Availability of ORCIDs in publications archived in PubMed, MEDLINE, and Web of Science Core Collection

Christophe Boudry

Title Page

Author

Christophe Boudry, PhD [1,2]

Article title

Availability of ORCIDs in publications archived in PubMed, MEDLINE, and Web of Science Core Collection

Affiliations

[1] Normandie Univ, UNICAEN, Média Normandie. Esplanade de la Paix. 14032 Caen Cedex 5, France

[2] Unité régionale de formation à l'information scientifique et technique (URFIST), Ecole Nationale des Chartes, PSL Research University. 17, rue des Bernardins, 75005 Paris, France

ORCID

0000-0002-8730-8731

Corresponding author

C. Boudry

Média Normandie, Université de Caen Normandie

Esplanade de la Paix

CS 14032, 14032 Caen Cedex 5, France

E-mail: christophe.boudry@chartes.psl.eu

Declarations

Conflict of Interest: The author declare that he has no conflict of interest.

Acknowledgements

The author thanks ORCID’s support and Clarivate's product support and customer service for their assistance with this study. Thanks to Manuel Durand-Barthez for constructive comments on an earlier draft of this manuscript.

Abstract

The purpose of this paper was to assess ORCID availability in articles indexed in PubMed, MEDLINE (WoS Platform), and in the Web of Science Core Collection databases. The results showed an overall increase in the percentage of references with ORCIDs in these databases over time. Nevertheless, in PubMed over the period 2012-2020, only 13.9% of the articles had at least one ORCID, and only 4.3% of the authors had an ORCID. The analysis of journals indexed in PubMed show that only about half of all journals (51.6%) allow the use of ORCIDs in their articles during the submission process. The comparison of availability of ORCIDs in PubMed and MEDLINE show higher implementation of ORCIDs in MEDLINE due to differences in the methods used to collect ORCIDs (from publisher for PubMed and from ORCID registry for MEDLINE). These results suggest that entering ORCIDs by authors during the submission process is tedious and time consuming and hinders a larger presence of ORCIDs in PubMed. This study also shows that only using ORCIDs to collect researcher output is still unreliable in these bibliographic databases. This should convince decision-makers to establish recommendations encouraging all actors involved in research to consider more frequent use of ORCIDs.

Keywords

ORCID

Author identifier

PubMed

Web of Science

Country

Publisher

Introduction

The accurate identification of researchers and their scientific production is crucial for all actors involved in research (e.g. publishers, funders, universities, research evaluators, libraries) because many actions depend on the precision of this step (e.g. promotions, obtaining funds, publishing, or reviewing articles). Therefore, there are both individual and institutional benefits to the management of researcher identities (Craft 2020). Nevertheless, identifying authors and correctly attributing their work is still a challenging task through the current proliferation of online journal articles (Jinha 2010). Indeed, difficulties encountered in tracking scholarly and institutional publications are numerous and well known: disambiguating identical or similar names, name changes over time due to marriage or other circumstances, using aliases or author groups, and using different alphabets, abbreviation, or naming conventions (Fenner and Haak 2014; Granshaw 2019). Changes of researcher affiliations over time due to researcher mobility and/or lack of uniformity when declaring affiliations in articles are also well known difficulties (Mering 2017; Tran and Lyon 2017). To overcome these pitfalls, author identifiers have been developed. Scholarly repositories such as Research Papers in Economics (RePEc) in 1999, and arXiv in 2005 first included author identifiers (AIDs) (Warner 2010). Then, Scopus Author Identifier (ScopusID) in 2006 and Web of Science ResearcherID in 2008 were developed by bibliographic database providers (Elsevier and Thomson Reuters then Clarivate Analytics, respectively). The main drawbacks of these author identifiers is their link with a specific database, and consequently their lack of universality. To overcome this problem and offer an author identifier independent of scholarly repositories and bibliographic databases, Open Researcher & Contributor ID (ORCID) was launched in 2012. ORCID is an open, international, non-profit, cross-national, community-based project that is supported by its membership fees ("ORCID" 2020). ORCID allows researchers to enter any publication they wish into their profile and to control what data is entered. This author identifier has been widely promoted for its open source, cross-disciplinary and cross-national approach (Youtie et al. 2017). It has also been considered by some to increase the visibility of all scholarly activities of authors, reviewers, and editorial board members (Mašić et al. 2016). It is now the author identifier most used by researchers (Tran and Lyon 2017; Bello and Galindo-Rueda 2020) and is required by many services used by researchers (e.g. platform submissions, national or international agencies for grant-funding requests (Citrome 2016; Gasparyan et al. 2014)). Contrary to other author identifiers, this service is interoperable with most actors involved in research, allowing exchange of

information with other sites (e.g. CrossRef or Scopus for publications, or Publons for peer reviews) (Gasparyan et al. 2014; Arunachalam and Madhan 2016; L. L. Haak et al. 2018). This service has some flaws, however: such as the creation of the ORCID profile, which is not supervised or controlled. Researchers can create multiple profiles, leading to duplication. Furthermore, some authors have pointed out that ORCID is vulnerable to fraud and hacking (Leopold 2016). Inactive or outdated profiles have also been reported by (Memon and Azim 2019). Because privacy settings allow researchers to mask the content of their profiles, some profiles can be completely private (message "no public information is available") without content available (Craft 2020). Despite these limitations, this author identifier now seems to be in the best position to establish itself as a standard (Carter and Blanford 2017). Numerous articles have been published to explain the author identification problem and the usefulness of author identifiers. They describe how ORCID works (please see (Youtie et al. 2017) for a review of these articles). Some studies have been conducted to specifically assess the extent to which author identifiers are used by researchers (Mikki et al. 2015; Morgan and Eichenlaub 2018; Tran and Lyon 2017; Boudry and Durand-Barthez 2020).

ORCIDs have been integrated into some bibliographic databases (e.g. PubMed, Web of Science, and Scopus) to link authors to their output, allowing searches using ORCIDs as queries. This is especially important for bibliometric analysis (Butler 2012), but depends on the extent to which ORCIDs are implemented in bibliographic databases (Youtie et al. 2017). The analysis of availability of ORCIDs in bibliographic databases is thus important to assess whether using ORCIDs to track researchers' output is reliable. To the best of our knowledge, only one article has studied the implementation of ORCIDs in a bibliographic database (Web of Science) (Youtie et al. 2017).

The objectives of the present study are:

- to assess ORCID availability in articles indexed in the PubMed database, year by year, over the period 2012-2020 in the categories of articles and authors,

- to analyze the overall ORCID implementation levels in journals indexed in PubMed,

- to assess ORCID availability in articles indexed in the MEDLINE database using the WoS Platform provided by Clarivate Analytics, year by year, over the period 1966-2020 for the purposes of comparison with PubMed,

- to compare the ability of PubMed and MEDLINE to retrieve researchers' output using ORCIDs as queries,

- to assess ORCID availability in articles indexed in the Web of Science Core Collection (WoSCC) to analyze trends in disciplines and source types.

Materials and Methods

The number of existing ORCIDs from 2012 to 2020 were collected using ORCID annual reports found in the ORCID research repository ("ORCID research repository" 2020).

PubMed ("Home - PubMed - NCBI" 2020) is the most widely used bibliographic database in bio-medicine (Falagas et al. 2008). It has been available since 1996 and includes references from more than 5,200 scholarly journals published around the world. Its more than 30 million references include the MEDLINE database and other references such as books or references from PubMed Central. ORCIDs found in PubMed references are entered during the submission process by corresponding authors, and the distribution of article data (including ORCIDs) from publishers to PubMed is done automatically in Extensible Markup Language (XML) format (L. Haak et al. 2016).

The availability of ORCIDs in PubMed was assessed for categories of article and authors: references were randomly selected from PubMed according to their PubMed identifier (PMID) using a Hypertext Preprocessor language (PHP) function that generates random integers (the PHP random number generation function called mt_rand (min,max), which returns an integer between min and max). The minimum value used was 21,500,000, and the maximum value 32,226,052, allowing us to extract references published from 2011 to 2020 (32,226,052 corresponding to the highest PMID assigned in PubMed when the experiment was done). Then, the randomly generated integer was used to query PubMed by PMID using Efetch Entrez Programming Utilities to extract the reference of the corresponding article. Data were downloaded from PubMed in XML and were processed through developed PHP scripts. The collection of references was carried out from April 8th to April 20th, 2020. They were then imported to Microsoft Excel 2013 (Microsoft, Redmond, USA) for data processing as previously described (Boudry and Chartron 2017). For each article having one or more ORCIDs, the number was verified and their number was assessed. Then the number of authors, the date of publication, the name of the journal, and the country of publishers were analyzed. For the analysis of countries, England, Scotland, Northern Ireland, and Wales were grouped into the United Kingdom. The analysis was limited to the publication type "Journal

article". The ORCID registry was launched in October 2012 (L. L. Haak et al. 2012), but in PubMed there is no retrospective assignment of ORCIDs in indexed references. Therefore, the analysis was limited to the articles published from 2012 to 2020 (a check was done by analyzing 45,069 articles published in 2011, and no ORCIDs were found in these articles). An analysis of 508,934 randomly selected articles published from 2012 to 2020 was done (49,011 for 2012; 55,004 for 2013; 58,323 for 2014; 59,918 for 2015; 61,491 for 2016; 62,579 for 2017; 65,403 for 2018; 72,179 for 2019; 25,026 for 2020), corresponding to 5.87% of the articles included in PubMed within the same time span (n=8,666,054).

The list of currently indexed journals in PubMed was extracted from the NLM catalog ("Home - NLM Catalog - NCBI" 2020). A total of 5,248 journals was found. For each of the 5,248 journals the 100 most recently published articles referenced in PubMed were extracted using PHP scripts. Of these 5,248 journals, 196 did not have any articles indexed in PubMed and 4 had no authors. Because these 200 journals were not relevant for ORCID implementation analysis, they were not included in this study. Thus, a total of 5,048 journals and 504,800 articles were retained for analysis. For each of these 504,800 articles, the presence of one or more ORCIDs was assessed. The percentage of articles with ORCIDs and the percentage of authors with an ORCID was also evaluated for each journal.

The MEDLINE database is a subset of PubMed, searchable through search services such as Clarivate Analytics (WoS platform) that obtain data from the National Library of Medicine's (NLM) data distribution program ("MEDLINE, PubMed, and PMC (PubMed Central)" 2020). ORCIDs are integrated into MEDLINE in the same way as any of the databases proposed on the WoS platform, including MEDLINE and the Web of Science Core Collection (WoSCC). "ORCIDs are harvested from the ORCID registry and added to the Web of science platform monthly using metadata matching (author, journal, DOI…)" ("Web of Science: How ORCID works are matched to Web of Science records" 2020). ORCIDs are thus linked to the corresponding references in any of the databases on the Web of Science platform. ORCIDs can also be implemented from references present in ResearcherID profiles if ORCID profiles are public and associated with ResearcherID ("Web of Science: Inclusion of ORCID numbers" 2020). Please note that ORCIDs provided by the NLM data distribution program are not used in the MEDLINE WoS platform. Because collecting a very large number of references on the Web of science platform using PHP scripts is not possible and since the maximal number of references that can be uploaded is limited to 500, we were not able to assess the number of ORCIDs in the articles and authors categories. Thus, availability of

ORCIDs were assessed at a macro level, i.e. the percentage of references with at least one ORCID. To calculate this parameter, the search strategy used was "PY=(1966-2019) and AI=(0000*)" where PY stands for "Year Published" and AI for "Author Identifiers". Because the Web of Science platform retrospectively assigns ORCIDs to references using the ORCID registry, the restriction to articles published before 2012 was not applied. For the analysis and comparison of the entire group of databases (PubMed, MEDLINE, and WoSCC), the year 2020 was not included because it was incomplete when the analysis was done, and a bias could have been introduced in the results when including journals with variable indexing times (e.g. in PubMed, the time in indexing varies from a few days to several month)(Irwin and Rackham 2017). Therefore, the time span applied was 1966-2019, covering MEDLINE. The search strategy used to extract all references in the database with and without ORCIDs was "PY=(1966-2019)". All searches were limited to Journal Article as document type. For comparison purposes, the percentage of articles with at least one ORCID was also calculated in PubMed for the entire database. The query used to extract articles with at least one ORCID was "0000*[Author - Identifier] AND journal article [PT] AND 1966:2019[DP]" where PT stands for "Publication Type" and DP "Date of Publication". The search strategy used to extract all references in the database with and without ORCIDs was journal article [PT] AND 1966:2019[DP].

We compared the ability of PubMed and MEDLINE to retrieve researcher output using ORCIDs as queries, indirectly assessing the efficiency of these databases to implement ORCIDs in articles. To achieve this, 500 ORCIDs were randomly selected from those included in the 508,934 articles studied and were used as queries in PubMed, MEDLINE, and the ORCID registry. For each of these databases and each ORCID, the number of references found was manually collected. Each ORCID profile was checked to assess its confidentiality level (private versus public, private meaning that no public information was available).

Finally, in order to assess whether the implementation of ORCIDs varies according to discipline (sciences, social sciences, and art and humanities) and type of document (articles, reviews vs conference proceedings), the availability of ORCIDs in references indexed in the Web of Science Core Collection (WoSCC) provided by Clarivate Analytics was evaluated. The WoSCC is a multidisciplinary bibliographic database that includes five indexes referencing articles and proceedings: the Science Citation Index-Expanded (SCIE), the Social Science Citation Index (SSCI), the Arts & Humanities Citation Index (A&HCI), the Conference Proceedings Citation Index-Science (CPCI-S), and the Conference Proceedings

Citation Index-Social Science & Humanities (CPCI-SSH). The implementation of ORCIDs in these five indexes is done using the same methodology applied in MEDLINE and described above. However, the implementation of ORCIDs in these indexes can additionally be done from a publisher's PDF (when they are present). The search strategy used to extract all references with at least one ORCID in the five indexes studied was "PY=(1990-2019) and AI=(0000*)". The search strategy used to extract all references in the five indexes with and without ORCIDs was "PY=(1990-2019)". The analysis was restricted to "citable items" (Gorraiz et al. 2016): articles, proceeding papers, and reviews, and time span was 1990, covering the Web of Science Core Collection available in our institution.

Results

Number of articles and authors with ORCIDs in PubMed: temporal trends

We included 508,934 references in this study. Of these 508,934 references, 70,531 (13.9%) had at least one ORCID. The number of authors found was 3,012,625 (average of 5.9 authors per article), of which 139,912 (4.3%) had an ORCID. Considering all articles (with and without ORCIDs, n=508,934), the average number of ORCIDs per article was 0.26. When taking into account only articles with ORCIDs (n=70,531), the average number of ORCIDs per article was 1.8, and more than one-quarter of the authors had one ORCID (29.8%; 139,912 authors of 469,479).

As shown in Fig. 1, the number of existing ORCIDs found in the ORCID registry has been steadily increasing since its creation. The percentage of authors with ORCIDs (Fig. 1) and the percentage of articles with ORCIDs in PubMed has also been increasing, mainly from 2016 (Fig. 2). This shift is probably due to the time needed for the scientific community to discover, promote, and use ORCIDs during submission processes. In 2020, 38.2% of all articles had one or more ORCIDs (Fig. 2) and 12.1% of the authors had one ORCID in the PubMed references studied (Fig. 1).

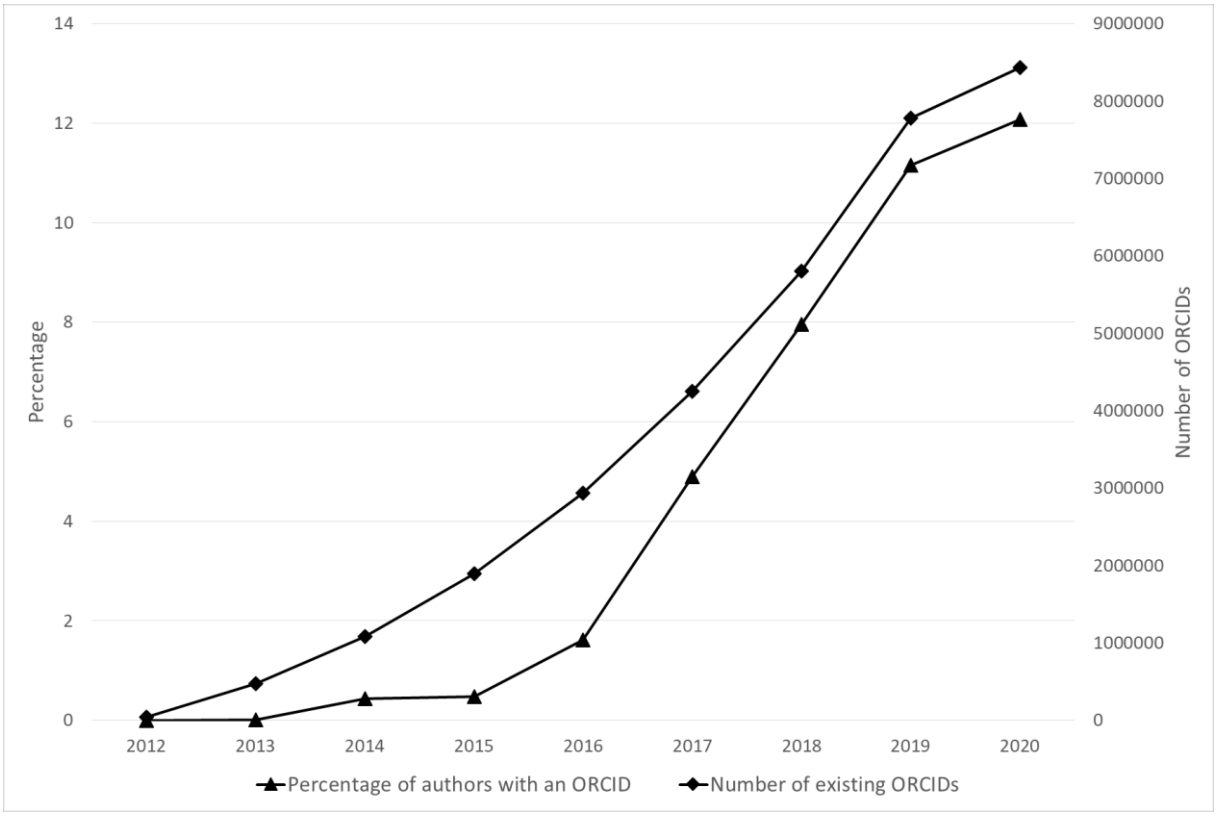


Fig. 1. Number of existing ORCIDs in the ORCID registry and percentage of authors with one ORCID in PubMed (2012-2020). Data were collected in April 2020.

Although the distribution of the number of authors was quite constant over the period studied (See supplementary Fig. 1), articles with few ORCIDs (mainly one ORCID, and to a lesser extent articles with 2 or 3 ORCIDs) have contributed more to the overall increase in the percentage of articles with ORCIDs compared to articles with more ORCIDs (Fig. 2). As an example, in 2020 (Fig. 2 and Supplementary Fig. 1), the percentage of articles with one author was 0.65%, whereas articles with one ORCID represented 22.1%. The percentage of articles with 10 or more authors was 35.3%, whereas articles with 10 or more ORCIDs represented only 0.33%. This shows that the number of authors of articles is inversely proportional to ORCIDs and indicates that the implementation of ORCIDs is not proportional to the number of authors during the submission process. There seem to be some obstacles during the submission process that prevent corresponding authors from exhaustively entering ORCIDs, especially when the number of authors is large.

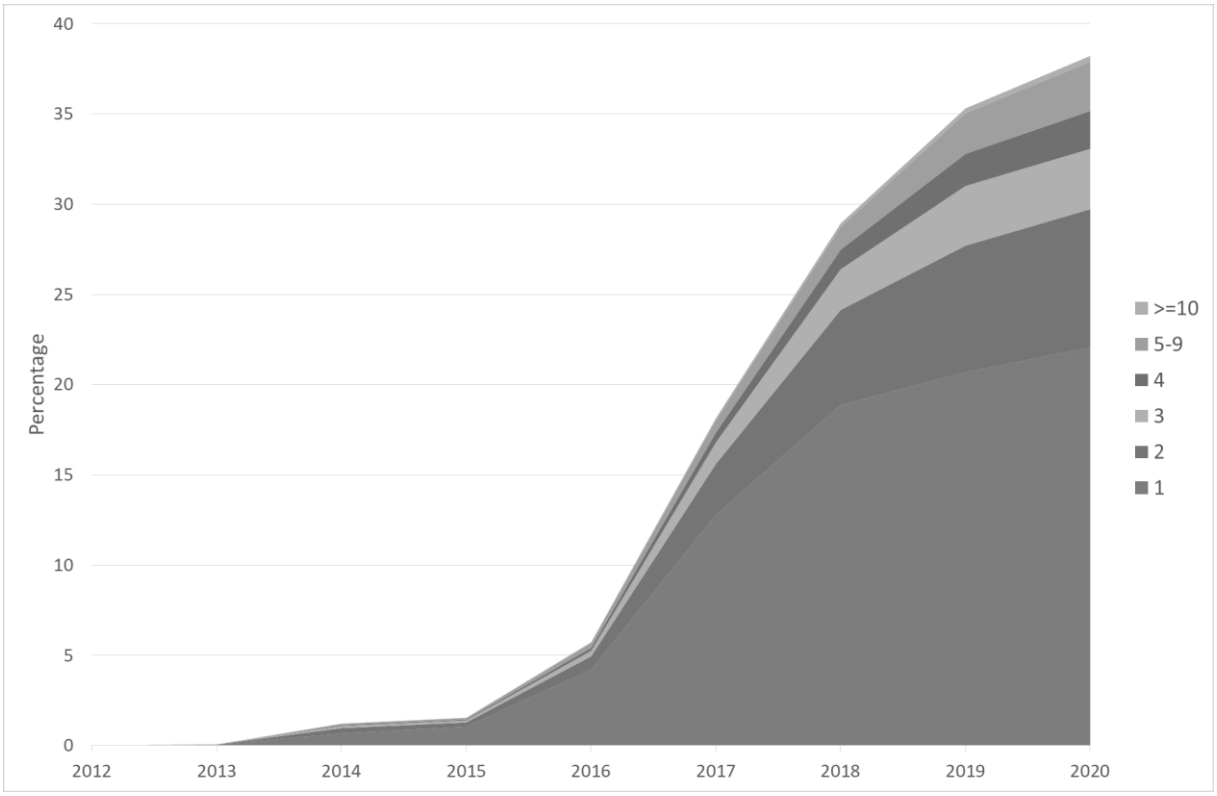


Fig. 2. Percentage of articles with ORCIDs (from 1 ORCID per article to more than 10 per article) (2012-2020)

Implementation of ORCIDs in PubMed: analysis of journals and geographical analysis of countries of publishers

Of the 5,048 journals studied, 2,604 (51.6%) had no ORCID in their 100 articles studied. This means that for these journals, there was probably no means for authors to enter ORCIDs during the submission process. For 2,444 (48.4%) journals, at least one article with one or several ORCIDs was found among the 100 articles studied, meaning that authors were allowed to enter ORCIDs during the submission process. Of these 2,444 journals, 523 (10.4%) had more than 90% of articles with at least one ORCID, and only 59 journals (1.2%) had at least one ORCID in each of the last 100 articles they had published (meaning that in this case entering at least one ORCID is mandatory when submitting an article in this journal).

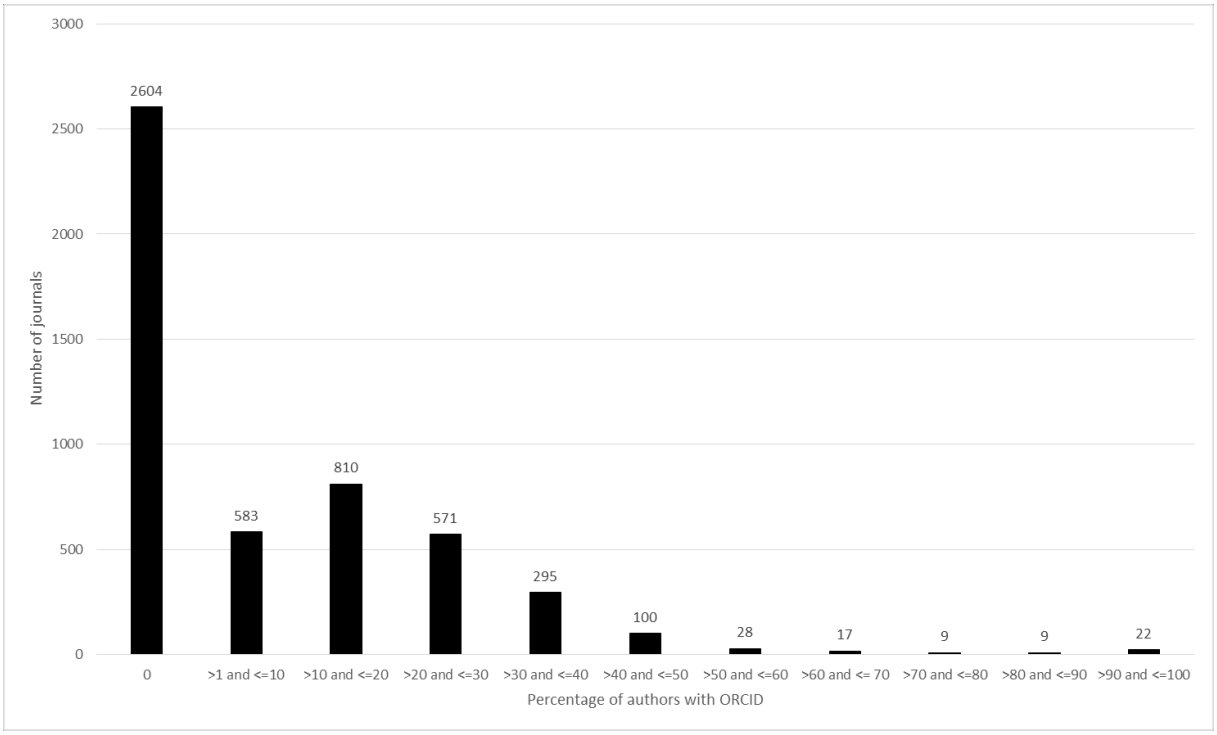


Fig. 3. Distribution of percentage of authors with ORCIDs in the 100 last articles published in the 5,048 journals studied

When considering the distribution of percentage of authors with ORCIDs in the 100 last articles published in the 5,048 journals studied, 2,604 journals (51.6%) did not include any authors having an ORCID. Only 85 (1.7%) of the journals included a percentage of authors with ORCIDs strictly amounting to higher than 50% (Fig. 3). This shows that the vast majority of journals do not encourage corresponding authors to enter as many ORCIDs as there are authors. It should be noted that only two journals (0.04% (*Revista latino-americana de enfermagem* and *Rev Saude Publica*) include 100% of authors with ORCIDs. This means that journals requiring an ORCID entry for each author during the submission process are extremely rare.

For the 5,048 journals studied, 71 countries of publishers were identified. As shown in Table 1, the percentage of journals allowing ORCID implementation is greatly variable with country of publishers and ranges from 0% for the Czech Republic (n=16 journals) to 91.3% for New Zealand (n=23 journals). One must note that in the ten countries of publishers with the highest percentages of ORCID implementation, nine have a national ORCID consortium. This suggests that having a national ORCID consortium raises publishers' awareness of ORCID, consequently increasing the number of journals implementing ORCIDs.

Table 1 Number of journals listed by countries of publishers (countries with more than 10 journals). Number and percentage of journals allowing ORCID implementation (with at least one ORCID implemented) listed by countries of publishers and ranked in decreasing order. National ORCID consortium indicates if an ORCID consortium exists in the country (Consortia are groups of 5 or more non-profit and/or governmental organizations taking a coordinated approach to ORCID implementation)

Country of publisher	Number of journals	Number of journals allowing ORCID implementation (%)	National ORCID consortium
New Zealand	23	21 (91.3)	Yes
Brazil	46	40 (87)	Yes
Denmark	29	22 (75.9)	Yes
Austria	12	9 (75)	Yes
Germany	275	202 (73.5)	Yes
United Kingdom	1344	949 (70.6)	Yes
Australia	74	49 (66.2)	Yes
Korea (South)	33	21 (63.6)	No
United States	1911	875 (45.8)	Yes
Sweden	10	4 (40)	Yes
Iran	11	4 (36.4)	No
Japan	110	40 (36.4)	Yes
Russia (Federation)	17	6 (35.3)	No
Singapore	12	4 (33.3)	No
Hungary	10	3 (30)	No
Canada	44	13 (29.5)	Yes
Switzerland	134	39 (29.1)	No
South Africa	11	3 (27.3)	Yes
Poland	41	11 (26.8)	No
Turkey	23	6 (26.1)	No
Netherlands	334	73 (21.9)	Yes
Italy	73	14 (19.2)	Yes
France	82	9 (11)	Yes
India	37	4 (10.8)	Yes
Greece	11	1 (9.1)	Yes
Mexico	11	1 (9.1)	No
China	79	5 (6.3)	No
Ireland	32	2 (6.3)	Yes
Spain	57	2 (3.5)	No
United Arab Emirates	40	1 (2.5)	No
Czech Republic	16	0 (0)	No

Number of ORCIDs in MEDLINE (WoS interface). Comparison with PubMed (entire database)

Over the period 1966-2019, the percentage of articles with at least one ORCID in MEDLINE was was 34.3% (8,829,890 articles with at least one ORCID among 25,773,086 references). This percentage was 3.9% for PubMed (entire database) over the same period (1,003,486 articles with at least one ORCID among the 25,954,423 references).

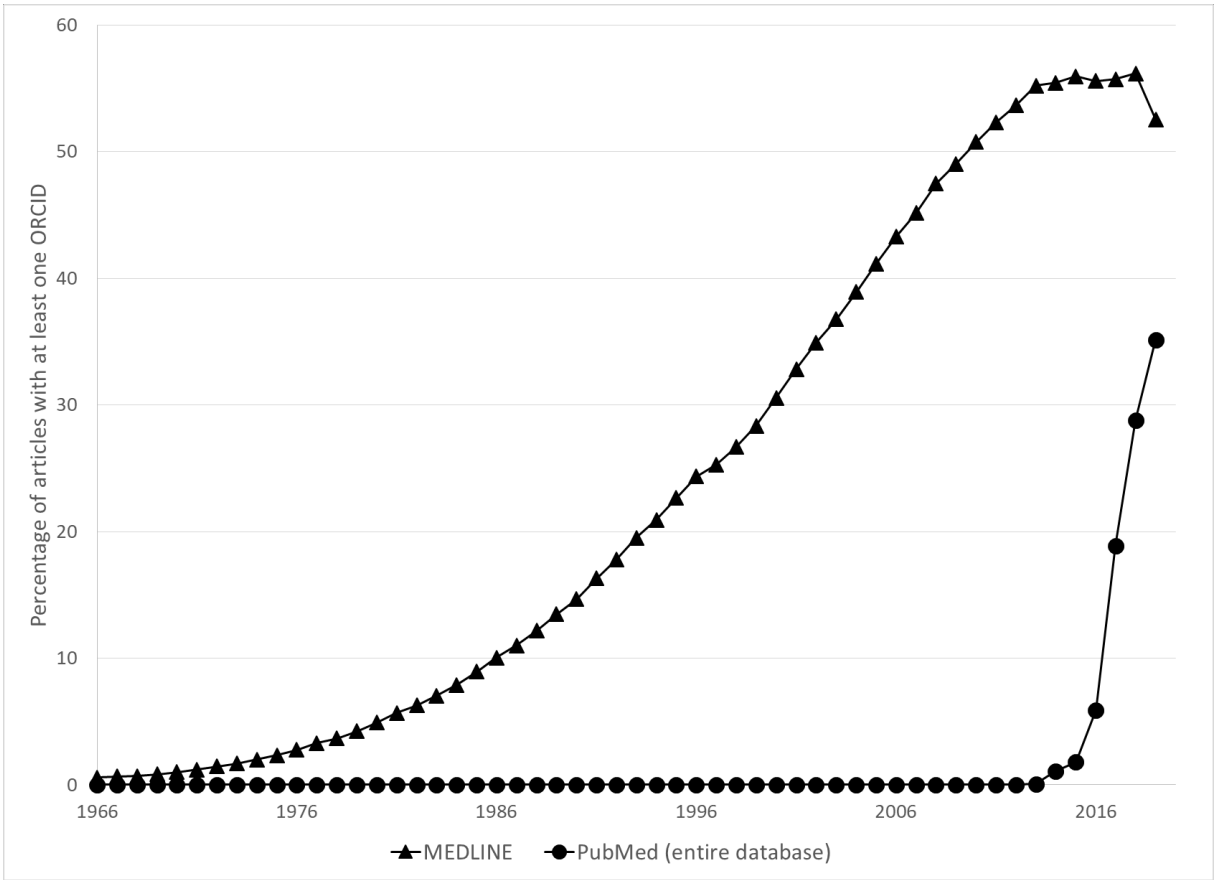


Fig. 4. Percentage of articles with at least one ORCID in MEDLINE and PubMed (entire database) (1966-2019)

Fig. 4. shows that the percentage of articles with at least one ORCID in MEDLINE is always higher than that of PubMed over the entire period studied. Thus, collecting ORCIDs from the ORCID registry (MEDLINE) is always more efficient compared to collecting ORCIDs entered by authors during the submission process (PubMed). Nevertheless, unlike PubMed, the percentage of articles with at least one ORCID decreased in MEDLINE for articles published in the last year (2019). This is probably due to the time needed for authors to manually enter a reference in their ORCID profile or the time needed (sometimes several

weeks) for Crossref to send articles to the ORCID registry after the publication date ("Auto-updates in third-party systems" 2018).

The results of the searches in PubMed and MEDLINE using 500 ORCIDs randomly selected from those included in the 508,934 articles studied are presented in Table 2.

Table 2 Number of articles per researcher found in PubMed, MEDLINE. The number of ORCIDs used as queries was 500

	Number of articles per researcher found in Pubmed	Number of articles per researcher found in MEDLINE
Mean	4.5	19.7
SD	6	45.1
Min	1	0
Max	50	523

In line, with results presented in Fig. 4, searching using ORCIDs as queries generally led to finding 4.4-fold more articles in MEDLINE compared to PubMed (average number of articles per researcher found in MEDLINE was 19.7 and was 4.5 in PubMed). It should be noted that, contrary to PubMed, a number of ORCIDs in MEDLINE (182; 36.4%) did not link to any article: 44 (8.8%) because the researchers did not enter any articles in their profile although the profile was public, 138 (27.6%) because the profile was private, and no public information was available.

Implementation of ORCIDs in the WOSCC

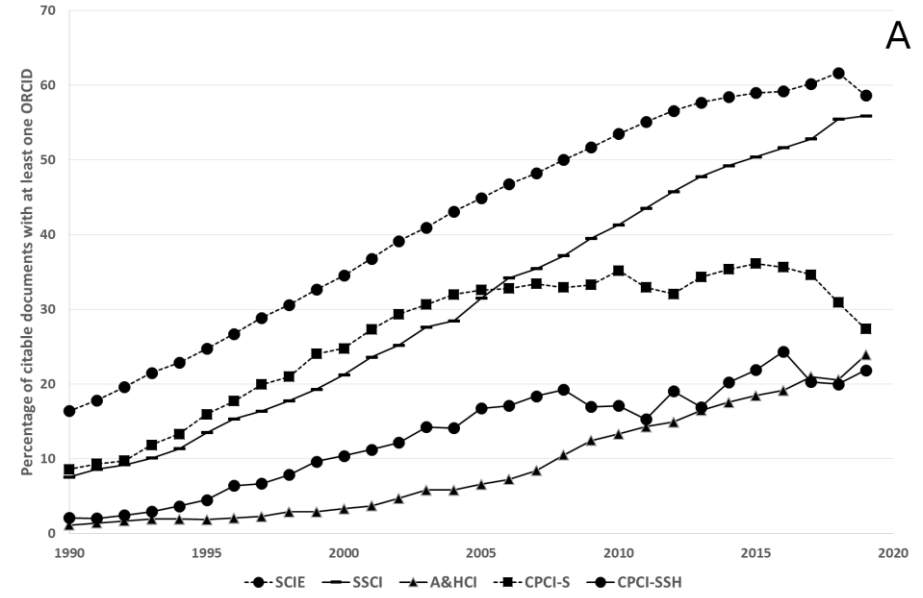
The results presented above relate only to one discipline (biology/medicine) and one document type (journal articles). To assess whether the implementation of ORCIDs in references indexed in bibliographic databases varies according to main disciplines (science, social sciences, and arts & humanities) and source types (contribution to journals or proceedings), the percentages of citable documents with at least one ORCID were evaluated in the indexes listed in Material and Methods. Over the period 1990-2019, the percentage of articles with at least one ORCID was 41.1% for the five indexes taken as a whole. For each index, these percentages were 46.9% for SCIE (13,652,553 articles with at least one ORCID among the 29,102,237 references), 38.8% for SSCI (1,511,582 articles with at least one ORCID among the 3,898,361 references), 10.5% for A&HCI (119,657 articles with at least

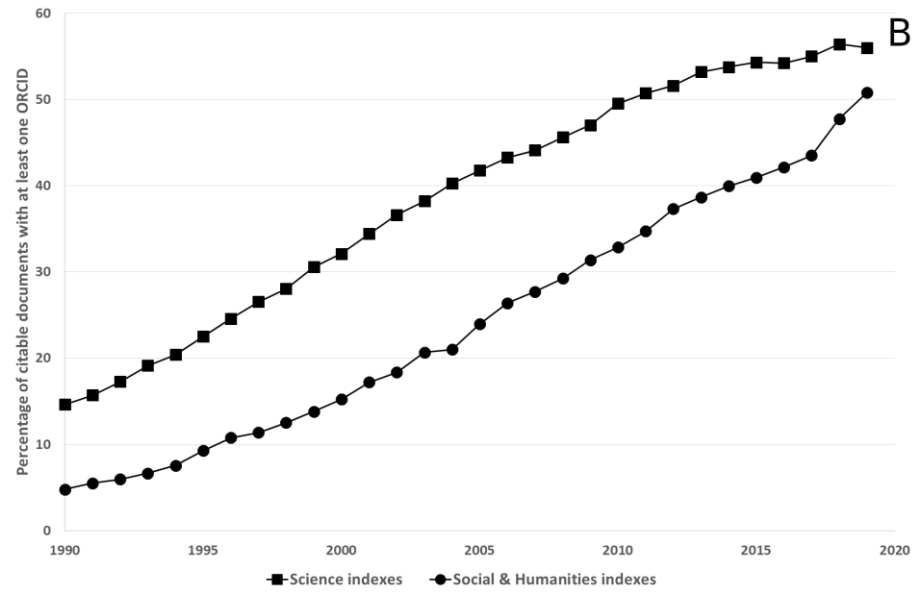
one ORCID among the 1,141,103 references), 28.4% for CPCI-S (2,264,063 articles with at least one ORCID among the 7,959,582 references), and 14.7% for CPCI-SSH (128,111 articles with at least one ORCID among the 868,720 references).

As shown in Fig. 5A, regardless of the index considered, the percentages of articles with at least one ORCID increased continuously from 1990 and reached more than 60% in 2018 for SCIE.

Fig. 5B shows the progression over time of the percentage of citable documents with at least one ORCID listed according to the main disciplines (science *vs.* social sciences and humanities). Results show the unequal distribution of ORCIDs between theses disciplines, with a higher implementation of ORCIDs in science compared to social sciences and humanities.

Fig. 5C shows the timeline of the percentage of citable documents with at least one ORCID grouped by source type (contributions to journals or proceedings) without disciplinary differentiation. From this point of view, the results show that the implementation of ORCIDs in references indexed in the five indexes studied, is more prevalent in journals compared to proceedings. It is noticeable that the percentage of citable documents in proceedings indexes remained almost constant over time from 2002, contrary to citable documents in journal indexes.


A
Percentage of citable documents with at least one ORCID
70
60
50
40
30
20
10
0
1990
1995
2000
2005
2010
2015
2020
SCIE
SSCI
A&HCI
CPCI-S
CPCI-SSH


B
Percentage of citable documents with at least one ORCID
60
50
40
30
20
10
0
1990
1995
2000
2005
2010
2015
2020
Science indexes
Social & Humanities indexes

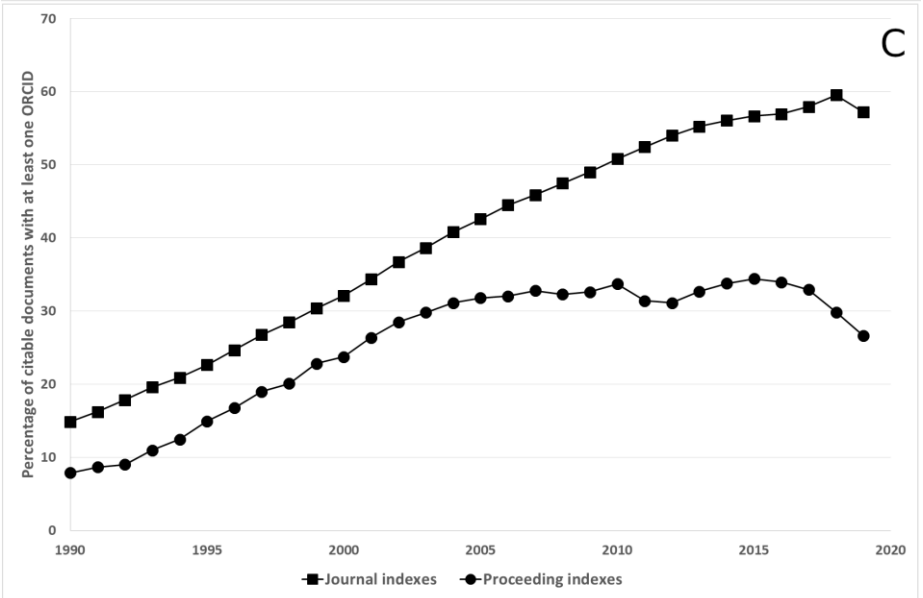

C
Percentage of citable documents with at least one ORCID
70
60
50
40
30
20
10
0
1990
1995
2000
2005
2010
2015
2020
Journal indexes
Proceeding indexes

Fig. 5. Percentage of citable documents (articles, proceeding papers, and reviews) with at least one ORCID (1990-2019). A: in SCIE, SSCI, A&HCI, CPCI-S, and CPCI-SSH. B: in Sciences indexes (SCIE and CPCI-S) and Social & Humanities indexes (SSCI, A&HCI and CPCI-SSH). C: in Journal indexes (SCIE, SSCI, and A&HCI) and Proceedings indexes (CPCI-S and CPCI-SSH)

Discussion

Implementation of ORCIDs in bibliographic databases has been rarely explored although it is a crucial and required step if we hope to find researcher output efficiently on these platforms using ORCIDs as queries. Only one study has assessed the availability of ORCIDs at a macro level using the percentage of articles with at least one ORCID in WoS without distinguishing between indexes, disciplines, or source types (Youtie et al. 2017). This study showed that 19% of articles indexed in WoS published between 2000 and 2016 have at least one associated ORCID and concluded that "gaps in coverage warn against relying solely on ORCID to collect researcher data". The authors also encourage future searches to "focus on the integration of ORCID information with WoS (or Scopus)", especially to "explain the anomalies in the presence or absence of ORCIDs in these indexes.

The aim of the present study was to assess the extent to which ORCIDs were implemented in PubMed, MEDLINE, and WoSCC databases, indirectly evaluating the possibilities offered by using ORCIDs to efficiently find researcher output on these platforms. To our knowledge, there are no similar studies examining the availability of ORCIDs in bibliographic databases. Furthermore, the methodology developed using PHP scripts to collect data on a large number of references, allowed us to access parameters that have never been evaluated before. These include the exact number of ORCIDs per article and the percentage of authors with ORCIDs in PubMed, or the percentage of articles with ORCIDs in journals indexed in PubMed.

This study had a number of limitations. The percentage of articles with at least one ORCID was the only parameter that could be calculated for MEDLINE and WoSCC databases. As pointed out by (Youtie et al. 2017), "the weakness of this macro-level search is that it is at the publication record level and not the author level" because solely the percentage of articles with at least one ORCID has been assessed. Thus, it was impossible to distinguish the implementation of ORCID among the authors of multi-authored papers and this methodology likely overestimates ORCID usage, because only one of the authors need to have an ORCID for the article to count as having an ORCID (Youtie et al. 2017). It could have been useful to

extend this analysis to the Scopus database (Elsevier), but unfortunately the wildcard operator was not available for the author identifier search field. Finally, this paper only examines the indexation of ORCIDs in bibliographic databases. Other points in the publishing and indexing process that are not studied in this paper may limit the presence of ORCIDs in bibliographic databases (e.g. some journals may be collecting ORCIDs but not passing them along to PubMed in their metadata submissions).

This study puts into evidence a relatively weak implementation of ORCIDs in articles indexed in the three bibliographic databases studied: over the period 1966-2019 the percentages of articles with at least one ORCID in MEDLINE and PubMed (entire database) were 34.3% and 3.9%, respectively, and over the period 1990-2019 the percentage of articles with at least one ORCID was 41.1% for the five indexes studied in WoSCC. The percentage of authors with ORCIDs in PubMed was also very low: 4.3% over the period 2012-2020. As a whole, with such weak percentages of articles with at least one ORCID and authors with ORCIDs, using solely ORCIDs to collect exhaustive researcher output is unreliable in these databases.

This study showed that entering ORCIDs by authors during the submission process is a bottleneck, limiting a larger presence of ORCIDs in PubMed. Indeed, the greater the number of authors of articles, the fewer the number of proportionally entered ORCIDs during the submission process. This is largely due to journals focusing their ORCID integration on the corresponding author even though a growing number of journals enable ORCID integration for co-authors ("Requiring ORCID in Publication Workflows" 2015). This can be also explained by the difficulty corresponding authors encounter when entering a large number of ORCIDs, a very tedious and time consuming process which can be considered as an undue burden (L. Haak et al. 2016). This is all the more true as the number of authors per article increases. If most corresponding authors are willing to collect 2 or 3 ORCIDs from their co-authors and enter these ORCIDs during the submission process, most of them understandably give up as the number of co-authors grows larger, if entering ORCIDs is not mandatory. To help authors enter co-authors' ORCIDs more easily, submission platforms should be linked to the ORCID registry so that the corresponding author's existing ORCID could appear automatically when their name is entered. Furthermore, as suggested by (L. Haak et al. 2016), instead of focusing the collection of ORCIDs during the submission process, which loads the burden on authors all at once, ORCID collection could be enabled through XML workflows as research is being carried out, e.g. with e-lab notebooks. Effort should also be made to increase the awareness of researchers so that it will not come as a surprise during the

submission process if entering ORCIDs is possible or mandatory (L. Haak et al. 2016). These suggestions could facilitate the implementation of more ORCIDs in PubMed.

The second limitation of the widespread implementation of ORCIDs in PubMed is that authors in more than half of the journals indexed in PubMed (51.6%) were not allowed to enter ORCIDs during the submission process. Furthermore, the results presented in this study suggest that the publishers' level of interest in ORCIDs depends on their geographical location. Targeted efforts must therefore be undertaken in countries where publishers implement ORCIDs least often in order to encourage these publishers to change their strategy regarding ORCIDs. As suggested by the results of this study, encouraging countries to create a national ORCID consortium should be part of these efforts.

The great majority of journals indexed in PubMed do not encourage corresponding authors to always enter one ORCID per author, and when entering ORCIDs is possible, this step is rarely mandatory: for only 1.2% of journals, entering at least one ORCID in each article published was mandatory to submit an article, and only 0.04% of journals required entering ORCIDs for every author. Submission platforms could play a particularly important role in urging journals to require ORCIDs during the submission process. Editorial strategies of journals could thus be upgraded to require authors to enter ORCIDs. For the moment, the two major submission platforms (Editorial Manager/Aries systems and ScholarOne Manuscripts/Clarivate analytics) let the journals choose whether mandatory entry of one or several ORCIDs will be requested during the submission processes ("Support for ORCID in Publishing Systems | ORCID Members" 2020). An intermediate step, before requiring entry of ORCIDs for every author, could be for all journals present on these two submission platforms to require entering at least one ORCID per article for each article submitted ("Requiring ORCID in Publication Workflows" 2015). This could have a significant influence on the journals not assigning ORCIDs and could encourage authors to go further in entering ORCIDs during the submission process.

This study has shown that the availability of ORCIDs differed between PubMed (entire database) and MEDLINE over the period 1966-2019: 3.9% of the articles had at least one ORCID in PubMed versus 34.3% in MEDLINE. Furthermore, using ORCIDs as queries to find researcher output leads to finding an average of 4.5 articles per researcher in PubMed and 19.7 in MEDLINE. These differences can be explained by the methods used to collect ORCIDs by PubMed and MEDLINE. ORCIDs found in PubMed references are those entered by corresponding authors during the submission process (L. Haak et al. 2016), whereas in

MEDLINE ORCIDs are collected from the ORCID registry and added monthly to the Web of Science platform using metadata matching (author, journal, DOI…) ("Web of Science: How ORCID works are matched to Web of Science records" 2020). ORCIDs can also be implemented from references present in ResearcherID profiles if ORCID profiles are public and associated with ResearcherID ("Web of Science: Inclusion of ORCID numbers" 2020). As shown in this study, collecting ORCIDs from the ORCID registry has the advantage of allowing the retrospective implementation of ORCIDs and enabling the number of ORCIDs to grow over time, which is not possible when ORCIDs are entered by authors during the submission process. Compared to results obtained by (Youtie et al. 2017), even though the period studied is not equivalent, our results seem to confirm the increasing implementation of ORCIDs as time goes by: 19% of the articles have at least one ORCID from 2000 to 2016 in (Youtie et al. 2017), and 41.1% have ORCIDs from 1990 to 2019 in this study. The other advantage of collecting ORCIDs from the ORCID registry is that by allowing access to all existing ORCIDs, more ORCID profiles will be available over time and more articles with ORCIDs will be present in MEDLINE. Nevertheless, an important limitation of collecting ORCIDs in the ORCID registry results from empty and private profiles. The results of this study showed that 8.8% of ORCIDs did not link to any article and 27.6% were private. Each of these empty or private profiles are clearly a limitation to the implementation of ORCIDs in MEDLINE. Moreover, this is also an obstacle for all parties involved in research when they are accessing information needed about researchers. Private profiles also prevent automatic data exchange with other actors involved in research. In order to minimize this problem, it could be useful to encourage researchers to make their profiles public. To achieve this goal, having a public profile could be mandatory when submitting articles or using services requiring an ORCID. Providing tools for authors in bibliographic databases allowing them to link their articles without ORCIDs to their ORCID profiles should also be promoted. This has the advantage of increasing the number of articles present in the ORCID profiles of researchers, consequently increasing the number of articles with ORCIDs in bibliographic databases that harvest data from the ORCID registry. The "ORCID article claiming" tool, developed and proposed by Europe PMC, seems to be a good example to follow for implementation in other bibliographic databases (Rossiter 2013; The Europe PMC Consortium 2015).

The analysis of the five WoSCC indexes studied has shown that the availability of ORCIDs is very variable, with more than 60% of the articles having at least one ORCID for SCIE and 20% for A&HCI in 2018. Differences also appear when grouping indexes by discipline

(science *vs.* social sciences and humanities), with more implementation of ORCIDs in sciences compared to social & humanities. These results are in accordance with those obtained by (Mikki et al. 2015) and can probably be explained by the fact that social sciences and humanities services using ORCID are less numerous than in science (e.g. submission platform). Therefore, the knowledge of researchers in social sciences and humanities with respect to ORCIDs is less extensive than researchers in sciences. The implementation of ORCIDs was lower in proceedings indexes compared to journal indexes. This can be explained by the automatic integration of proceedings into ORCID profiles by CrossRef, which are less numerous compared to articles because the proportion of proceedings with DOIs is smaller (Gorraiz et al. 2016) . We can also put forward the hypothesis, without being certain, that a higher proportion of articles than proceedings are entered in ORCID profiles because researchers may be less motivated to enter proceedings in their profiles.

To conclude, this study showed that solely using ORCIDs to collect researcher output is still unreliable in the three bibliographic databases studied, particularly in PubMed. Nevertheless, the availability of ORCIDs continues to increase over time, and it is very important to advance in this direction regardless of the solutions implemented: collecting ORCIDs during the submission process, collecting ORCIDs from the ORCID registry or other Author Identifiers (e.g. ResearcherID), or even, improving journal and indexing processes. Results presented in this study should convince decision-makers to establish recommendations encouraging editorial services, publishers, institutions, and professional and scholarly associations to consider more frequent use of ORCIDs.

References

Arunachalam, S., & Madhan, M. (2016). Adopting ORCID as a unique identifier will benefit all involved in scholarly communication. *National Medical Journal of India*, *29*(4), 227–234.

Auto-updates in third-party systems: Crossref. (2018). *ORCID*. https://support.orcid.org/hc/en-us/articles/360006971293. Accessed 5 November 2020

Bello, M., & Galindo-Rueda, F. (2020). Charting the digital transformation of science: Findings from the 2018 OECD International Survey of Scientific Authors (ISSA2).

Documents de travail de l'OCDE sur la science, la technologie et l'industrie, (2020/03). https://doi.org/10.1787/1b06c47c-en

Boudry, C., & Chartron, G. (2017). Availability of digital object identifiers in publications archived by PubMed. *Scientometrics*, *110*(3), 1453–1469. https://doi.org/10.1007/s11192-016-2225-6

Boudry, C., & Durand-Barthez, M. (2020). Use of author identifier services (ORCID, ResearcherID) and academic social networks (Academia.edu, ResearchGate) by the researchers of the University of Caen Normandy (France): A case study. *PLOS ONE*, *15*(9), e0238583. https://doi.org/10.1371/journal.pone.0238583

Butler, D. (2012). Scientists: your number is up. ORCID scheme will give researchers unique identifiers to improve tracking of publications. *Nature News*, *485*(7400), 564. https://doi.org/10.1038/485564a

Carter, C. B., & Blanford, C. F. (2017). All authors must now supply ORCID identifiers. *Journal of Materials Science*, *52*(11), 6147–6149. https://doi.org/10.1007/s10853-017-0919-7

Citrome, L. (2016). Open researcher and contributor ID: ORCID now mandatory for Wiley journals. *International Journal of Clinical Practice*, *70*(11), 884–885. https://doi.org/10.1111/ijcp.12912

Craft, A. R. (2020). Managing Researcher Identity: Tools for Researchers and Librarians. *Serials Review*, *46*(1), 44–49. https://doi.org/10.1080/00987913.2020.1720897

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, *22*(2), 338–342. https://doi.org/10.1096/fj.07-9492LSF

Fenner, M., & Haak, L. (2014). Unique identifiers for researchers. In *Opening science* (pp. 293–296). Springer. http://link.springer.com/chapter/10.1007/978-3-319-00026-8_21. Accessed 23 January 2017

Gasparyan, A. Y., Akazhanov, N. A., Voronov, A. A., & Kitas, G. D. (2014). Systematic and open identification of researchers and authors: focus on open researcher and contributor ID. *Journal of Korean Medical Science*, *29*(11), 1453–1456. https://doi.org/10.3346/jkms.2014.29.11.1453

Gorraiz, J., Melero-Fuentes, D., Gumpenberger, C., & Valderrama-Zurián, J.-C. (2016). Availability of digital object identifiers (DOIs) in Web of Science and Scopus. *Journal of Informetrics*, *10*(1), 98–109. https://doi.org/10.1016/j.joi.2015.11.008

Granshaw, S. I. (2019). Research identifiers: ORCID, DOI, and the issue with Wang and Smith. *Photogrammetric Record*, *34*(167), 236–243. https://doi.org/10.1111/phor.12290

Haak, L., Donohoe, P., Kiermer, V., Atkins, H., Lees-Miller, J., & Raybould, C. (2016). *ORCID iD Throughput in Publishing Workflows*. *Journal Article Tag Suite Conference (JATS-Con) Proceedings 2016 [Internet]*. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK350150/. Accessed 14 May 2020

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: a system to uniquely identify researchers. *Learned Publishing*, *25*(4), 259–264. https://doi.org/10.1087/20120404

Haak, L. L., Meadows, A., & Brown, J. (2018). Using ORCID, DOI, and Other Open Identifiers in Research Evaluation. *Frontiers in Research Metrics and Analytics*, *3*. https://doi.org/10.3389/frma.2018.00028

Home - NLM Catalog - NCBI. (2020). https://www.ncbi.nlm.nih.gov/nlmcatalog. Accessed 28 April 2020

Home - PubMed - NCBI. (2020). https://www.ncbi.nlm.nih.gov/pubmed/. Accessed 28 April 2020

Irwin, A. N., & Rackham, D. (2017). Comparison of the time-to-indexing in PubMed between biomedical journals according to impact factor, discipline, and focus. *Research in social & administrative pharmacy: RSAP*, *13*(2), 389–393. https://doi.org/10.1016/j.sapharm.2016.04.006

Jinha, A. E. (2010). Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, *23*(3), 258–263. https://doi.org/10.1087/20100308

Leopold, S. S. (2016). Editorial: ORCID is a Wonderful (But Not Required) Tool for Authors. *Clinical Orthopaedics and Related Research*, *474*(5), 1083–1085. https://doi.org/10.1007/s11999-016-4760-0

Mašić, I., Begić, E., Donev, D. M., Gajović, S., Gasparyan, A. Y., Jakovljević, M., et al. (2016). Sarajevo Declaration on Integrity and Visibility of Scholarly Publications. *Croatian Medical Journal*, *57*(6), 527–529. https://doi.org/10.3325/cmj.2016.57.527

MEDLINE, PubMed, and PMC (PubMed Central): How are they different? (2020). FAQs, Help Files, Pocket Cards, U.S. National Library of Medicine. https://www.nlm.nih.gov/bsd/difference.html. Accessed 14 May 2020

Memon, A. R., & Azim, M. E. (2019). Open Researcher and Contributor Identifier and other author identifiers: Perspective from Pakistan. *Journal of the Pakistan Medical Association*, *69*(6), 888–891.

Mering, M. (2017). Correctly Linking Researchers to Their Journal Articles: An Overview of Unique Author Identifiers. *Serials Review*, *43*(3–4), 265–267. https://doi.org/10.1080/00987913.2017.1386056

Mikki, S., Zygmuntowska, M., Gjesdal, Ø. L., & Al Ruwehy, H. A. (2015). Digital Presence of Norwegian Scholars on Academic Network Sites-Where and Who Are They? *PloS One*, *10*(11), e0142709. https://doi.org/10.1371/journal.pone.0142709

Morgan, M., & Eichenlaub, N. (2018). Author identifier analysis: Name authority control in two institutional repositories. In *Proceedings of the International Conference on Dublin Core and Metadata Applications* (Vol. 2018-September, pp. 113–116). Presented at the Proceedings of the International Conference on Dublin Core and Metadata Applications.

ORCID. (2020). https://orcid.org/. Accessed 17 May 2020

ORCID research repository. (2020). https://orcid.figshare.com/browse. Accessed 22 April 2020

Requiring ORCID in Publication Workflows: Open Letter. (2015). https://orcid.org/content/requiring-orcid-publication-workflows-open-letter. Accessed 15 May 2020

Rossiter, P. (2013). Linking Articles Available in Europe PMC to your ORCID. http://blog.europepmc.org/2013/08/linking-articles-available-in-europe.html. Accessed 27 October 2020

Support for ORCID in Publishing Systems | ORCID Members. (2020). https://members.orcid.org/api/vendors/publisher-tools. Accessed 15 May 2020

The Europe PMC Consortium. (2015). Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic acids research*, *43*(Database issue), D1042-8. https://doi.org/10.1093/nar/gku1061

Tran, C. Y., & Lyon, J. A. (2017). Faculty use of author identifiers and researcher networking tools. *College and Research Libraries*, *78*(2), 171–182. https://doi.org/10.5860/crl.78.2.171

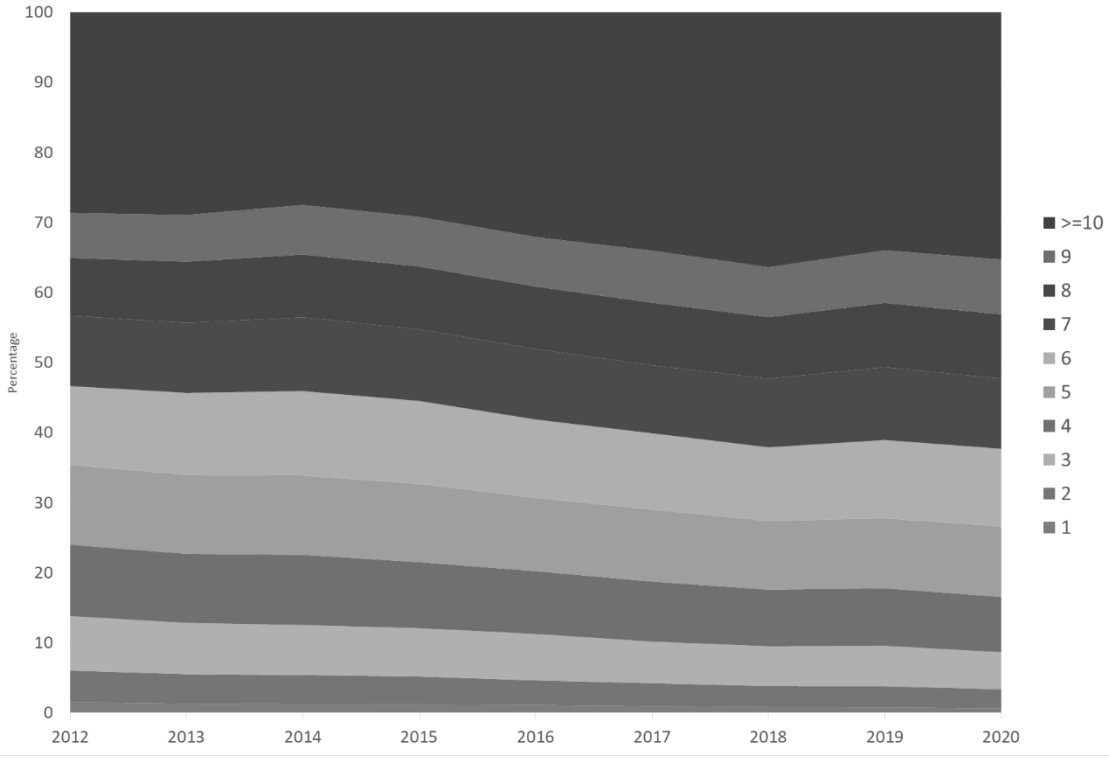
Warner, S. (2010). Author identifiers in scholarly repositories. *Journal of Digital Information*, *11*(1), 1–10.

Web of Science: How ORCID works are matched to Web of Science records. (2020). https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-How-ORCID-works-are-matched-to-Web-of-Science-records?language=en_US&r=5&ui-force-components-controllers-recordGlobalValueProvider.RecordGvp.getRecord=1. Accessed 30 April 2020

Web of Science: Inclusion of ORCID numbers. (2020). https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Inclusion-of-ORCID-numbers?language=en_US. Accessed 30 April 2020

Youtie, J., Carley, S., Porter, A. L., & Shapira, P. (2017). Tracking researchers and their outputs: new insights from ORCIDs. *Scientometrics*, *113*(1), 437–453. https://doi.org/10.1007/s11192-017-2473-0

Supplementary material



Supplementary Fig 1. Number of authors per article (2012-2020) for the 508,934 references studied in PubMed