Health Policy 124 (2020) 1387–1394

Contents lists available at ScienceDirect

Health Policy

journal homepage: www.elsevier.com/locate/healthpol

Measuring research in the big data era: The evolution of performance measurement systems in the Italian teaching hospitals

Frank Horenberg[a,b,*], Daniel Adrian Lungu[a,b], Sabina Nuti[a,b]

[a] Health and Management Laboratory (MeS Lab), Institute of Management and Department EMbeDS, Scuola Superiore Sant'Anna, Piazza Martiri della Libertà, 33, Pisa, Italy
[b] Sant'Anna School of Advanced Studies, Health and Management Laboratory (MeS Lab), Piazza Martiri della Libertà, 33, 56127 Pisa PI, Italy

ARTICLE INFO

Article history:
Received 2 July 2019
Received in revised form
21 September 2020
Accepted 4 October 2020

Keywords:
Teaching hospitals
Research productivity
Performance evaluation
Impact factor
Field-weighted citation impact

ABSTRACT

Background: In the healthcare system, Teaching Hospitals(THs) not only provide care, but also train healthcare professionals and carry out research activities. Research is a fundamental pillar of THs' mission and relevant for the healthcare system monitored by Performance Evaluation Systems. Research activities can be measured using citation index services and this paper highlights differences between two services based on bibliometrics, describes opportunities and risks when performance indicators rely on data collected, controlled and validated by external services and discusses the possible impact on health policy at a system and provider level.
Methods: A bibliometric analysis was done on data between 2014–2016 from ISI Web of Science and Scopus of 18.255 physicians working in 26 Italian THs. Quantity was defined as the number of publications and quality as Impact Factor or Field-Weighted Citation Impact.
Results: Overall, 41.233 and 66.409 documents were extracted from respectively ISI Web of Science and Scopus. While benchmarking results, significant differences in ranked position both in metrics emerged.
Discussion: Utilizing secondary data sources to measure research activities of THs allows benchmarking at an (inter)national level and overcoming self-referment. To utilize indicators for multiple governance purposes at the system and provider level, indicators need to be profoundly understood, require formalizations in data validation, internal analysis and a sharing process among health professionals, management and policymakers.

1. Background

Performance evaluation systems (PESs) are crucial for accountability and serve as a feedback and guidance tool for the managerial level of organizations [1]. PESs are used to evaluate how well organizations are managed and to measure the value that organizations deliver to customers and other stakeholders [2,3]. From the 1980s, PESs have broadened the kind of indicators monitored, but maintained the focus on financial ones [4,5]. Alongside, developments in information and communications technology (ICT) facilitated data availability, completeness, and accessibility and the evolution of the so-called Big Data turned useful to enrich the PES information [6–8].

Still now PESs in economic sectors that are profit-oriented are mainly focused on measures regarding profit and revenues, while this is not the case in healthcare where the goal is to produce value for patients and the population [9–11]. Within the healthcare sector non-financial indicators are crucial and PESs, mostly in public universal coverage healthcare systems where revenues are based on a per capita quota, are designed and implemented to be able to measure on one side outcomes, quality of care and life, identify issues, and on the other hand resources made available by society.

In order for PESs to be effective in public universal coverage healthcare system, it should be characterized by the following elements [12,13]:

- Multi-dimensionality: Indicators should include multiple dimensions (process, quality of care, equity, etc.);

Abbreviations: IRPES, Inter-Regional-Performance Evolution System; LHA, local health authority; PES, performance evaluation system; TH, teaching hospital; WoS, ISI Web of Science.

* Corresponding author at: Sant'Anna School of Advanced Studies, Health and Management Laboratory (MeS Lab), Piazza Martiri della Libertà, 33, 56127 Pisa PI, Italy.
E-mail addresses: horenbergfrank@gmail.com (F. Horenberg), danieladrian.lungu@santannapisa.it (D.A. Lungu), sabina.nuti@santannapisa.it (S. Nuti).

https://doi.org/10.1016/j.healthpol.2020.10.002

- Evidence-based: on research or clinical practice;
- Shared design: all stakeholders, and especially health professionals, should be involved in the design and the fine-tuning process of the PES and the indicators;
- Systematic benchmarking: allows to overcome self-referentiality and to measure avoidable variation and space for improvement;
- Transparent disclosure: stimulates data peer-review and makes professional reputation leverage possible;
- Timeliness: allows policymakers to make decisions promptly.

With these premises, PESs in healthcare have been evolving overcoming the organizational boundaries of single providers [14]. Moreover, among the above elements, the most crucial is to rely on benchmarking which facilitates and triggers organizational improvement processes to increase effectiveness based on reputation [15].

Considering these relevant features for the healthcare sector, this paper focuses on the PESs adopted by teaching hospitals (THs) within the healthcare system. These kind of institutions, even if they may be no different from other hospitals in terms of quality of care [16], fulfill a special role in the healthcare system because their mission is not just providing care, but also to train healthcare professionals and to carry out research activities.

As medical knowledge continuously evolves, THs are at the center of innovations in healthcare with respect to treatments and cures. They are in charge of conducting research and adding new knowledge to scientific literature. For this reason, research is a fundamental pillar of the THs' mission and therefore a relevant component to be included in PESs, not just for the single provider but for the whole healthcare system. The ability of THs to perform research activities guarantees them a role of reference and guide in the processes of improving care at regional or national level. The more healthcare professionals know how to be on the frontiers of clinical research, the more likely it is that care will be aligned with the best and most updated clinical protocols benefiting for patients. It is in fact proven that the best hospitals are those where more research is carried out and in turn healthcare outcomes improve [17,18] and patients benefit from access to new and innovative treatments that would not be otherwise available [19,20]. Furthermore, research activities should guide processes to improve quality of care generation knowledge which leads to updated and trained staff to establish teams of experts and centers of excellence [23–26].

Measuring performance of research activities thus becomes a relevant topic for the whole healthcare system and for each TH that operates in it. However, is an endeavor as research activities result in both intangible (e.g. knowledge, experience) and tangible outputs (e.g. scientific articles, products) and accurate measurements depend on many preconditions [27,28].

A measure that is frequently used is scholarly output [27], by counting the number of published documents. This proxy can be accessed via readily available data sources from publishers, journals, citation indexing services, and other secondary data sources. It is a simple measure that can be detected internally by each TH, but which also has an external value: the articles have been published and therefore recognized by the scientific community as significant contributions to the evolution of science.

Quantifying research using the number of published articles can therefore be the first step to measure research performance but does not provide any information about the value and impact of these published works. Other metrics are therefore needed to measure the quality of these works to provide context and the impact within the research community [29]. In this perspective, citations can be a reference element, since it is an indirect positive evaluation that the work has been read and taken into consideration - and therefore mentioned - by colleagues.

In order to provide context and assess the impact of scholarly output, a quantitative method can be employed, namely a bibliographic analysis [30]. Nowadays, various commonly used metrics can be used to assess value i.e. downloads and views, citations, impact factor, h-index and field-weighted citation impact (FWCI), altmetrics (storage, links, bookmarks, conversations) and many others [31–34]. These metrics can be utilized both on a journal or individual researcher level. Most bibliometrics are calculated, managed, and tracked using citation index services such as *Scopus*, and *ISI Web of Science* (WoS) and can be accessed via the Internet.

These different citation index services are an access point to different repositories which store and categorize scholarly output. However, each of these services differs in their coverage, method of tracking, and available metrics [35]. Since data provided by these services are managed and controlled by external parties, new opportunities and challenges arise for PESs when used to evaluate and benchmark performance.

- What are the main differences between the commonly used scholarly metrics extracted from citation index services and derived using bibliographic analyses?
- What are the opportunities and risks when PESs and their corresponding governance tools are relying on data collected, controlled, and validated by external sources?
- What is the impact in terms of health policies at a system level?

This paper, therefore, describes the differences by focusing on the quantity and quality of research performance of THs that result from the use of two different scientific citation indexing services on the web, *Scopus* and *ISI Web of Science Core Collection*, by performing a bibliometric analysis of 26 THs in Italy. Moreover, results are contextualized by discussing the possible implications on PESs when measuring research performance through alternative external citation index services.

The next section conceptualizes different metrics used to evaluate scientific production provided by both *Scopus* and *WoS* and a description of the implementation of research performance in a regional performance evaluation system in Italy. The third section describes the methodology and comparing metrics from both services. Findings are contextualized in the final discussion and conclusion sections.

1.1. Measuring scientific performance

As early as in 1927 Gross et al. identified the problem of disseminating literature and in 1955 Eugene Garfield proposed to utilize a citation index for scientific literature to eliminate the uncritical citation of fraudulent, incomplete, or obsolete data [31,36]. Later, in 1961, as a founder of the Institute for Scientific Information (ISI) Garfield launched the Science Citation Index as a tool for researchers, librarians and scholars to manage the large number of library collections. Over time, the purpose of the citation index changed, now known as *impact factor* which was intended to describe journal impact based on the number of citations [37]. Although this metric was never designed or intended to be used as an evaluation indicator, in practice it is often used to indicate the quality of individual scientific work [38]. The scientific community has often expressed concerns regarding the biased impact factor of journals deriving from asymmetric - left skewed - distribution of paper citations [39,40] and causing quite some controversy within the research community [38,41–43].

Although many alternatives evaluation metrics have been proposed by the research community, impact factor remains dominant in usage. However, Larivière et al. proposed a simple and robust methodology to defer the citation distributions that underlie the Journal impact factor creating more transparency [44]. This pro-

posed method was adopted by the Journal Citation Reports (JCR) in 2018 and seems to be an earnest first attempt to address the concerns of the scientific community [45].

Another well-known problem with impact factor is the skewness in specific research fields. For example, Narin et al. reported that research in biochemistry and molecular biology were cited about five times as often as pharmacy articles [46]. In order to correct for this phenomenon, the Dutch publisher Elsevier implemented their own metric, namely the field-weighted citation impact metric. The FWCI shows how the number of citations of a single paper compares with the average number of citations by similar publications indexed in Scopus [47] resolving the issue of different research behavior across disciplines.

Another well know metric group are the Altmetrics which use multiple data sources such as social media, number of readings and downloads to assess the impact of the paper both inside and outside the scientific community. The ability to measure impact of scientific work outside the scientific community is a valuable trait. However, the actual use of Altmetrics needs to be further conceptualized to become a metric on its own while still lacking a clear definition, an ever-evolving framework, low data transparency, and origin [48].

Evaluating the large body of available bibliography, it becomes clear that it is not possible to measure scientific performance by simply using one measure and while other measures such as the *h-index* and *g-index* are increasingly used to evaluate researchers' performance, using different metrics, emphasizing both productivity, quality and context, inside and outside the research community is imperative.

1.2. Evaluating research performance in Italy

The Italian National Healthcare System follows a Beveridge model, mainly financed through general taxation and based on the principle of universal coverage. Resources are collected at a national level and allocated to the twenty regions on age-adjusted per capita basis. The responsibility for the organization and provision of care has been decentralized at a regional level, and regions allocate resources to Local Health Authorities (LHAs) who are responsible for the delivery of all healthcare services in their geographical area, directly through public providers or accredited private hospitals. THs are autonomous bodies from the LHA, can be public or private, and are usually managed jointly by the regional administration and a university. This shared responsibility in managing has a relevant impact on their organizational culture as they can be considered double professional bureaucracies [49]. Within the regional healthcare system, they play a relevant role because they oversee training of future health professionals and because they are in charge of leading innovation processes based on the research activities that they carry out.

Starting from 2005, the Management and Health Lab of the Scuola Superiore Sant'Anna has developed a multidimensional healthcare performance evaluation system (PES), initially adopted by Tuscany's regional administration. Over time, the PES was adopted by an increasing number of regions and in 2008 the Network of Italian Regions was formed. The Network expanded and nowadays twelve regions have adopted the same Inter-Regional-Performance Evolution system (IRPES) [13] to benchmark their performance using more than three hundred shared indicators. In 2014, the Network of THs was founded, aimed at benchmarking THs performance using a PES to take into account the specific characteristics and mission of THs within the regional healthcare system considering around sixty indicators [16,50,51].

Reporting on these indicators are considered an important management tool by all THs and the Italian government and regions use them for several issues as monitoring and assessing performance, allocate financial resources for research, and also to evaluate the General Managers' performance [52]. Given that research is one of the three pillars of THs' mission, within these sixty indicators some are focused on research performance.

Table 1 provides an overview of the indicators included in the IRPES which are used to evaluate research activities.

Table 1
Indicators included in the IRPES regarding research evaluation of teaching hospitals.

Number	Description of indicator
1	Average impact factor per physician
2	Average number of publications per physician
3	Percentage of publications with an average impact factor higher than the benchmark specialty impact factor reported in ISI
4	Percentage of publications with a median impact factor higher than the benchmark specialty impact factor reported in ISI
5	Median impact factor per specialty
6	Median impact factor variation per specialty

2. Methods

The bibliometric analysis can be performed using two well-known publicly accessible citation indexing services, namely, *ISI web of knowledge* and *Scopus*. Metrics about the scholarly output can be extracted from these repositories by simply providing author first name, last name, and optionally their affiliated organization.

The possibility of using these search engines, external to the internal detection systems, has always been perceived as an opportunity to have čertainänd validated data, a fundamental characteristic to guarantee the strength, rigor and reputation of the PES itself.

However, even these systems show some criticalities. For example, when an author is affiliated with multiple organizations or duplicate names are affiliated with the same organizations manual correction is required. Data extraction, article de-duplication of (co-)authors, and validation were done in two different manners for both repositories including all scientific documents in these databases which have been published between 2014–2016. A detailed protocol of the data extraction can be found as supplementary material; *Appendix A (in Supplementary material) – Extraction of data*. THs were responsible for providing names of researchers affiliated with their organization.

2.1. Data extraction WoS

Documents published in *ISI Web of Science* between 2014–2016 were extracted and validated by an external affiliated party, Research Value SRL, in May 2018. Data was sent to corresponding authors for internal validation on completeness using random sampling methods. All available metrics were extracted from ISI-WoS, including; title, discipline, document type, affiliated authors, ISSN, ISBN, year, edition, page numbers, subject categories, DOI, PubMed identification number, number of citations and journal's impact factor.

2.2. Data extraction Scopus

Scholarly output production between 2014–2016 in *Scopus* was extracted utilizing internally written scripts using Elsevier's API developers' program. The script was divided into two main functionalities, *Match* and *Extract*, and was executed in December 2018. The first part queries the *Scopus* database author names to obtain a unique identification code used in all Elsevier's products such as *Scopus* and *Scival*. Only when a unique match was found based on name and affiliation a *Match* was deemed successful. When the

Table 2
Number of published documents in both Scopus and WoS database with their respective difference and change in ranking when benchmarked per TH.

Teaching hospital	Number of physicians	Number of documents **Scopus**	Number of documents **WoS**	Difference in documents n (%)	Difference in position when benchmarked
AO Padova	854	5854	3579	2275 (38,9%)	▲1
AOU Bologna	868	4605	2727	1878 (40,8%)	▲1
AOU Careggi	1026	4343	2684	1659 (38,2%)	▲1
S. Raffaele - MI	526	4140	3583	557 (13,5%)	▼-3
AOU Verona	876	3910	2247	1663 (42,5%)	▲1
Fondaz.IRCCS Ca Granda	780	3818	2470	1348 (35,3%)	▼-1
IRCCS S. Martino	862	3334	2223	1111 (33,3%)	◄0
AOU Pisana	952	3295	2222	1073 (32,6%)	◄0
P.O. Spedali Civili Brescia	1039	3184	1766	1418 (44,5%)	◄0
Ist. Clin. Humanitas - Rozzano	679	2928	1305	1623 (55,4%)	▲4
AOU Pol. Bari	895	2878	1708	1170 (40,7%)	▼-1
IRCCS Policlinico San Matteo	571	2304	1485	819 (35,5%)	▼-1
AOU Parma	696	2208	1374	834 (37,8%)	▼-1
AOU Osp. Riun. Ancona	758	2049	1184	865 (42,2%)	▲3
Osp. S.Gerardo - Monza	735	1961	871	1090 (55,6%)	▲5
AOU Senese	573	1958	1309	649 (33,1%)	▼-3
AO Perugia	576	1882	1304	578 (30,7%)	▼-2
AOU Modena	489	1850	1199	651 (35,2%)	▼-2
Osp. L. Sacco - Milano	559	1623	859	764 (47,1%)	▲2
ASUI Udine	745	1573	1017	556 (35,3%)	▼-1
AOU Ferrara	523	1542	1020	522 (33,9%)	▼-3
Osp. S. Paolo - Milano	736	1441	742	699 (48,5%)	▲1
Osp. di Circolo e Fond. Macchi	589	1274	801	473 (37,1%)	▼-1
ASUI Trieste	539	1165	640	525 (45,1%)	◄0
OO.RR. Foggia	412	811	570	241 (29,7%)	◄0
AO Terni	397	479	344	135 (28,2%)	◄0
Total	*18.255*	*66.409*	*41.233*	*25.176 (37,9%)*	

search resulted in multiple possible authors a manual validation was done by the authors, selecting, or merging the researcher profile(s).

The second part, *extracts* published work from Scopus and Scival. All available metrics were extracted from *Scopus*, including; title, DOI, ISSN, Journal name, type of publication, cover data, number of citations, affiliation organization. As Scopus does not allow to track any value metrics, the FWCI per author using the Elsevier identification number was extracted from Scival.

A detailed description of the full script can be found in the supplementary material; *Appendix A (in Supplementary material) – Extraction of data.* Full script details used to obtain data from Scopus and statistical procedures can be requested via the corresponding author. Statistical analysis was performed using R version 3.5.2.

3. Results

After extracting the scholarly output of all 26 THs, a total of 66.409 and 41.233 documents are included for analysis from Scopus and respectively WoS from a total of 18.255 authors. Descriptive statistics about the THs can be found in Appendix B *(in Supplementary material) – Details Teaching Hospitals*. Documents are categorized as articles (69,2 % Scopus; 75,3 % Wos), reviews (13,6 % Scopus; 14,29 % WoS), Letters (7,1 % Scopus; 9,4 % Wos), editorials (1,89 % Scopus), book(chapters) (2,97 % Scopus) or other (5,2 % Scopus; 1,0 % WoS). The preceding two categories are only indexed in Scopus. Table 2 and Fig. 1 compare the total number of documents per TH in WoS and Scopus published between 2014–2016. Table 3 shows an overview of the total number of documents per THs between WoS and Scopus excluding book(chapters) and editorials which are not indexed in WoS to provide a more accurate comparison.

A Wilcoxon signed rank test was performed to compare the difference of indexed documents in both databases, indicating a significant difference ($p < 0.005$) in the documents indexed in Scopus ($M = 2.060$, $SD = 1.122$) and WoS ($M = 1.267$, $SD = 863$). When ranking THs based on the scholarly output as shown in Table 2, almost all organizations are benchmarked at a different position. On average, institutes change two positions either positive or negative. The biggest positive change in the ranking when looking at Scopus is the Teaching hospital "AO San Gerardo di Monza" moving from the 20th position the 15th position.

Although the quality metric extracted from *Scopus* and *WoS* cannot be directly compared with each other since WoS measures impact factor and Scopus measures FWCI, investigating the quality of the published documents shows a difference in ranking when benchmarked. A detailed overview can be found in Table 4, showing both impact factor and FWCI of the institutes and their respective ranking when benchmarked. On average, institutes change five positions either positive or negative. The biggest positive change in the ranking can be seen with "AO San Gerardo di Monza" moving from the 19th position the 3rd position. However, some organizations also move down in the ranking. AOU Careggi is placed on 18th position when ranking the organization with FWCI but is ranked 7th when benchmarking with impact factor. None of the organizations remain at the same position when comparing the benchmark on Impact factor or FWCI.

Finally, Fig. 2 shows the relationship between quality and quantity between the published works. Calculating the Spearman's rho shows a low but positive correlation between the quality (FWCI) of produced documents and the number of documents per researcher.

4. Discussion

This paper describes the differences in performance of 26 THs in Italy by focusing on the quantity and quality of their research. The goal of this paper was to perform a bibliometric analysis focusing on two commonly used performance metrics, impact factor and FWCI using Scopus and ISI Web of Science Core Collection to identify main differences and potential opportunities and challenges for PESs as a strategic tool at the provider and system level.

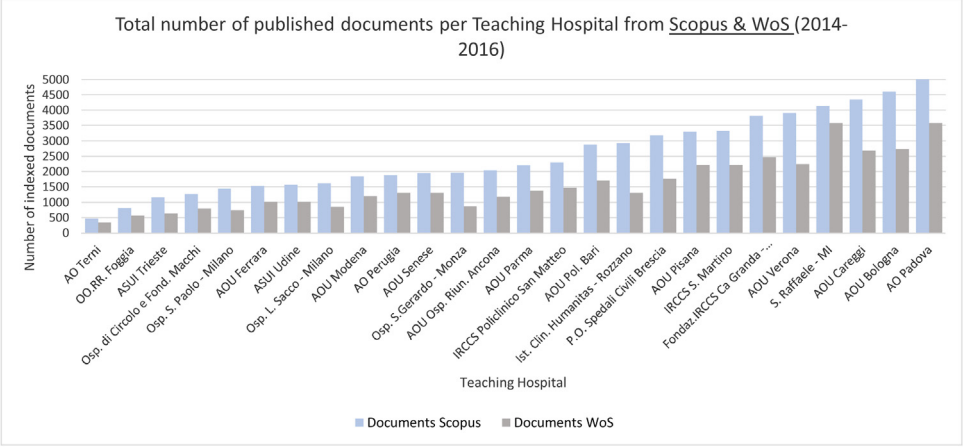

Fig. 1. Number of published documents in both Scopus and WoS database with their respective difference when benchmarked per TH.

Table 3
Number of published documents in both Scopus and WoS database with their respective difference and change in ranking when benchmarked per TH, excluding book(chapters) and editorials.

	ISI WoS	**Scopus** (articles & reviews)	**Scopus** (articles, reviews & conference papers)	**Scopus** (all documents)
Number of published documents (2014–2016)	41.233	55.805	56.962	66.409
Difference (%)		14.572 (26,1 %)	15.729 (27,6 %)	25.176 (37,9 %)

Table 4
Quality of published documents in both Scopus and WoS database with their ranking and respective change in ranking when benchmarked per TH. Quality is defined as the average impact factor or Field-weighted citation impact of all authors affiliated to the TH.

Teaching hospital name	Impact Factor	Field-Weighted Citation Impact	Ranking **Scopus**	Ranking **WoS**	Difference in position when benchmarked
AOU Bologna	13,49	2,66	1	4	▲3
Ist. Clin. Humanitas - Rozzano	10,66	2,54	2	9	▲7
Osp. S.Gerardo - Monza	5,84	2,38	3	19	▲16
S. Raffaele - MI	35,70	2,14	4	1	▼-3
AO Perugia	10,57	2,07	5	10	▲5
AO Padova	18,03	1,97	6	2	▼-4
AOU Pisana	8,99	1,93	7	12	▲5
AOU Modena	10,19	1,91	8	11	▲3
IRCCS Policlinico San Matteo	12,06	1,88	9	5	▼-4
Fondaz.IRCCS Ca Granda	14,27	1,85	10	3	▼-7
P.O. Spedali Civili Brescia	7,41	1,84	11	15	▲4
AOU Senese	8,68	1,82	12	13	▲1
AOU Verona	10,96	1,79	13	6	▼-7
AOU Osp. Riun. Ancona	6,23	1,78	14	18	▲4
AOU Ferrara	7,42	1,77	15	14	▼-1
ASUI Udine	5,40	1,77	16	21	▲5
IRCCS S. Martino	10,68	1,73	17	8	▼-9
AOU Careggi	10,69	1,62	18	7	▼-11
AO Terni	3,52	1,60	19	24	▲5
OO.RR. Foggia	4,38	1,59	20	23	▲3
Osp. di Circolo e Fond. Macchi	5,03	1,57	21	22	▲1
Osp. L. Sacco - Milano	5,67	1,56	22	20	▼-2
AOU Parma	6,99	1,54	23	17	▼-6
Osp. S. Paolo - Milano	3,43	1,52	24	26	▲2
AOU Pol. Bari	7,00	1,48	25	16	▼-9
ASUI Trieste	3,51	1,33	26	25	▼-1

4.1. Bibliometric analysis

Extracting the scholarly output showed a significant discrepancy between the extracted data from the two repositories. When extracting the full scholarly output of the 18.255 authors in the sample, Scopus resulted in 37,9 % more documents (66.409 Vs. 41.233). To a certain degree, this difference can be explained. First, apart from reviews, articles, conference papers, book chapters, Scopus also indexes books and editorials in their database. WoS does not include these two categories into their core collection, thus explaining 9,7 % of the variation. Second, Scopus is known to be more extensive in their coverage including over 71 million records and covering over 23,700 peer-reviewed journals [53] while WoS includes just over 20.000 peer-reviewed journals [54]. It is, therefore, possible that some articles are not indexed in both databases. Third, since authors are searched using only name, surname, and affiliation and it is possible that authors are indexed differently in both databases resulting in a mismatch when extracting information. However, at this stage, we are unable to provide an exact quantification of this observed variation.

When comparing the quality indicator of both databases and benchmarking THs based on impact factor and FWCI, none of the organizations remain at the same position. Interestingly, data shows a positive correlation between the quality and quantity of the

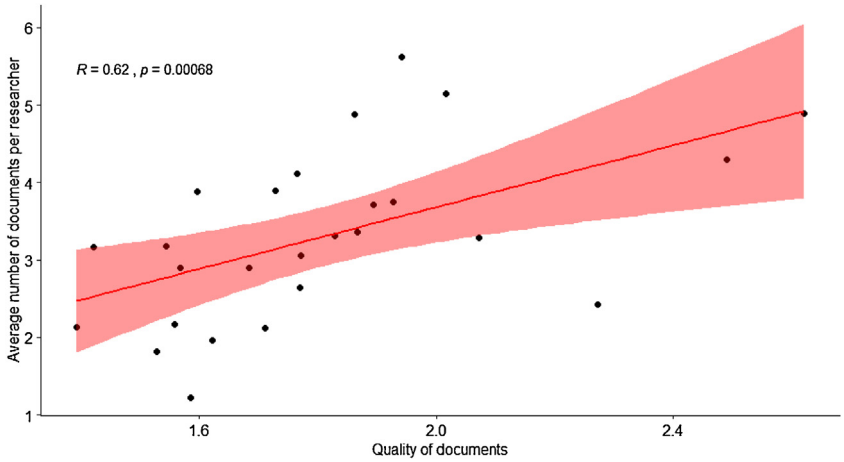

Fig. 2. Correlation between quality (FWCI) and published documents per researcher in Scopus.

works published by the THs. Haslam & Laham, hypothesized that researchers in more prestigious institutions may follow a strategy where the focus is more on the quality of published papers and less on the quantity [55]. Our results contradict this hypothesis proving a positive relationship between quantity and quality of published works indicating that THs publishing more documents also produce higher quality documents. However, it should be noted that with the current number of observations the relationship is weak and might not sustain with an increased sample size.

Archambault et al., 2009 provides evidence that indicators of scholarly production and citations at the country level are stable and largely independent of the database reported and no significant bibliographic differences between Scopus and WoS are found [35]. We were able to compare results on the individual researcher, now suggesting that a significant difference exists between both repositories, rejecting the findings of Archambault. However, Archambault was unable to investigate the scholarly output on an individual level and focused on an institutional level.

4.2. Impact on performance evaluation systems

The presented findings can have important implications for the current use of performance evaluation systems in the healthcare sector.

In traditional PESs data are measured, calculated and validated by the organizations themselves, using benchmarking in order to compare results with others, on different levels such as individuals, departments, and organizations [14]. Using secondary big data sources opens new opportunities to benchmark outside the organizational boundaries with other organizations on a national and international level.

This reduces the role of each single TH in the collection of data and reassures the Regional or National Health System about the reliability of the data itself, as there is a reduced risk of opportunistic data manipulation. The benchmarking process, at a first glance, appears more robust.

Services to consult bibliographic information are publicly available and easily accessible via the Internet. However, data in these systems are not managed, owned, and often not validated by the organization themselves, but by external parties such as Clarivate Analytics and Elsevier which have partly a commercial interest. Especially since numerous studies have provided evidence about inaccurate information, falsification, and fabrication of data in citation index services which affect and influence the bibliometric measures [56–59]. Additionally, metrics measuring the same construct namely *quality* often differ from each other and are all subjected to their own advantages and disadvantages making comparisons challenging.

These indicators are an important management tool used by the Italian government, the Regions and the THs. They use them for monitoring and assessing performance, allocating financial resources for research, and evaluating the General Managers' performance. We want to underline that choosing one of these databases is not sufficient nor reliable to base important health policy decisions on without including contextual information.

The differences between the two metrics found in the results, in fact, highlight the intrinsic weakness of these metrics which, to be effective, require a significant work to critically assess the meaning using contextual information. Validation of the origin of the metric is a key step in the age of Big Data before assessing the meaning of the metric itself. Moreover, increasing the scope of benchmarked organizations provides new insights to policymakers, and can support beneficial strategies when using PESs such as naming and shaming [60] or rewarding organizations for higher performance [61]. This goal can only be achieved if data are reliable. The fact that the research indicators are based on systems such as WoS and Scopus does not guarantee *per se* the pursuit of this condition.

Health systems must accompany the use of these metrics with a continuous sharing process with all the stakeholders of the system and first of all with the researchers themselves [62]. This same sharing process represents the first mechanism to align efforts and commitment towards pursuing the overall mission of the healthcare system and it is the basis of the relationship of trust and esteem that allows to feed and promote improvement processes.

Finally, other several issues should be mentioned possibly influencing the presented results. First, although a representative sample size of 18.255 authors was used, authors were not able to validate each individual researcher. Names of researchers were provided by all THs in the IRPES network, but authors were not able to validate to what extent these researchers were actively working for the THs or provide any descriptive statistics about these authors. Second, data from WoS was extracted and primarily validated by an external commercial party, Research Value SRL. Due to commercial interests' authors were unable to assess the extraction procedure to validate accuracy. Authors were able to validate the extraction from Scopus by accessing the developers' platform from Elsevier. Since the authors were not able to compare the extraction accuracy it is possible that the difference in the number of articles from Scopus is attributed to a higher accuracy when querying Scopus. Third, the scholarly output from WoS was extracted in May 2018 and Scopus 9 months later. The effect on the number of papers would be minimal, however, the quality metrics might be affected due to

this delay since impact factor and FWCI rely on the total number of citations. It is, therefore, possible that the FWCI is positively skewed compared with IF.

Future research should address the topics mentioned above by aligning the extraction method of WoS and Scopus and perform extraction simultaneously. Additionally, by expanding the IRPES more THs can be included to improve generalizability on a national level. Next, a detailed study should be performed to analyze each document type separately, since reviews have, in general, a higher impact than most other document types such as articles, letters, notes [41]. Finally, other quality metrics can be included into the analysis to further contextualize FWCI with other quality indicators by looking at, but not limited to, cross-checking grants, collaborations with other research institutes, and percentage of papers publish in top 5 percentile journals.

5. Conclusion

To our knowledge, no prior research was performed to identify and highlight the differences of research performance of THs with respect to quantity and quality metrics using their published works while including a large sample of individual physicians. Utilizing secondary Big Data sources for performance management is, on the one hand, useful because they allow benchmarking at a national and international level, but on the other hand, using different data sources to measure the same construct of quality and quantity, clearly lead to different results when benchmarked against each other.

Research activities are an objective to be pursued and is part of the mission of both the Healthcare System as a whole and the providers who operate within the System. Among the providers, in the first place there are the THs, with their triple-fold mission of research, care and training. Following their mission, THs have an intrinsic motivation to deliver high performance on all three pillars. Measuring the performance of the research activities is essential but complex. Web-based tools allow to ensure a benchmarking process on different levels and can be effectively used at a Healthcare System level for different governance purposes such as planning, designing incentives for research, and allocating resources. Web-based tools have weaknesses and require a formal internal data and validation process to avoid self-referral. This can be overcome by setting up a transparent process shared among health professionals, hospital management and policymakers, which can contribute and in turn improve research performance.

Ethics approval

Not Applicable.

Consent to participate

Participating teaching hospitals in the network of measuring performance provided authors with their affiliated employed researchers. Final analysis was performed on an organizational level and employed researchers were not involved, contacted or analyzed on an individual level in any way during the study.

Consent for publication

Responsible region representatives have approved the final results.

Availability of data and material

The datasets, scripts or any other supplementary material used and analyzed during the current study are available from the corresponding author on reasonable request. Data obtained from *SciVal®* database, Elsevier B.V., http://www.scival.com

Funding

FH is working as a fellow in a project (www.healthPros-h2020.eu) that has received funding from the European Union's Horizon 2020 research and innovation programme under *the Marie Skłodowska-Curie* grant agreement No 765141. The overall project is partly financed by Italian regions within the IRPES.

Authors' contributions

Study conception was created by SN; study design was created by FH and DAL. Acquisition of data was performed by FH, DAL. Analysis and interpretation of data was performed by FH. Drafting of the manuscript was performed by FH and DAL. SN was involved in critical revisions of the manuscript and contributed in writing the background, discussion and conclusion paragraphs. All authors have read and approved the submitted manuscript.

Declaration of Competing Interest

The authors report no declarations of interest.

Acknowledgments

Authors would like to thank the participation of the regional network in providing us with input for data collection. This paper is part of a project (www.healthpros-h2020.eu) that has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 765141.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.healthpol.2020.10.002.

References

[1] Marr B, Gray D. Strategic performance management. Taylor Francis; 2006, 240 p.
[2] Moullin M. Performance measurement definitions: linking performance measurement and organisational excellence. International Journal of Health Care Quality Assurance 2007;20(3):181–3.
[3] Moullin M. Delivering excellence in health and social care: quality, excellence, and performance measurement. Open University Press; 2002.
[4] Bourne M, Mills J. Designing, implementing and updating performance measurement systems. International Journal of Operations Production & Management 2000;20(7):754–71.
[5] Kaplan D, Robert S, Norton D. The balanced scorecard: translating strategy into action. Boston: Harvard Business School Press; 1996.
[6] Neely A. The performance measurement revolution: why now and what next? International Journal of Operations Production & Management 1999;19(2):205–28.
[7] Kale S, Tamakuwala H, Vijayakumar V, Yang L, Rawal Kshatriya BS. Big data in healthcare: challenges and promise. In: Smart innovation, systems and technologies. Springer; 2020. p. 3–17.
[8] Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncology 2019;20:e262–73. Lancet Publishing Group.
[9] Porter ME. What is value in health care? New England Journal of Medicine [Internet] 2010;363(26):2477–81, http://dx.doi.org/10.1056/NEJMp1011024. Dec 23 [cited 2019 Sep 19]; Available from:.
[10] Gray M. Population healthcare: designing population-based systems. Journal of Royal Society Medicine [Internet] 2017;110(5):183–7, http://dx.doi.org/10.1177/0141076817703028. May 12 [cited 2019 Jul 3]; Available from:.
[11] Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Affairs 2008;27(3):759–69.
[12] Nuti S, Noto G, Vola F, Vainieri M. Let's play the patients music systems in healthcare. Management Decision 2018;56(10):2252–72.

[13] Nuti S. Making governance work in the health care sector : evidence from a 'natural experiment' in Italy. Policy Law 2016;11(February 2015):17–38.
[14] Bititci U, Garengo P, Nudurupati S. Performance measurement: challenges for tomorrow. International Journal of Management Reviews 2012;14:305–27.
[15] Bevan G. Reputations count : why benchmarking performance is improving health care across the world. Heal Economics Policy Law 2019;(14):141–61.
[16] Nuti S, Ruggieri T, Podetti S. Do university hospitals perform better than general hospitals? A comparative analysis among Italian regions. BMJ Open 2016;6(011426).
[17] Prasad V, Goldstein JA. US news and world report cancer hospital rankings: do they reflect measures of research productivity? PLoS One 2014;9(9):1–6.
[18] Krzyzanowska MK, Kaplan R, Sullivan R. How may clinical research improve healthcare: outcomes? Annals Oncology 2011;22(Suppl. 7):10–5.
[19] Majumdar SR, Chang WC, Armstrong PW. Do the investigative sites that take part in a positive clinical trial translate that evidence into practice? American Journal of Medicine 2002;113(November (2)):140–5.
[20] Kanavos P, Sullivan R, Lewison G, Schurer W, Eckhouse S, Vlachopioti Z. The role of funding and policies on innovation in cancer drug development. Ecancermedicalscience 2010;Vol. 4:1–139.
[23] Janni W, Kiechle M, Sommer H, Rack B, Gauger K, Heinrigs M, et al. Study participation improves treatment strategies and individual patient care in participating centers. Anticancer Research 2006;26(September (5 B)):3661–7.
[24] Birkmeyer JD, Stukel TA, Siewers AE, Goodney PP, Wennberg DE, Lucas FL. Surgeon volume and operative mortality in the United States. New England Journal of Medicine [Internet] 2003;349(November (22)):2117–27, http://dx.doi.org/10.1056/NEJMsa035205 [cited 2020 May 9]; Available from:.
[25] Birkmeyer JD, Dimick JB, Birkmeyer NJO. Measuring the quality of surgical care: structure, process, or outcomes?1 1No competing interests declared. Journal of American College Surgery 2004;198(April (4)):626–32.
[26] du Bois A, Rochon J, Pfisterer J, Hoskins WJ. Variations in institutional infrastructure, physician specialization and experience, and outcome in ovarian cancer: a systematic review. Gynecology Oncology 2009;112:422–36.
[27] Abramo G, D'Angelo CA. How do you define and measure research productivity? Scientometrics 2014;101(2):1129–44.
[28] Kreiman G, Maunsell JHR. Nine criteria for a measure of scientific output. Front Computing Neuroscience 2011;5(48):1–6.
[29] Dewett T, Denisi A. Exploring scholarly reputation it's more than just productivity. Scientometrics 2004;60(2):249–72.
[30] Broadus RN. Toward a definition of "bibliometrics". Scientometrics 1987;12(5–6):373–9.
[31] Gross EM. College libraries and chemical education. Science (80-) 1927;66(1713):385–9.
[32] Sternberg RJ, States U. Journal of applied research in memory and cognition evaluating merit among scientists. Journal of Applied Research Memoir Cognitive 2018;7(2):209–16.
[33] Grech V. Increasing importance of research metrics : journal Impact Factor and h-index H-index. International Urogynecological Association 2018;29:619–20.
[34] Cockriel WM, Mcdonald JB. The influence of dispersion on journal impact measures. Scientometrics 2018;116(1):609–22.
[35] Archambault E, Campbell D, Gingras Y, Larivière V. Comparing bibliometric statistics obtained from the web of science and Scopus. Journal of American Society Information Science Technology 2009;60(7):1320–6.
[36] Garfield E. Citation indexes for science. Science (80-) 1955;6:31–2.
[37] Garfield E. The history and meaning of the journal impact factor. Journal of American Medical Association 2006;295(1):1–4.
[38] Garfield E. How can impact factors be improved? BMJ 1996;313:411–3.
[39] Mutz R, Daniel H. Skewed citation distributions and bias factors : solutions to two core problems with the journal impact factor. Journal of Informatics 2012;6(2):169–76.
[40] Seglen P. The skewness of science. Journal of American Society Information Science 1992;43(9):628–38.
[41] Van Leeuwen T, Moed HF, Reedijk J. Critical comments on Institute for Scientific Information impact factors: a sample of inorganic molecular chemistry journals. Journal of Information Science 1999;25(6):489–98.
[42] Seglen PO. Why the impact factor of journals should not be used for evaluating research. BMJ 1997;314:498–513.
[43] Alberts B. Impact factor distortions. Science (80-) 2013;340(6134):787.
[44] Lariviere V, Kiermer VV, MacCallum CJ, Mcnutt M, Patterson M, Pulverer B, et al. A simple proposal for the publication of journal citation distributions. bioRxiv 2016:1–23.
[45] Minnick J. The 2018 JCR is here! Clarivate analytics; 2018.
[46] Narin F, Hamilton SK. Bibliometric performance measures. Scientometrics 1996;36(3):293–310.
[47] Research metrics guidebook. Elsevier; 2018.
[48] Haustein S. Grand challenges in altmetrics: heterogeneity, data quality and dependencies. Scientometrics 2016;108(1):413–23.
[49] Mintzberg H, Up Saddle River Structure in Fives; Designin Effective Organizations; 1992.
[50] Nuti S, D'Orio G, Gamba MP. Il sistema di valutazione della performance dei sistemi sanitari regionali; I risultati delle Aziende Ospedaliero-Universitarie a confronto; 2017.
[51] Nuti S, Ruggieri T. La valutazione della performance delle Aziende Ospedaliero-Universitarie. Finalità, metodi e risultati a confronto 2016:109.
[52] Vainieri M, Vola F, Gomez G, Nuti S. How to set challenging goals and conduct fair evaluation in regional public health systems. Insights from Valencia and Tuscany Regions. Health Policy (New York) 2016;120(11):1270–8.
[53] An eye on global research. Elsevier; 2018.
[54] Carloni M. Web of science core collection descriptive document; 2018.
[55] Haslam N, Laham SM. Quality, quantity, and impact in academic publication. European Journal of Society Psychology 2010;40:216–20.
[56] Franceschini F, Maisano D, Mastrogiacomo L. Empirical analysis and classification of database errors in Scopus and Web of Science. Journal of Informatics 2016;10(4):933–53.
[57] Franceschini F, Maisano D. The museum of errors / horrors in Scopus. Journal of Informatics 2016;10(1):174–82.
[58] López-cózar ED, Robinson-garcía N, Torres-salinas D. The google scholar experiment : how to index false papers and manipulate bibliometric indicators. Journal of Association Information Science & Technology 2014;65(3):446–54.
[59] Bartneck C, Kokkelmans S. Detecting h -index manipulation through self-citation analysis. Scientometrics 2011;87:85–98.
[60] Bevan G, Wilson D. Does 'naming and shaming' work for schools and hospitals? Lessons from natural experiments following devolution in England and Wales. Public Money Management [Internet] 2013;33(July (4)):245–52, http://dx.doi.org/10.1080/09540962.2013.799801 [cited 2019 Jul 26]; Available from:.
[61] Vainieri M, Lungu DA, Nuti S. Insights on the effectiveness of reward schemes from 10 - year longitudinal case studies in 2 Italian regions. International Journal of Heal Plan Management 2018;33(2):474–84.
[62] Nuti S, Bini B, Ruggieri TG, Piaggesi A, Ricci L, Grillo Ruggieri TG, et al. Bridging the gap between theory and practice in integrated care: the case of the diabetic foot pathway in Tuscany. International Journal of Integrative Care 2016;16(May (2)):9.