts, misclassification)

necessitate careful validation. Document type inconsistencies are typical among the three databases, which can affect publication activity analyses if not cross-verified with original publisher data (Haupka et al., 2024; Visser et al., 2021).

In terms of accessibility and usability, OA is fully open access, enabling reproducible and transparent bibliometric studies without licensing barriers. Scopus and WoS are subscription-based, with robust interfaces and advanced search functionalities, but access is restricted to paying institutions. Unlike Scopus and WoS, OA's open data model significantly enhances reproducibility and transparency, enabling community verification and facilitating tool development, such as openalexR (Aria et al., 2024). Scopus offers more comprehensive related search functions (by references, authors, keywords) than WoS, which only provides related search by references (Char & Ajiferuke, 2011).

In terms of implications for research assessment, citation impact scores vary significantly across databases, with Scopus generally providing higher counts than WoS, especially in social sciences, humanities, and engineering (Stahlschmidt & Stephen, 2022). OA offers higher citation counts and is suitable for most fields; however, differences at the university and institutional levels suggest caution in cross-database comparisons. A study comparing OA with WoS, Scopus, and Dimensions on field-normalized citation counts found strong agreement at the paper level but low agreement for university-level average scores, indicating OA is a viable free alternative for individual publication analysis but should not be used to compare universities using aggregated data (Scheidsteger et al., 2025).

Database-specific structural biases (favoring Western, English-language, and STEM research) persist in Scopus and WoS, limiting global representativeness and disadvantaging non-Western and humanities research. OA improves inclusivity but still underrepresents certain regions (e.g., Africa, some emerging countries) and suffers from metadata gaps that can distort evaluation outcomes for non-Western institutions. Evaluation outcomes are affected by these biases, with non-Western institutions facing cumulative disadvantages in research visibility.

The choice of a bibliographic database has profound implications for research evaluation, funding, and institutional assessment. While OA democratizes access and enhances inclusivity, its metadata quality issues must be addressed to avoid distorted outcomes. Scopus and WoS, despite their established roles, perpetuate structural biases that hinder global research equity and the assessment of humanities research.

Cost-free OA offers a valuable alternative for those in resource-limited settings or those seeking to evaluate open-access and regional journals. Scopus and WoS remain essential for high-stakes evaluation in contexts where data curation and established metrics are prioritized, but users must be aware of their biases and coverage limitations. Cross-database comparisons require careful normalization and validation due to divergent coverage and data quality.

Awareness of database-specific biases is critical to ensure fair and equitable research evaluation and funding allocation. Supporting the development of hybrid or integrative approaches that combine open access coverage with rigorous data curation can enhance the reliability and inclusivity of research assessment.

This study aims to deepen the understanding of the differences among OA, Scopus, and WoS from a research assessment perspective. We therefore measure research performance at both the individual and aggregate levels across the three data sources. Our focus is on

scholars in "Management Engineering"[1]—a field in which we are directly involved. Leveraging our disciplinary expertise, we examine cases that exhibit the most significant shifts in performance rankings to uncover their underlying causes. Through this analysis, we aim to identify the strengths and limitations of OA in comparison to the other two databases, thereby contributing to enhanced data curation practices and a more informed selection of data sources for research evaluation.

Methods

From a research assessment perspective, broader publication coverage should, in principle, yield more refined citation analyses. However, coverage alone is not sufficient. Reliable evaluation also depends on accurate author's name and affiliation disambiguation (to ensure that publications are attributed to the true authors, and the latter to the true institution), accurate document-type classification (to ensure like is compared with like), robust subject categorization (so benchmarking occurs within the same research discipline), and precise, comprehensive citation extraction, a technically demanding task. Therefore, OA's more exhaustive coverage does not automatically translate into greater usefulness for research evaluation, and its effectiveness still requires verification (Thelwall & Jiang, 2025).

Data and indicators

The proposed analysis focuses on Italian professors who conduct research in "Management Engineering", like the three authors of this study.

Data were extracted from the database of Italian university personnel, kept by the Ministry of Universities and Research (MUR).[2] This source contains, as of 31/12 of each year, information on each academic, including affiliation, academic rank, and field of research. Italian academics are classified in one, and only one, discipline called Scientific Disciplinary Sector (SDS), for a total of 370 SDSs, which are grouped into fields (called University Disciplinary Areas (UDAs), for a total of 14 UDAs. "Management Engineering" is one of the 42 SDSs of the UDA "Information and Industrial Engineering". Currently, around 300 professors work in this discipline. Our dataset includes 223 professors with at least three years of service during the observation period, i.e., 2019–2023. About two-thirds are male. In terms of academic rank, there are 108 full professors, 87 associate professors, and 28 assistant professors. They work in 37 different universities, with significant variability across universities: in particular, there are 56 (25.1%) at the Polytechnic of Milan, followed by the Polytechnic of Bari, and the University of Padua (both with 12 professors). Nine universities employ only one professor in this discipline.

For each of the 223 professors in the dataset, the census of scientific publications was conducted using:

[1] In Italy, the discipline "Management Engineering" (Ingegneria Economico-Gestionale) extends beyond the traditional boundaries of "Industrial Engineering", blending engineering principles with managerial domains such as operations, logistics, marketing, finance, strategy, innovation, etc.

[2] http://cercauniversita.cineca.it/php5/docenti/cerca.php, retrieved 17 March 2026.

- The author IDs available in both Scopus and OA, manually identified and validated.
- An author name disambiguation (AND) algorithm developed for WoS by D'Angelo et al. (2011).[3]

After an initial step of separately retrieving authorship data for 2019–2023 from each database, the data was cross-referenced to identify and correct any false negatives and false positives.

Only certain document types were considered eligible for evaluation:

- For WoS: article, review, proceedings paper, letter,[4] book review, book chapter.
- For Scopus: article, review, conference paper, letter, book chapter, data paper. "Notes" and "short surveys" were also considered eligible if they referred to publications that were tagged as one of the six types above in WoS.
- For OA: articles, reviews, letters, and book chapters. All other document types were excluded unless they were referred to in publications tagged as one of the six types above in WoS.

Overall, the final bibliometric dataset consists of 3634 WoS authorships, 4077 Scopus authorships, and 5299 OA authorships.

The comparative analysis is based on two indicators. The first performance indicator is the total fractional impact (TFI), at the individual level, defined as:

$$TFI = \sum_{i=1}^{N} c_i \mathrm{f_i} \tag{1}$$

where c_i = year/category normalized citations received by publication i at 31 March 2025; f_i = fractional contribution of the professor to publication i, i.e., the inverse of the number of co-authors in the byline of publication i; N = total number of eligible publications authored by each professor in 2019–2023.

A more accurate indicator of a publication's future impact is not citations alone but a weighted combination of year- and field-normalized citations and a journal's impact factor (Abramo et al., 2019). However, because OA metadata does not include journal impact factors, we cannot apply this indicator in the present comparative analysis.[5]

The second indicator, the simple publication count, constitutes a component of the TFI. It represents the total number of eligible publications (N) produced by each professor during the observation period and is frequently referred to in the literature as a measure of research productivity.[6]

At the meso level, institutional performance for both indicators is derived by averaging the indicator values observed for all professors belonging to the respective institution.

[3] The harmonic average of precision and recall (F-measure) of authorships, as disambiguated by the algorithm, is around 97% (2% margin of error, 98% confidence interval).

[4] While some "research letters" clearly function as full research outputs, other letters may be editorial or commentary-type publications, and their treatment can vary across databases. We have not distinguished between letter types.

[5] Actually, an even stronger predictor of future citation impact should also account for broader, non-scientific factors (Abramo et al., 2025).

[6] In our view, N is not a meaningful performance indicator, as aggregating publications of differing value or impact lacks interpretive validity (Abramo & D'Angelo, 2014).

The comparative analyses will be carried out considering both the distribution of indicator scores and the resulting rankings, in particular the relative percentiles, obtained by considering the top-ranked professor/institution as 100, the bottom-ranked as 0.

Normalized citations: differences across databases

The value of c_i in [1] is calculated according to the standard specifications of each bibliometric database. In WoS, c_i corresponds to the Category Normalised Citation Impact (CNCI), a metric provided by Clarivate Analytics, that compares the actual citations received by a publication with the expected citation rate for documents of the same subject category, publication year, and document type. In Scopus and OA, c_i is represented by the Field-Weighted Citation Impact (FWCI), a similar indicator, though the field classification systems differ substantially.

In WoS, the classification scheme comprises 254 subject categories, with each publication inheriting the category of its hosting source (typically a journal). Scopus follows the same principle, but its scheme includes 333 "ASJC" categories. In contrast, OA applies classifiers at the publication level rather than the journal level. Each publication is assigned to a single "topic," determined from the text of the publication itself rather than the journal's field. These 4,516 topics are grouped into 252 subfields, and in OA, the normalisation of citations within the FWCI is performed at the subfield level.

Another important consideration concerns the year of publication. OA indexes publications based on the online first date, i.e., the year associated with DOI registration, whereas Scopus and WoS generally use the source year, which refers to the year of the journal issue in which the article is formally published (often corresponding to the print version).

Results

The results of our comparative analysis reveal a nuanced picture of how the three bibliographic databases capture the research performance of Italian professors in Management Engineering. Below, we will analyse, both at the individual and university levels, the differences between the distributions of the two indicators obtained using data extracted by the three databases. We will then analyse the impact of these differences on the resulting rankings.

Descriptive analysis of indicator score distributions at the individual level

The patterns visible in the figures highlight both substantial overlaps and meaningful divergences across sources, which have important implications for research evaluation.

Beginning with Fig. 1, the scatter matrix of output indicators provides a visual summary of pairwise comparisons among the three databases. The diagonal plots, showing density estimates of the distribution of outputs within each source, indicate a broadly right-skewed pattern: a majority of academics publish relatively modest numbers of works, while a minority account for very high output levels. This skewness is consistent with Lotka's law and general productivity distributions in science (Lotka, 1926). Off-diagonal scatterplots reveal strong linear relationships between sources (Pearson WoS vs Scopus correlation: 0.95; WoS vs OA: 0.91; Scopus vs OA: 0.92). In particular, the clustering along the diagonals demonstrates that WoS and Scopus record highly similar outputs for most professors,

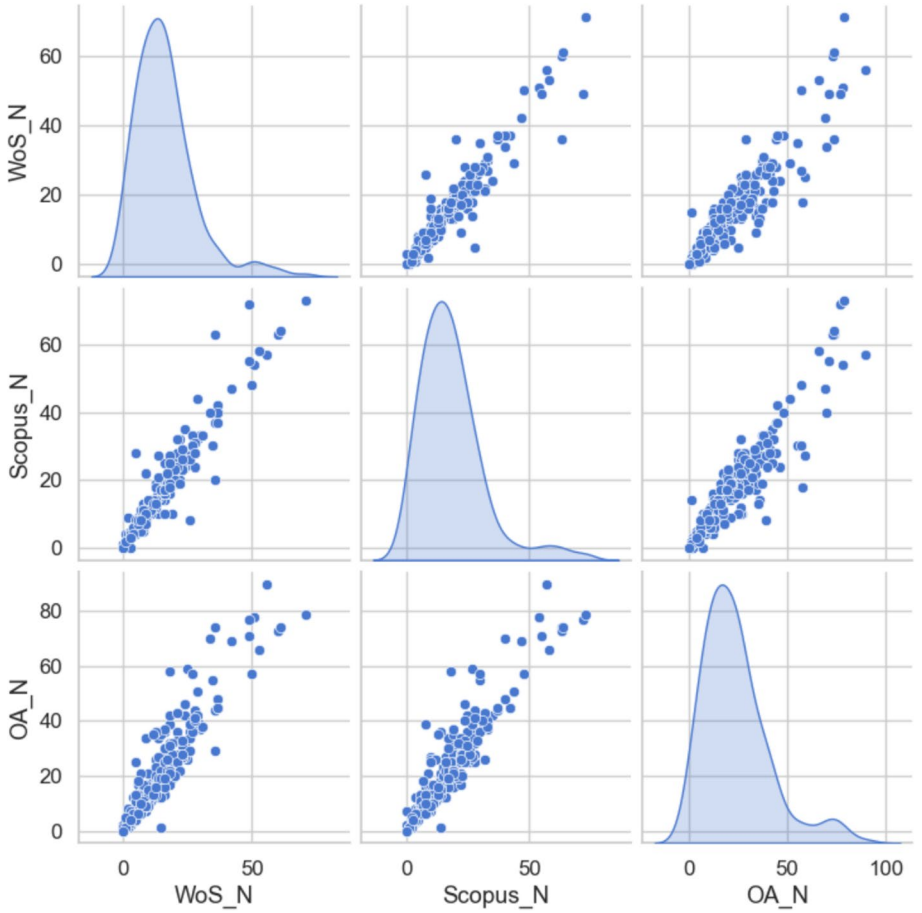


Fig. 1 Scatter matrix showing the N distributions recorded in the three databases, for the 223 professors in the dataset

while comparisons involving OA show slightly greater dispersion. This suggests that OA may capture additional publications not indexed in WoS or Scopus, leading to divergence, especially for prolific authors.

A similar dynamic is evident in Fig. 2, which reports scatter matrices for the TFI indicator. Once again, diagonal density plots show that most professors accumulate modest impact scores, with only a small proportion achieving very high normalized citation counts. The pairwise scatterplots confirm strong but not perfect concordance between sources. Compared to the output distributions in Fig. 1, the TFI results in Fig. 2 exhibit greater dispersion across sources, reflecting the additional complexity introduced by citation normalization. Particularly in comparisons involving OA, outliers become more pronounced, suggesting differences in subject classification schemes or citation coverage across databases. While correlations remain high (Pearson WoS vs. Scopus correlation: 0.94; WoS vs. OA: 0.89; Scopus vs. OA: 0.90), these discrepancies underscore the sensitivity of impact-based metrics to database-specific metadata and coverage choices.

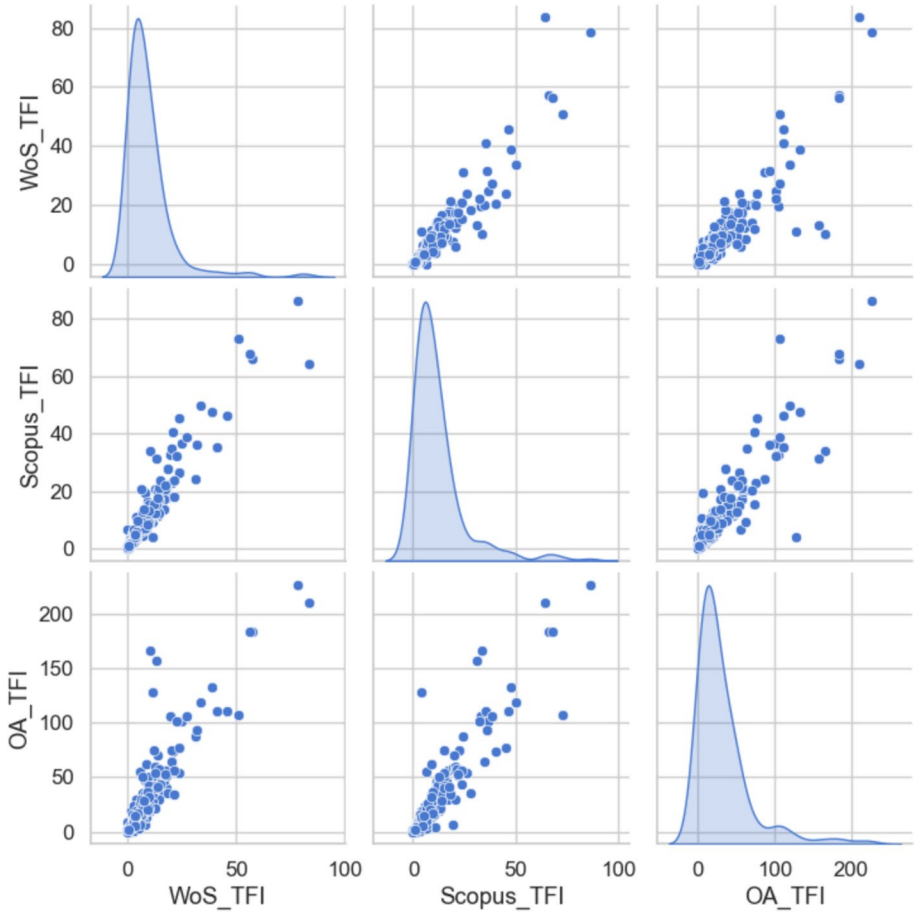


Fig. 2 Scatter matrix showing the TFI distributions recorded in the three databases compared, for the 223 professors in the dataset

Figures 3 and 4 provide violin plots that directly compare the distributional shapes of *N* and TFI scores across the three sources. In Fig. 3, the violin plots for N show striking similarities in median values and interquartile ranges across WoS and Scopus, while OA displays a somewhat broader distribution. The heavier tails in OA suggest that specific individuals appear with much higher N values relative to the other two databases, likely due to its more exhaustive coverage of open-access and non-traditional journals and document types. This pattern illustrates how OA's more inclusive indexing policy can inflate N measures for some professors, potentially altering their comparative ranking.

Turning to Fig. 4, which compares TFI distributions, the contrasts become sharper. Here again, WoS and Scopus present relatively similar distributional shapes, whereas OA exhibits heavier upper tails. This suggests that for specific individuals, OA records substantially higher citation-normalized fractional impact than the commercial databases. Possible explanations include broader citation linkages in OA, its text-based classification system that may place works in higher-citation subfields, or the inclusion of works not covered elsewhere. Moreover, OA contains a substantially higher share of uncited publications,

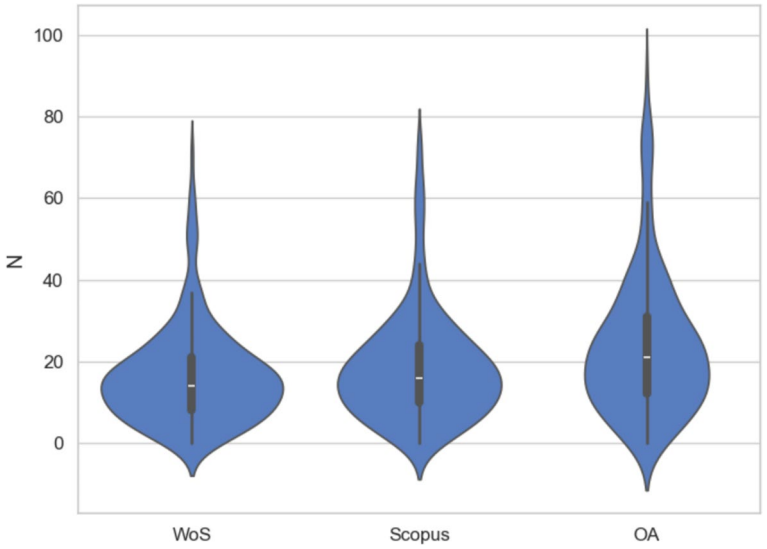


Fig. 3 Violin plots showing the N distributions recorded in the three databases compared, for the 223 professors in the dataset

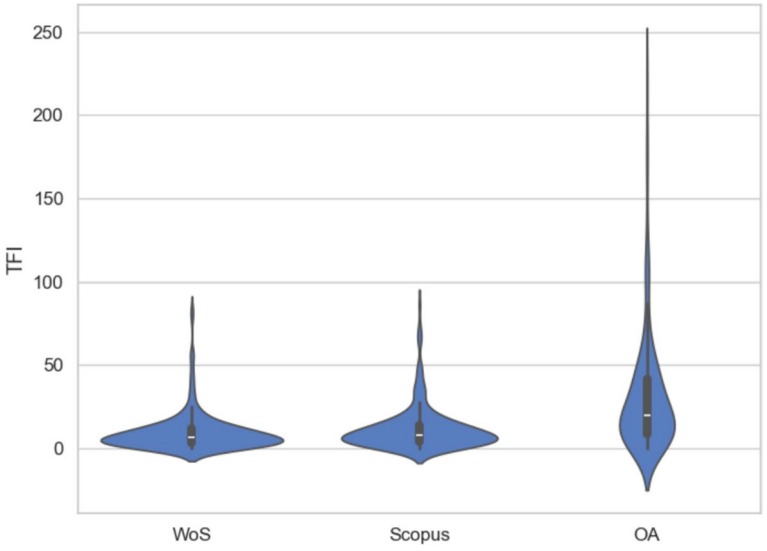


Fig. 4 Violin plot of the TFI distributions recorded in the three databases compared, for the 223 professors in the dataset

which reduces the mean citation rate of the field and consequently inflates normalized indicators. This raises concerns about comparability, as rankings and assessments based on OA data may systematically favor particular profiles of professors while disfavoring others. It is also worth noting the significant difference in the shape of the broadest part of the violin, with WoS and Scopus showing a higher density of values around the distribution median,

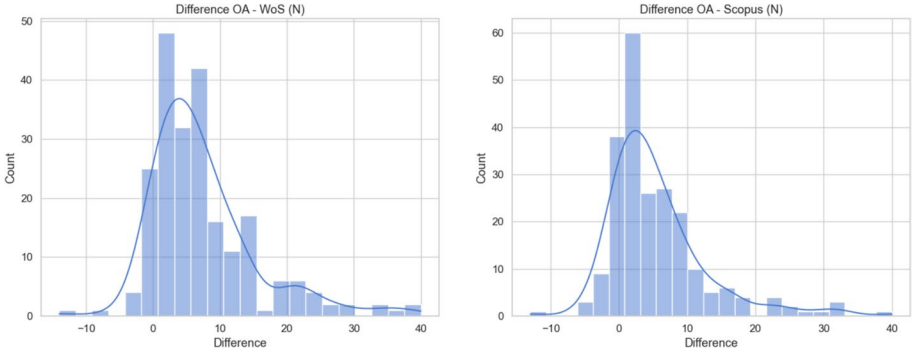


Fig. 5 Frequency of N differences recorded for pairs of databases (OA vs. WoS on the left; OA vs. Scopus on the right), for the 223 professors in the dataset

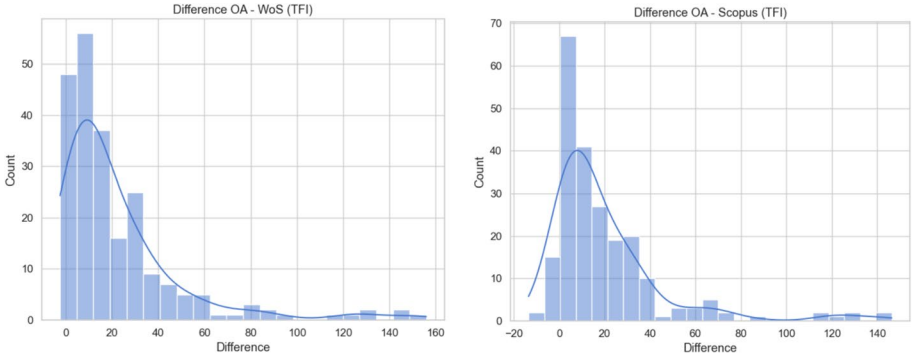


Fig. 6 Frequency of TFI differences recorded for pairs of databases (OA vs. WoS on the left; OA vs. Scopus on the right), for the 223 professors in the dataset

compared to OA, which shows greater dispersion. This becomes very clear when considering TFI rather than simple N. Even slight differences in classification or citation coverage can propagate into larger discrepancies in normalized impact indicators, which explains the somewhat weaker alignment of WoS/Scopus on one side with OA on the other.

Figures 5 and 6 explore these divergences further by presenting histograms of pairwise differences in N (Fig. 5) and TFI (Fig. 6). In Fig. 5, the frequency distributions of N differences reveal that for the majority of professors, differences between databases are minor, clustering closely around zero. However, the histograms also show non-negligible right tails, indicating a minority of academics for whom OA records substantially higher outputs than WoS or Scopus. These individuals may benefit from OA's broader coverage of conference proceedings, regional journals, or open-access outlets not indexed elsewhere. However, it is also interesting to note the presence of a few cases with negative differences, indicating the possibility that for some professors, OA indexes fewer works than commercial data sources.

In Fig. 6, the histograms of TFI differences show longer right tails compared to the output histograms. The left panel, plotting the differences between OA and WoS, does not indicate the presence of negative values, unlike the right panel, which suggests that in

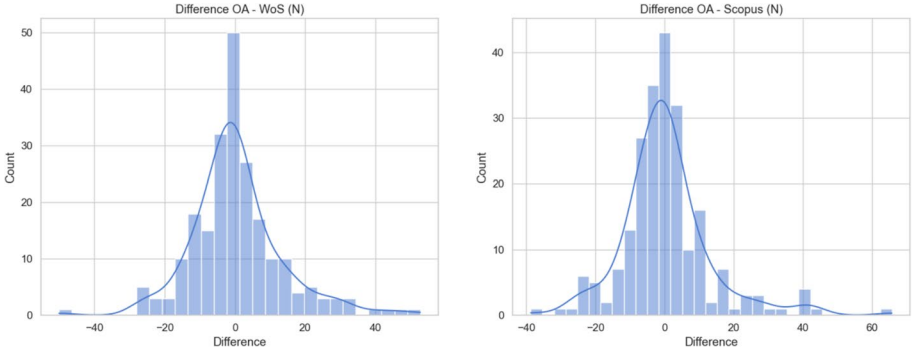


Fig. 7 Frequency of N percentile differences recorded for pairs of databases (OA vs. WoS on the left; OA vs. Scopus on the right), for the 223 professors in the dataset

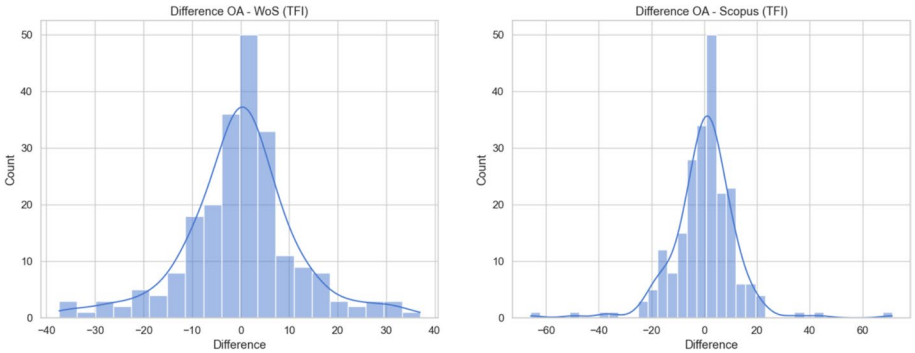


Fig. 8 Frequency of TFI percentile differences recorded for pairs of databases (OA vs. WoS on the left; OA vs. Scopus on the right), for the 223 professors in the dataset

Scopus, the TFI value is higher than that recorded in OA for a non-negligible number of professors.

The intraclass correlation coefficient (ICC) analysis based on agreement (Shrout & Fleiss, 1979) reveals a high level of concordance among WoS, Scopus, and OA for the N indicator (ICC = 0.83, 95% CI 0.64–0.91), indicating that these databases generally produce comparable output estimates for individual professors. In contrast, agreement for the TFI indicator is moderate (ICC = 0.45, 95% CI 0.22–0.62), suggesting that impact-based metrics are more sensitive to variations across databases. These findings underscore that the choice of bibliographic source can substantially affect research evaluations, particularly those focused on scholarly impact.

Ranking variations at the individual level across databases

Overall, the findings from Figs. 1, 2, 3, 4, 5, 6 underscore two important points: (i) the distributions of the two indicators measured through the three compared databases appear to be significantly correlated, but (ii) they are also significantly different, especially concerning the right tails. Therefore, after analysing the distributions of the indicator scores, it is certainly interesting to analyse the differences in the resulting rankings. In this regard,

Table 1 Descriptive statistics of percentile variations (for N and TFI) recorded for pairs of databases, for the 223 professors in the dataset

	Output		TFI	
	OA vs WoS	OA vs Scopus	OA vs WoS	OA vs Scopus
Min	− 50.0	− 38.7	− 37.4	− 65.8
Max	52.7	65.7	37.0	71.2
Mean*	9.0	8.6	8.1	8.2
Median*	5.9	5.4	5.8	5.4
St. dev.*	9.5	9.7	8.4	9.5

*Computed considering absolute values of percentile variations

Table 2 Contingency matrices of TFI quartiles for pairs of databases, for the 223 professors in the dataset

		OA			
		Q1	Q2	Q3	Q4
WoS	Q1	45	11		
	Q2	10	36	9	1
	Q3	1	9	34	11
	Q4			12	44

		OA			
		Q1	Q2	Q3	Q4
Scopus	Q1	41	14		1
	Q2	13	33	9	1
	Q3	2	9	36	8
	Q4			10	46

similarly to Figs. 5, 6, 7 and 8 show, respectively, histograms of pairwise differences in N and TFI percentiles. In Fig. 7, the distribution of N percentile differences illustrates that the majority of professors experience relatively modest shifts between databases, clustering around zero. However, the tails extend significantly, with a non-trivial number of individuals moving by more than ± 30 or even ± 40 percentile points.

The two plots appear very similar, even if when comparing OA with Scopus, the distribution of ranking variations shows a longer right tail. Figure 8, focusing on TFI, reveals a similar pattern, with a substantial subset of professors undergoing very large shifts, in some cases over 60 percentile points, when contrasting OA with Scopus.

The descriptive statistics presented in Table 1 quantify these patterns. Minimum and maximum values confirm the potential magnitude of ranking volatility, with differences ranging from − 65.8 to + 71.2 percentile points for TFI and nearly as high for N. These are extreme cases, but the means and medians, calculated on absolute values, reveal that even in ordinary cases the average shift is around 8–9 percentile points. This is not trivial, as a change of this order can easily move a professors across evaluation thresholds, such as quartiles or deciles, which are often used for assessment and resource allocation. The standard deviations further underscore a high variability, confirming that while most individuals remain relatively stable, database choice can have profound consequences for a non-negligible fraction of the population.

Table 2 presents the contingency matrices of TFI quartiles across databases, illustrating these effects categorically. The comparison between WoS and OA reveals that while 159 professors remain consistently in the same TFI quartile in both databases, 64 experience a shift of at least one quartile, with 32 moving up and 32 moving down. In particular, one professor (John Doe), who ranks in the second quartile in WoS, ends up in the fourth

quartile in OA. At the same time, Jane Doe, who ranks in the third quartile in WoS, rises to the first quartile in OA.

A similar degree of volatility is evident in the comparison between Scopus and OA, with 154 professors placed in the same TFI quartile in both circumstances. Additionally, 34 professors improved their ranking when evaluated using OA data compared to Scopus data, while 33 professors worsened it. In particular, the aforementioned John Doe ranks in the top quartile in Scopus, while, as already mentioned, he is in the bottom quartile with OA data.

Such cross-tabulations make clear that quartile-based evaluations, widely used in institutional contexts, are highly sensitive to database selection. For institutions or funding agencies that rely on quartiles as benchmarks, these discrepancies raise critical questions of fairness and reproducibility.

Ranking variations at the institutional level

We now aim to conduct a comparative analysis at the meso level, i.e., by aggregating data at the university level, to understand how institutional rankings vary depending on the database considered. In particular, we will consider the performance of a university as the average performance of its affiliated professors.

In our case, we will aggregate the data starting from the affiliations indicated in the official MUR list. For an evaluator who does not have access to this list, the aggregation must necessarily be based on the affiliations indicated in the bibliometric database used. In this regard, in Scopus, the "affiliations" of 5 of the 223 professors in the dataset (equal to 2.2%) are different from those of the MUR. In OA, this occurs in 5.4% of cases. A separate discussion applies to WoS, which uses a unique identifier (the researcher_ID) that is not available in all publications, nor for all authors of a publication (which is why we applied the AND developed by the authors to the WoS data). With reference to the dataset constructed in this work, in 9.8% of publications, there is no WoS researcher_ID associated with authors in the byline. In the remaining 90.2%, researcher IDs map on average to less than 60% of the total authors in the byline.

As mentioned, in our case, we will aggregate performance based on the official affiliations of professors in the dataset, regardless of how they may have signed their publications.

For this purpose, Fig. 9 shows the dispersion of rankings resulting from the "N" indicator, for universities with at least two professors in the dataset. The comparison among the three rankings clearly shows substantial variability in university performance depending on the bibliographic source. While some institutions, such as Napoli "Parthenope" and Mediterranea "J. Monnet", consistently appear at the top, others exhibit notable rank shifts — for example, Modena e Reggio Emilia moves from the bottom positions in WoS and Scopus to seventh place in OA.

The plot in Fig. 10, which refers to the TFI indicator, confirms that institutional rankings vary depending on the database used. While some universities cluster closely across the three databases, others exhibit pronounced divergence. Institutions employing only a few professors in the discipline are particularly susceptible to ranking volatility, as the exceptional performance of one or two individuals can substantially alter the institutional average when databases differ in publication and citation coverage. For example, the University of Parma, with three professors, ranks 3rd in Scopus, 10th in WoS, and 13th in OA. Similarly, the University of Rome "Tor Vergata", with six evaluated professors, shows considerable variability—ranking 11th in Scopus, 13th in WoS, and 20th in OA. By contrast,

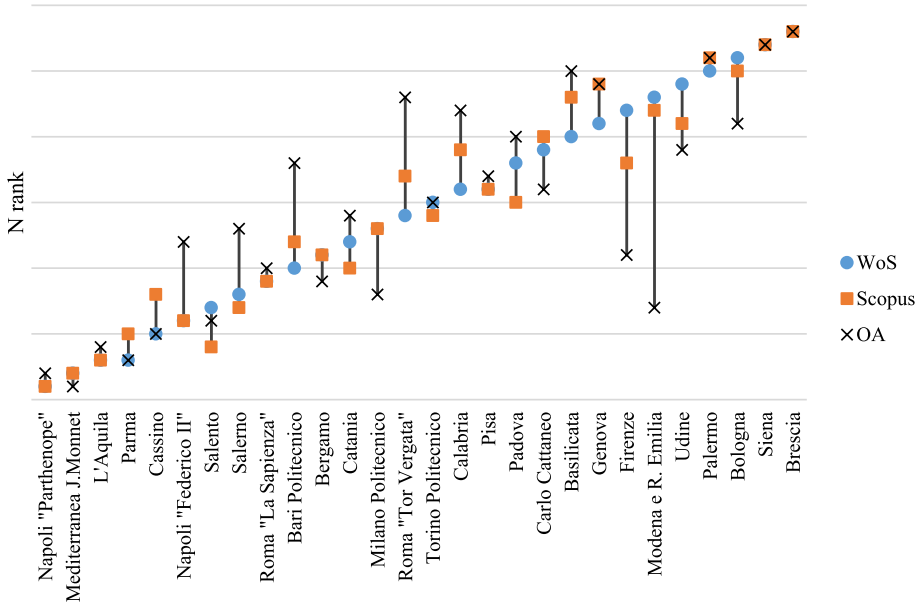


Fig. 9 Dispersion of N rankings of 28 Italian universities with at least two evaluated professors in the "Management Engineering" discipline. When the WoS and Scopus ranks coincide, one of the two labels is omitted from the graph due to overlap

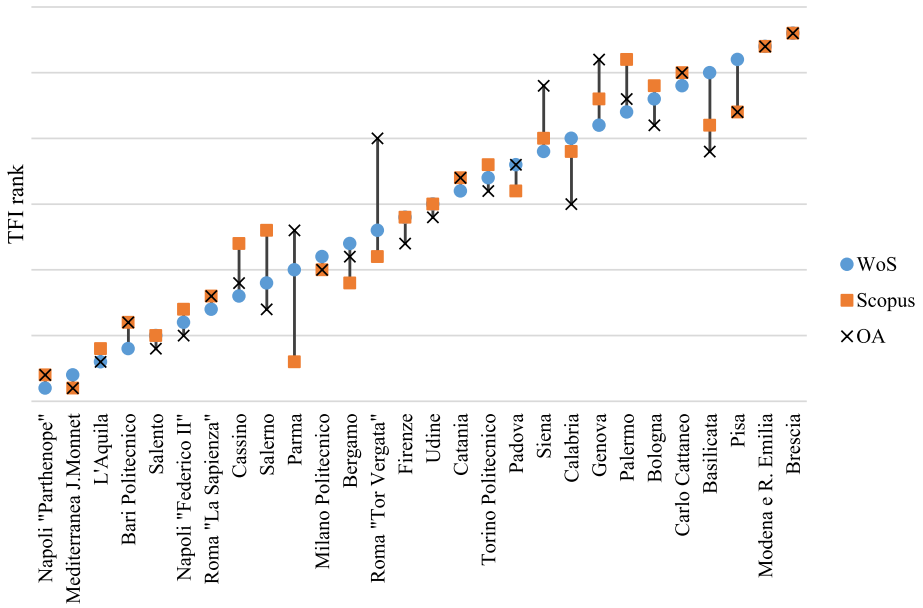


Fig. 10 Dispersion of TFI rankings of 28 Italian universities with at least two evaluated professors in the "Management Engineering" discipline. N.B. When WoS and Scopus ranks coincide; the relative labels do not appear in the graph since they overlap

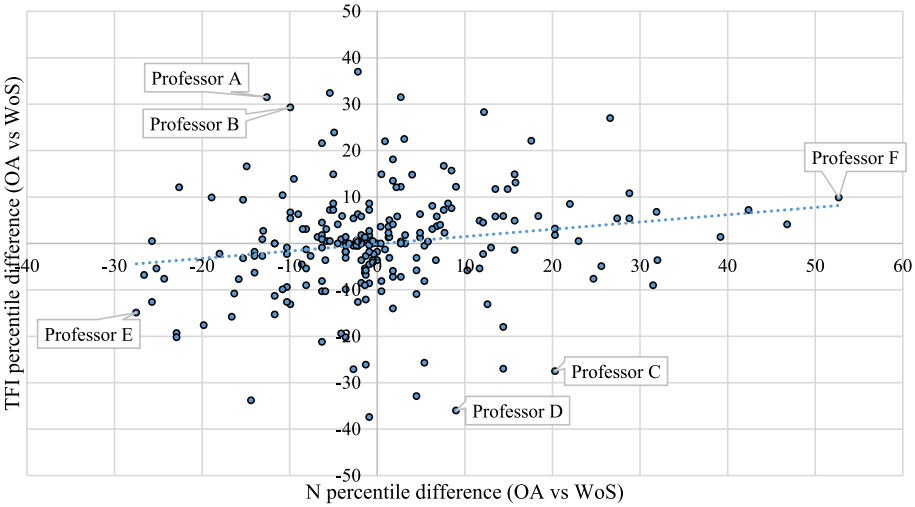


Fig. 11 Scatterplot of percentile differences between OA and WoS recorded for the two indicators (N vs. TFI) for the 223 professors in the dataset

larger universities (those employing ten or more professors) display smaller rank fluctuations, with minimum–maximum differences of up to three positions—still notable given the total of 28 institutions considered.

These discrepancies emphasize that bibliometric evaluations are highly dataset-dependent, also at the meso level and should be interpreted with caution when comparing institutional research performance.

Determinants of ranking variations

We now turn to the individual-level analysis to examine the factors underlying differences in indicator scores and the resulting rankings. We begin by asking whether the variations observed for the two indicators are correlated. In fact, one could assume that part of the variability recorded for TFI is due to the variability of N, which is a fundamental component of TFI. To this end, Fig. 11 plots the N percentile differences against TFI percentile differences for the 223 professors in the dataset, obtained using OA instead of WoS as the data source.[7] The Pearson correlation coefficient between the two distributions is negligible (rho = 0.17); the scatterplot shows a broad dispersion across all four quadrants. This means that for some individuals OA simultaneously records higher/lower N and higher/lower TFI, while for others it records higher N but lower TFI, or vice versa. The lack of systematic alignment indicates that discrepancies between databases are not uniform but depend on individual publication and citation profiles. The presence of points at extreme values, with professors experiencing simultaneous significant shifts in both indicators, illustrates that some careers can appear radically different depending on which database is consulted. At the same time, the presence of numerous points in the second and fourth

[7] We conducted the same analysis (omitted here for reasons of space) contrasting Scopus and OA, obtaining exactly the same evidence.

quadrants suggests that any advantage or disadvantage associated with the use of OA in terms of N does not necessarily correspond to a similar effect in terms of TFI. Accordingly, we analyse two cases that are particularly illustrative of this phenomenon:

- Professor A and Professor B, who, in the transition from WoS to OA, would record a, significant positive jump in rank for TFI (+31 and +29 percentiles, respectively), but a worse position in terms of N (− 13 and − 10);
- Professor C and Professor D, who, on the contrary, would see their TFI ranking worsen in the transition from WoS to OA (− 28 and − 36 percentiles, respectively), but their N ranks improve (+20 and +10).

The first two, transitioning from WoS to OA, experience a minimal increase in N (+18/22%), which, in comparative terms, causes them to fall in the national ranking, as the average increase in N for all 223 professors in the dataset is 46%.

The opposite is true for Professor C and Professor D, who have 90 and 300% higher N in OA than in WoS. The situation is different for TFI. In this case, two factors are at play: for the first two subjects, the increase in TFI is primarily due to the higher value of c_i for publications indexed in both databases. For example, for Professor A, the publication with https://doi.org/10.1080/14778238.2020.1851615 has a WoS CNCI of 0.8 and an OA FWCI of 11.7. The same applies to Professor B, who authored the publication with https://doi.org/10.3390/su12229575, which has a WoS CNCI of 1.9 and an OA FWCI of 25.6. Same publication, with an impact 14 times higher in OA than in WoS.

For Professor C and Professor D, the situation is different: the TFI measured in WoS for Professor C is strongly influenced by the publication with https://doi.org/10.1007/s11365-018-0526-3, which has a CNCI of 15; however, it is not present in the OA dataset because its DOI version was indexed in 2018. The same applies to Professor D, whose best publication (https://doi.org/10.1016/j.jfbs.2017.12.001) in WoS was indexed in 2019 with a CNCI of 7, whereas in OA it was indexed in 2018 and therefore does not contribute to his TFI measured with this database.

To complete the analysis, it is worth examining the cases of Professors E and F, characterized by the largest negative and positive changes in N percentiles, respectively. The former loses almost 28 percentile points when moving from WoS to OA, as he is author of 15 publications in the first source and 12 in the second, with the remaining three all indexed in OA for 2018. The latter, on the other hand, gains over 50 percentile points in terms of N, being listed as the author of 9 publications in WoS and 34 in OA. The difference between the two counts consists of 11 book chapters and 14 articles. Among the latter, in terms of venues, 7 appear in proceedings collections, 6 in institutional databases, and only one is published in a journal. In this regard, across the whole dataset, there are 1,350 publications indexed in OA but not in WoS overall, of which 967 have a venue clearly recorded in OA. Half of these are hosted in just 20 sources, the most recurring of which are:

- Academy of Management Proceedings: 88 publications.
- Virtual Community of Pathological Anatomy: 80 publications.
- SSRN Electronic Journal: 70 publications.

This analysis, therefore, shows that the variations are mainly attributable to:

- For N, the greater coverage of OA, compared to WoS,

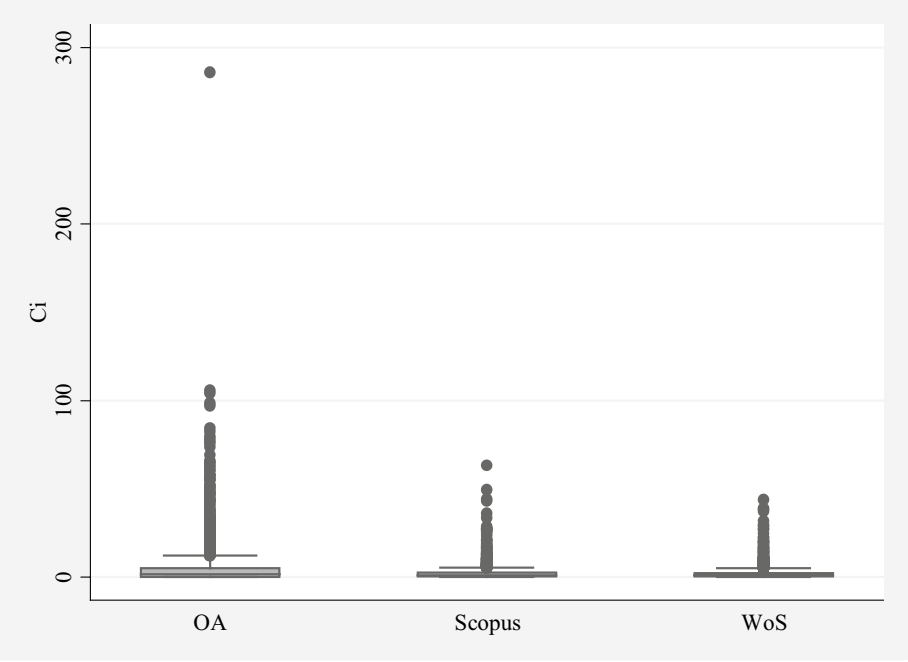


Fig. 12 Boxplot of normalized citations (ci) of publications indexed in all three databases

- For TFI, the different scaling of citations, which in some cases inflates the impact measured in OA, compared to WoS.

In any case, differences in the indexing year also produce significant variations in both N and TFI. They, therefore, can substantially affect the comparative evaluation of an author's bibliometric performance over time. In this regard, the 26 publications included in the WoS census but not in the OA census, with a CNCI greater than 5, were all indexed in OA in 2018 and therefore absent from the TFI calculation in that database because they fell outside the observation period.

We would like to further explore the issue of inflation in the impact of OA publications. In this regard, Fig. 12 shows the box plot of c_i registered for publications simultaneously indexed in all three databases.[8] It is worth noting the longer right tail of the OA distribution, with one point completely off the scale. This is the publication with https://doi.org/10.1080/09647775.2020.1790029, credited with 142 citations received in WoS, 174 in Scopus, and 216 in OA.

However, while the WoS CNCI is 9 and the FWCI in Scopus is 33.7, the OA FWCI is 286.0. This disproportion is due to the peculiarities of discipline labels associated with this publication:

- "Humanities, Multidisciplinary" in WoS.

[8] For reasons of greater robustness, this analysis is restricted to the 1957 publications indexed in all three repositories (i.e., sharing the same DOI).

Table 3 Descriptive statistics of c_i for publications indexed in all three databases

	WoS	Scopus	OA
Median	1.073	1.180	1.682
Q3	2.339	2.500	4.981
IQR	1.893	2.020	4.895
Max	44.1	63.4	286.0
Q3 + 1.5 IQR	5.2	5.5	12.3
Q3 + 3 IQR	8.0	8.6	19.7
Outliers 1.5	8.6%	7.8%	8.8%
Outliers 3	3.7%	3.6%	4.2%

- "Business and International Management" in Scopus.
- "Museology" in OA.

While at first glance this third label is certainly more specific than the first two, on the other hand, this "specificity" distorts the scaling: "Museology" articles in OA in 2020 receive, on average, less than 1 citation five years after publication; "Humanities, Multidisciplinary" articles in WoS in 2020 receive almost 16.

The descriptive statistics shown in Table 3 reveal marked differences in normalized citation (c_i) distributions across the three bibliometric databases. OA consistently reports higher central tendency and dispersion measures than both WoS and Scopus, with a median citation value (1.682) roughly 40–60% higher than in the other sources. Similarly, OA's upper quartile (Q3 = 4.981) and interquartile range (IQR = 4.895) indicate a substantially more skewed distribution, suggesting the presence of highly cited items driving the spread. The extreme maximum value (286.0) and higher thresholds for outlier detection (e.g., Q3 + 1.5 IQR = 12.3) reinforce this pattern of heavier tails. While the overall proportion of outliers remains relatively comparable across datasets, OA still shows a slightly greater incidence of extreme values at each threshold, particularly for the most distant points (4.2 vs. 3.6–3.7%, considering Q3 + 3 IQR). These differences could reflect:

- Variation in coverage across data sources. OA greater coverage is associated with a lower average impact of indexed publications, resulting in lower scaling factors, which inflate c_i values of publications in the dataset.
- Different disciplinary labelling of publications, resulting in distortions in impact standardization.

In this regard, we now examine the disciplinary taxonomies employed by the three databases in closer detail. Table 4 shows the top 15 labels for publications in the dataset.[9] A certain overlap exists between the three schemes, as the word "Management" appears in all three labels at the top of the table. However, we also register a higher "concentration" in WoS and a lower one in OA (with Scopus having intermediate values). In particular, the

[9] In OA, the association between a publication and a subject category is one-to-one, whereas in WoS and Scopus it is not. In this analysis, for WoS/Scopus publications in multi-category/ASJC journals, a single label identified as 'prevalent' in terms of the total number of publications in the dataset was assigned.

Table 4 Main disciplinary labels in the three databases, and relative share of publications in the dataset

WoS main label	%	Scopus main label	%	OA main label	%
Management	36.8	Strategy and Management	25.2	Strategy and Management	15.6
Environmental Sciences	10.5	General Business, Management and Accounting	9.4	Management Information Systems	8.1
Business	9.7	Management of Technology and Innovation	9.3	Economics and Econometrics	7.6
Economics	5.7	Renewable Energy, Sustainability and the Environment	6.6	Management of Technology and Innovation	7.6
Engineering, Industrial	5.5	Business and International Management	5.5	Industrial and Manufacturing Engineering	5.7
Information Science & Library Science	4.2	Computer Science Applications	4.1	Marketing	4.8
Multidisciplinary Sciences	3.0	Economics and Econometrics	3.5	Sociology and Political Science	4.3
Education & Educational Research	2.6	Management Science and Operations Research	3.0	Organizational Behavior and Human Resource Management	3.8
Computer Science, Interdisciplinary Applications	2.1	General Engineering	2.9	Management Science and Operations Research	3.1
Green & Sustainable Science & Technology	2.0	Marketing	2.4	Political Science and International Relations	2.8
Business, Finance	1.6	Multidisciplinary	2.3	Accounting	2.5
Health Care Sciences & Services	1.2	Education	1.8	Information Systems	1.9
Computer Science, Information Systems	1.1	Computer Networks and Communications	1.8	Statistics, Probability and Uncertainty	1.6
Energy & Fuels	1.0	Business, Management and Accounting (miscellaneous)	1.7	Education	1.5
Operations Research & Management Science	1.0	General Computer Science	1.6	Building and Construction	1.5

Table 5 Coverage of bibliometric databases, for the aggregated publication portfolio of 223 professors in the dataset

Set	No. of publ. (N)	Set	No. of publ. (D)	Share (N/D)
WoS NOT OA	123	OA	3,706	3.3%
Scopus NOT OA	136	OA	3,706	3.7%
WoS AND OA	2,282	OA	3,706	61.6%
Scopus AND OA	2,563	OA	3,706	69.2%
OA NOT WoS	1,349	WoS	2,405	56.1%
OA NOT Scopus	1,037	Scopus	2,699	38.4%
WoS AND OA	2,282	WoS	2,405	94.9%
Scopus AND OA	2,563	Scopus	2,699	95.0%

first label ("Management") features 36.8% of WoS publications, while "Strategy and Management" features 25.2% of Scopus publications and 15.6% of OA publications.

The top five labels cumulatively account for 68.2% of total publications in WoS, 55.9% in Scopus, and 44.6% in OA. Overall, the publications in the dataset span 89 categories in WoS, 93 ASJC codes in Scopus, and 135 subfields in OA.

Given the expertise of the authors in the field, it seems rather unusual the very high OA ranking of:

- Sociology and Political Science (4.3% of total OA publications).
- Statistics, Probability and Uncertainty (1.6% of total OA publications).

The first of these two subfields has an identical label in Scopus, where the corresponding ASJC code accounts for only 0.1% of total publications. At the same time, the WoS subject category "Sociology" accounts for only 0.1% of total WoS publications in the dataset.

As for the OA subfield "Statistics, Probability and Uncertainty", in WoS and Scopus, the corresponding label should be "Statistics and Probability", which accounts for 0.1% of WoS publications and 0.3% of Scopus publications, respectively.

The trade-off between publication coverage and citation coverage

To better understand the implications of relying on open bibliographic databases for research evaluation, we compared OA with WoS/Scopus for the aggregated publication portfolio of professors in our dataset, along two key dimensions: database publication coverage, which directly affects the number of publications attributed to each professor, and average citations, which influence the assessment of scholarly impact. By examining the overlap and exclusivity of publications across these databases, we aim to highlight how the choice of source can substantially shape bibliometric evaluations.

For this purpose, the data in Table 5 reveal that OA exhibits broader coverage, with 3,706 unique publications (column 4) against 2699 in Scopus and 2405 in WoS.[10]

[10] The figures here differ from total authorships reported in "Data and indicators" section because of multiple counting of publications co-authored by two or more researchers in the dataset.

Table 6 Average citations received by the aggregated publication portfolio of 223 professors in the dataset, by bibliometric source

Set	Citations source	Avg citations
WoS	WoS	32.4
Scopus	Scopus	35.9
OA	OA	28.5
WoS NOT OA	WoS	2.7
WoS AND OA	WoS	34.0
WoS AND OA	OA	48.2
Scopus NOT OA	Scopus	3.0
Scopus AND OA	Scopus	37.6
Scopus AND OA	OA	42.9
OA NOT WoS	OA	3.3
OA NOT Scopus	OA	1.7

The set of publications indexed in both OA and WoS accounts for 61.6% of all OA publications and 94.9% of all WoS publications. Similarly, publications indexed in both OA and Scopus represent 69.2% of all OA publications and 95.0% of all Scopus publications.

Publications indexed in OA but not in WoS (Scopus) represent 56.1% (38.4%) of the total WoS (Scopus) publications, highlighting the additional coverage provided by OA. Conversely, publications indexed in WoS (Scopus) but not in OA account for only 3.3% (3.7%) of total OA publications, indicating that the vast majority of records in traditional databases are also captured by OA.

Table 6 shows the average citation counts for different publication subsets, allowing an assessment of how database coverage and the choice of citation source affect the perceived scholarly impact.[11]

Publications indexed in Scopus tend to receive slightly more citations (35.9) than those in WoS (32.4), while publications in OA show a lower average (28.5). This suggests that traditional databases (WoS and Scopus) capture publications that are, on average, more cited, whereas OA's broader coverage includes additional outputs that tend to attract fewer citations. The difference highlights a potential trade-off in bibliometric evaluation: using OA increases coverage, but may lower average citation metrics due to the inclusion of less-cited works.

This is confirmed by data registered for publications indexed in both OA and commercial databases (WoS or Scopus), which exhibit substantially higher citation averages compared to publications exclusive to OA or to WoS/Scopus alone. For instance, WoS-and-OA publications receive on average 34.0 citations on WoS and 48.2 on OA, while WoS

[11] To ensure maximum consistency, the citation data used in this analysis were re-extracted simultaneously from all three databases approximately nine months after the initial extraction employed for the calculation of the TFI. This re-extraction revealed an unexpected issue affecting OA data: nearly 12% of the original records appeared to be missing. Further inspection showed that these records had not disappeared but were instead associated with different OA identifiers. This issue is primarily due to OA work-based data model, whereby records corresponding to preprints, early-access papers, or other provisional versions may be reassigned to different IDs as they are merged into finalized journal articles.

publications not in OA average only 2.7 citations. Similarly, OA-exclusive publications have a very low citation average (3.3). These results suggest that although OA significantly increases coverage, the additional publications tend to attract fewer citations, indicating a trade-off between broader coverage and perceived impact in bibliometric evaluations.

Robustness analysis on the shared publication set

The robustness of the results was further examined by restricting the analysis to the subset of 217 individuals whose publications are indexed in all three databases (WoS, Scopus, and OA). Within this subset, the comparison was carried out using two alternative specifications: first, considering the full set of publications associated with each database, and second, restricting the analysis to the shared publication set, defined as the subset of publications identified in all three databases through DOI matching. This approach allows us to assess whether the results are sensitive to database-specific coverage. Table 7 reports the correlations between the individual TFI indicators derived from the three databases under the two specifications. The results show a very high level of agreement across databases in both cases, with only marginal differences between the full and shared publication sets. In the full dataset, the correlation between WoS and Scopus reaches 0.940, while the correlations between WoS and OA and between Scopus and OA are 0.885 and 0.900, respectively. When the analysis is restricted to the shared publication set, the correlations remain almost unchanged: the WoS–Scopus correlation increases slightly to 0.945, the WoS–OA correlation remains essentially stable (0.883), and the Scopus–OA correlation increases to 0.948. The very limited variation observed between the two specifications suggests that the relative positioning of authors is largely unaffected by the inclusion of publications indexed only in one database.

Further evidence emerges from the quartile transition matrices reported in Fig. 13, which compare the distribution of authors across quartiles of the TFI indicator derived from WoS and Scopus with those obtained from OA. In the full publication set, the heatmaps show a strong concentration along the main diagonal, indicating that most authors are assigned to the same quartile across databases. A similar pattern is observed when the analysis is restricted to the shared publication set, where the distribution of authors across quartiles remains largely unchanged. The persistence of the diagonal structure across the two specifications indicates that the agreement in quartile classifications is not substantially affected by differences in database coverage, and that the ranking of authors remains broadly stable even when only publications indexed in all three sources are considered.

Finally, we investigated whether discrepancies in publication year attribution across databases could explain part of the differences observed in the individual-level indicators. For each author, we calculated the proportion of shared publications whose publication year differs between OA and WoS, and between OA and Scopus, and examined their relationship with the corresponding differences in TFI values. The results, reported in Fig. 14,

Table 7 Correlation coefficients (full and shared publications sets)

Comparison	Full publication set	Shared publication set
WoS – Scopus	0.940***	0.945***
WoS – OA	0.885***	0.883***
Scopus – OA	0.900***	0.948***

Significance level: *** $p < 0.01$

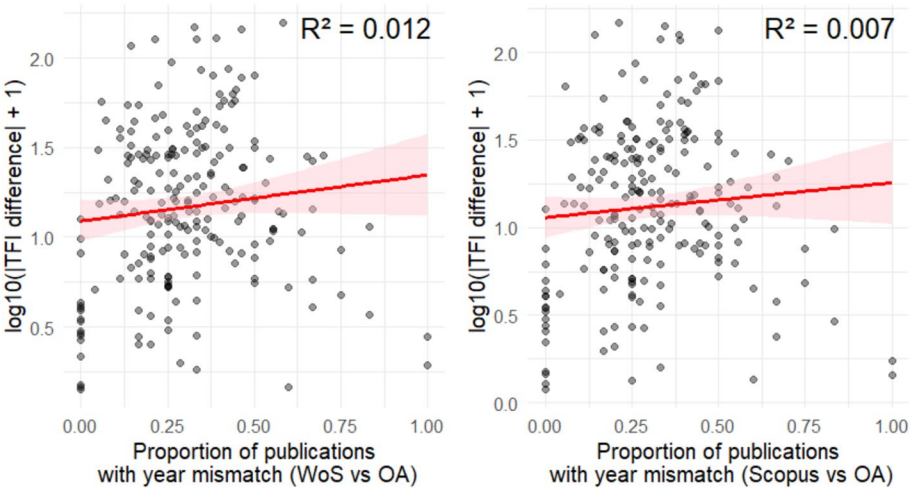


Fig. 13 Contingency matrices of TFI quartiles for pairs of databases, for the full and shared publications sets

indicate a very weak association. The regression models explain only a negligible share of the variance, with $R^2 = 0.012$ in the WoS-OA comparison and $R^2 = 0.007$ in the Scopus-OA comparison. The differences in year attribution across databases have only a minimal impact on the resulting author-level indicators.

The robustness checks show that restricting the analysis to the shared publication set produces results that are very similar to those obtained using the full publication set. The high correlations between indicators, the stability of quartile classifications, and the minimal impact of publication year discrepancies collectively indicate that the main findings of the study are not driven by differences in database coverage, but instead reflect a broadly consistent representation of author-level citation performance across the three sources.

Discussion

We now discuss what our results imply for evaluation, focusing on the relative advantages and drawbacks of OA compared with WoS and Scopus at both the researcher and institutional levels. Two robust patterns emerge. Output counts (N) are highly consistent across the three databases, indicating that, on average, researcher productivity is similarly captured regardless of the source. By contrast, impact (TFI) shows only moderate agreement, implying that database design choices (classification, normalization, and indexing conventions) materially affect impact-based assessments and the associated percentiles and quartiles.

Our findings should be interpreted in relation to recent comparative studies of OpenAlex and commercial databases. Scheidsteger et al. (2025) report strong agreement in field-normalized citation impact at the paper level, but substantially weaker agreement when indicators are aggregated at the university level. Our results are consistent with this pattern. While we observe high concordance across databases for output counts (N), agreement decreases markedly for impact (TFI), particularly when expressed in percentile and

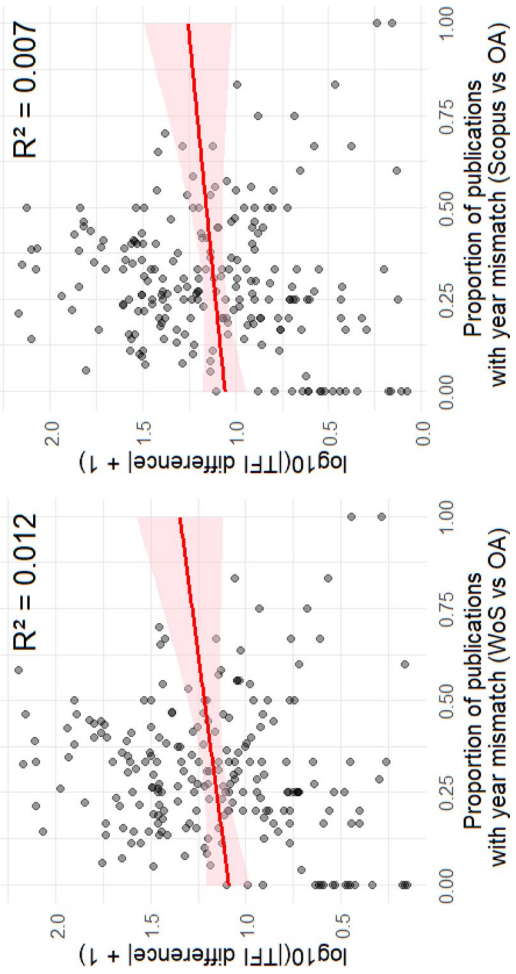


Fig. 14 Linear regression of TFI vs the share of publications whose indexing year differs between OA and WoS (left panel) or Scopus (right panel)

quartile-based rankings. This suggests that aggregation amplifies database-specific differences in classification systems, normalization procedures, and indexing conventions.

Similarly, Thelwall and Jiang (2025) conclude that OpenAlex can provide robust citation counts, but caution against its uncritical use in evaluative contexts due to metadata and classification issues. Our evidence refines this conclusion by showing that, even when citation counts are comparable or higher in OpenAlex, differences in field taxonomies and normalization denominators can generate more right-skewed impact distributions and greater rank volatility. In this sense, our analysis extends prior publication-level comparisons to the level of individual and institutional performance assessment.

Unlike most previous contributions, which focus primarily on coverage, metadata completeness, or publication-level citation indicators, our study directly examines the consequences of database choice for researcher and institutional performance rankings within a controlled national and disciplinary setting. The observed magnitude of quartile transitions provides concrete evidence of the practical implications of database selection in high-stakes evaluation exercises.

The evidence on the coverage–quality trade-off ("The trade-off between publication coverage and citation coverage" section) helps interpret these divergences. Importantly, these differences should not necessarily be interpreted solely as measurement problems or inconsistencies in bibliographic data. They may also reflect the fact that the databases effectively observe partially different publication universes. Because OpenAlex indexes a broader range of venues and document types, the reference distributions used for citation normalization may differ across sources. Consequently, normalized impact indicators are computed relative to slightly different baselines. From this perspective, part of the divergence observed in impact distributions and rankings reflects differences in scope rather than purely differences in data quality or indicator construction. At the portfolio level, OA captures almost the entirety of the WoS and Scopus sets (about 95% overlap in both cases), while adding a large number of additional records. OA indexes 3706 unique publications for the aggregated portfolio, compared with 2405 in WoS and 2699 in Scopus; OA also contains 1349 publications not found in WoS and 1037 not found in Scopus (Table 5). However, the additional coverage is not neutral in terms of quality as proxied by citations. Publications exclusive to OA are, on average, weakly cited (3.3 citations), and the overall average citations of the OA portfolio (28.5) are lower than those observed in WoS (32.4) and Scopus (35.9) (Table 6).

This trade-off is consequential for impact indicators that depend on normalization. On the one hand, OA often reports higher citation counts for publications shared with commercial databases (e.g., WoS-and-OA items average 34.0 citations in WoS but 48.2 in OA), suggesting broader citation linkages and/or different citation extraction (Table 6). On the other hand, the inclusion of many low-cited OA-only items changes the reference distribution used for scaling (and may interact with OA's fine-grained taxonomy), which can amplify the dispersion of field-normalized impact and contribute to the heavier right tails and outliers observed for c_i and TFI.

OA offers clear strengths for assessment. First, its broader coverage can deliver a more comprehensive census of research outputs, which is valuable for portfolio mapping and for contexts where commercial databases under-represent certain venues or formats. Second, OA's indexing-year convention, typically aligned with DOI registration/online-first availability, may better reflect the year in which outputs become publicly citable. Third, OA's open data model facilitates transparency, auditing, and reproducibility, enabling shareable workflows and easier replication of institutional assessments.

At the same time, the coverage–quality trade-off and OA's normalization mechanics warrant caution in high-stakes, rank-based settings. The additional OA-only records include heterogeneous venues (e.g., proceedings collections, institutional databases, and preprint-like outlets), which may be desirable for mapping activity but can dilute portfolio averages and increase sensitivity to database-specific choices. Moreover, as shown by the extreme FWCI cases, topic/subfield assignments to low-cited categories can inflate normalized impact for individual papers and, by aggregation, for researchers and small units. This helps explain why shifts in N and TFI are weakly coupled and why quartile transitions occur for a non-trivial share of scholars when switching sources.

For evaluators and decision makers, several practical steps can mitigate these risks without sacrificing OA's benefits. Where feasible, triangulate across at least two databases and disclose sensitivity (percentile deltas, quartile transitions, and min–max ranks). When using OA as the sole source, explicitly define the eligible publication set and consider reporting results for both a 'core' portfolio (intersection with WoS/Scopus, or otherwise curated journal/proceedings lists) and an 'extended' portfolio (including OA-only items) to make the coverage–quality trade-off visible. Audit affiliation and identity metadata against authoritative registries, and flag papers driving extreme normalized scores for manual inspection of subject assignment and indexing year. Finally, treat small-unit rankings with care by using minimum-size rules or uncertainty reporting, given the pronounced volatility observed when averages are computed over a few individuals.

In summary, OA advances openness, coverage, and reproducibility, but its broader and more heterogeneous coverage changes the citation baseline and interacts with OA-specific classification and indexing conventions.

Conclusions

This study compared OA, WoS, and Scopus for evaluating Italian professors in Management Engineering from 2019 to 2023, covering 223 academics from the national MUR registry. Publication counts (N) show high agreement across sources, but normalized impact (TFI) aligns only moderately, with heavier right tails and greater dispersion when computed from OA. These differences reflect database-specific choices in field taxonomies, normalization procedures, citation extraction, and indexing-year conventions.

OA largely subsumes the commercial databases while adding many additional records, and the added coverage is associated with lower average citation levels. At the aggregated portfolio level, OA contains 3,706 unique items versus 2405 (WoS) and 2699 (Scopus), and OA-only publications are weakly cited on average. This broader, more heterogeneous set lowers the portfolio's average citations in OA relative to WoS/Scopus, while OA may simultaneously report higher citation counts for the shared core set. Together, these features can alter the scaling baseline for field-normalized indicators and contribute to the observed dispersion and rank volatility.

These results must be read in light of several limitations. The analysis focuses on one national context and one disciplinary sector; results may differ in other fields or countries. The observation and citation windows are fixed (2019–2023, with citations normalized as of March 31, 2025). We operationalized assessment with N and TFI; other indicators could behave differently. Residual misattributions in authorship, affiliations, and indexing-year alignment cannot be ruled out, and OA, as a rapidly evolving infrastructure, may change in ways that alter some findings.

Future research should extend the design across disciplines and national systems, explicitly testing multilingual and regional publishing patterns. It should also quantify how much of the OA–WoS/Scopus divergence is explained by (i) coverage expansion into low-cited venues, (ii) differences in citation extraction and linkage, (iii) taxonomy/denominator effects in normalization, and (iv) indexing-year conventions. Developing uncertainty-aware rankings, small-N safeguards, and systematic audits of subject assignments for papers driving extreme normalized scores would be especially valuable.

For managers and policymakers, the implications concern the sensitivity of evaluation outcomes to the choice of bibliographic data source. In high-stakes settings, reliance on a single database may warrant caution. Where feasible, comparing results across multiple sources can serve as a precautionary practice to assess the robustness of evaluation outcomes and to disclose their sensitivity (percentile deltas, quartile transitions, and min–max ranks). When OA is used, conventions should be specified and harmonized in advance: define the eligible publication set and decide explicitly whether OA-only items are in scope, given the documented coverage–quality trade-off. Affiliations should be validated against authoritative lists before aggregation, and small-unit comparisons should be supported by minimum cohort sizes or uncertainty reporting. Crucially, volume and impact should be paired but not conflated: N is not, by itself, a performance indicator.

Another structural aspect that may contribute to differences in normalized impact indicators concerns the way citation links are constructed within each bibliographic environment. OpenAlex constructs citation counts based on the works indexed within its own database, whereas Web of Science and Scopus may also process references pointing to publications not themselves indexed as source items. To prevent this from having an impact on our comparison, we made sure that both the WoS CNCI and Scopus FWCI were calculated based on citations received only by publications indexed within each database environment.

To conclude, OA enables open and reproducible assessment and substantially broadens coverage, but the additional records tend to be less cited and can change normalization baselines while interacting with OA-specific classification and indexing conventions. Consequently, impact-based and rank-based decisions are particularly sensitive to OA-specific mechanics. Responsible evaluation should therefore treat database choice explicitly and, where possible, complement OA-based analyses with comparisons across alternative data sources or with transparent robustness checks that make the coverage–quality trade-off explicit.

The evolving landscape of bibliometric databases demands a nuanced, context-sensitive approach to research evaluation. Harnessing the strengths of OA, Scopus, and WoS, while addressing their respective limitations, will be essential for fostering a more equitable, reproducible, and comprehensive global research ecosystem.

Author contributions GA: Conceptualization, Investigation, Supervision, Writing—original draft, Writing—review and editing. TC: Investigation, Data curation, Methodology, Visualization, Writing—original draft. CAD: Conceptualization, Investigation, Data curation, Methodology, Writing—original draft.

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

Data availability The analysis presented in this manuscript is based on bibliometric data from openalex, WoS (Clarivate Analitycs) and Scopus (Elsevier). Access to data from these last two bibliometric sources cannot be granted by the authors.

Declarations

Competing interests The authors declare no competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Abramo, G., & D'Angelo, C. A. (2014). How do you define and measure research productivity? *Scientometrics, 101*(2), 1129–1144.

Abramo, G., D'Angelo, C. A., & Felici, G. (2019). Predicting long-term publication impact through a combination of early citations and journal impact factor. *Journal of Informetrics, 13*(1), 32–49.

Abramo, G., Cicero, T., & D'Angelo, C. A. (2025). Enhancing the prediction of publications' long-term impact using early citations, readerships, and non-scientific factors. *Journal of Informetrics, 19*(4), Article 101725.

Aria, M., Le, T., Cuccurullo, C., Belfiore, A., & Choe, J. (2024). openalexR: An R-tool for collecting bibliometric data from OpenAlex. *The R Journal, 15*(4), 167–180.

Asubiaro, T. V., & Onaolapo, S. (2023). A comparative study of the coverage of African journals in Web of Science, Scopus, and Crossref. *Journal of the Association for Information Science and Technology, 74*, 745–758.

Besançon, L., Cabanac, G., Labbé, C., & Magazinov, A. (2024). Sneaked references: Fabricated reference metadata distort citation counts. *Journal of the Association for Information Science and Technology, 75*(12), 1368–1379.

Brooks, J. (2023). *Leiden rankings to add open-source version in 2024*. Research Professional News.

Céspedes, L., Kozlowski, D., Pradier, C., Sainte-Marie, M. H., Shokida, N. S., Benz, P., Poitras, C., Ninkov, A. B., Ebrahimy, S., Ayeni, P., Filali, S., Li, B., & Larivière, V. (2025). Evaluating the linguistic coverage of OpenAlex: An assessment of metadata accuracy and completeness. *Journal of the Association for Information Science and Technology, 76*(6), 884–895.

Char, D. C., & Ajiferuke, I. (2011). The usefulness of related functions in Web of Science and Scopus. *Evidence Based Library and Information Practice, 6*(1), 41–52.

Clavey, M., 2024. *La recherche française parie sur OpenAlex pour briser l'emprise d'Elsevier et Clarivate*. Retrieved 17 March 2026.

CoARA. (2022). Retrieved March 17, 2026, from https://coara.eu/

Culbert, J. H., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2025). Reference coverage analysis of OpenAlex compared to Web of Science and Scopus. *Scientometrics, 130*(4), 2475–2492.

CWTS. (2023). *Introducing the Leiden ranking open Edition*. Leiden Madtrics. Retrieved March 17, 2026, from https://www.leidenmadtrics.nl/articles/introducing-the-leiden-ranking-open-edition

D'Angelo, C. A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology, 62*(2), 257–269.

Dalmeet, S. C. (2022). Massive open index of scholarly papers launches. *Nature*. https://doi.org/10.1038/d41586-022-00138-y

Declaration on Research Assessment (DORA). (2012). *San Francisco declaration on research assessment*. Retrieved March 17, 2026, from https://sfdora.org/read/

Haupka, N., Culbert, J. H., Schniedermann, A., Jahn, N., & Mayr, P. (2024). Analysis of the publication and document types in OpenAlex, Web of Science, Scopus, PubMed and Semantic Scholar. *arXiv*. https://doi.org/10.48550/arXiv.2406.15154

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature, 520*, 429–431. https://doi.org/10.1038/520429a

Jason, 2021. *We're building a replacement for Microsoft Academic Graph*. OurResearch blog. Retrieved March 17, 2026, from https://blog.openalex.org/were-building-a-replacement-for-microsoft-academic-graph/

Khanna, S., Ball, J., Alperin, J. P., & Willinsky, J. (2022). Recalibrating the scope of scholarly publishing: A modest step in a vast decolonization process. *Quantitative Science Studies, 3*(4), 912–930.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*(12), 317–324.

Maddi, A., Maisonobe, M., & Boukacem-Zeghmouri, C. (2025). Geographical and disciplinary coverage of open access journals: OpenAlex, Scopus, and WoS. *PLoS ONE, 20*(4), e0320347.

Parsons, S., Azevedo, F., Elsherif, M., Guay, S., Shahim, O. N., Elsherif, M. M., Govaart, G. H., Norris, E., O'Mahony, A., Parker, A. J., Todorovic, A., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E., … Pennington, C. R. (2022). A community-sourced glossary of open scholarship terms. *Nature Human Behaviour, 6*(3), 312–318.

Pölönen, J., & Kulczycki, E. (2025). Multilingualism is important for all fields of science: Evidence from Finland and Poland. *Language and the knowledge economy* (pp. 151–175). Routledge.

Price, G., 2024. *Funding: OurResearch receives $7.5 million grant from Arcadia to establish OpenAlex, a milestone development for Open Science*. Library Journal infoDOCKET. Retrieved March 17, 2026, from https://www.infodocket.com/2024/03/14/funding-ourresearch-receives-7-5-million-grant-from-arcadia-to-establish-openalex-a-milestone-development-for-open-science/

Priem, J., Piwowar, H., Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. Preprint retrieved from https://arxiv.org/abs/2205.01833

Scheidsteger, T., Haunschild, R., & Bornmann, L. (2025). How similar are field-normalized citation impact scores obtained from OpenAlex and three popular commercial databases? An empirical comparison based on large German universities. *Scientometrics, 130*(7), 3537–3569.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Stahlschmidt, S., & Stephen, D. (2022). From indexation policies through citation networks to normalized citation impacts: Web of Science, Scopus, and Dimensions as varying resonance chambers. *Scientometrics, 127*, 2413–2431.

Suber, P. (2004). Open access overview. Retrieved March 17, 2026, from https://legacy.earlham.edu/~peters/fos/overview.htm

Thelwall, M., & Jiang, X. (2025). Is OpenAlex suitable for research quality evaluation and which citation indicator is best? *Journal of the Association for Information Science and Technology, 76*(12), 1660–1681.

Vicente-Saez, R., & Martinez-Fuentes, C. (2018). Open science now: A systematic literature review for an integrated definition. *Journal of Business Research, 88*(C), 428–436.

Visser, M., van Eck, N. J., & Waltman, L. (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies, 2*(1), 20–41.

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., van Eck, N. J., van Leeuwen, T. N., van Raan, A. F. J., Visser, M. S., & Wouters, P. (2012). The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology, 63*(12), 2419–2432. https://doi.org/10.1002/asi.22708

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & Johnson, B. (2015). *The metric tide: Report of the independent review of the role of metrics in research assessment and management*. UK Research and Innovation.

Zhang, L., Cao, Z., Shang, Y., Sivertsen, G., & Huang, J. (2024). Missing institutions in OpenAlex: Possible reasons, implications, and solutions. *Scientometrics, 129*, 5869–5891.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.