Scientometrics
https://doi.org/10.1007/s11192-020-03845-3

Accuracy of *PubMed*-based author lists of publications and use of author identifiers to address author name ambiguity: a cross-sectional study

Paul Sebo[1] · **Sylvain de Lucia**[2] · **Nathalie Vernaz**[3]

Received: 16 August 2020 / Accepted: 15 December 2020

Abstract
Objective: To assess the accuracy of *PubMed*-based author lists of publications and use of author identifiers to address author name ambiguity. *Methods*: In this Swiss study conducted in 2019, 300 hospital-based senior physicians were asked to generate a list of their publications in *PubMed* and complete a questionnaire (type of query used, number of errors in their list of publications, knowledge and use of *ORCID* and *ResearcherID*). *Results:* 156 physicians (52%) agreed to participate, 145 of whom published at least one article (mean number of publications: 60 (SD 73)). Only 17% used the advanced search option. On average, there were 5 articles in the lists that were not co-authored by participants (advanced search: 1.0 (SD 2.6) vs. 5.9 (SD 13.9), *p* value 0.02) and 3 articles co-authored by participants that did not appear in the lists (advanced search: 1.5 (SD 2.0) vs. 3.6 (SD 8.4), *p*-value 0.05). Although 82% were aware of *ORCID*, only 16% added all their articles (39% and 6% respectively for *ResearcherID*). *Conclusions:* When used by senior physicians, the advanced search in *PubMed* is accurate for retrieving authors' publications. Author identifiers are only used by a minority of physicians and are therefore not recommended in this context, as they would lead to inaccurate results.

Keywords Accuracy · List of publications · *PubMed* · Author identifier · *ORCID* · *ResearcherID* · Author name ambiguity · Homonymy · Synonymy

✉ Paul Sebo
paulsebo@hotmail.com

[1] Primary Care Unit, University of Geneva, Geneva, Switzerland

[2] Department of Community Health and Medicine, Geneva University Hospital, Geneva, Switzerland

[3] Medical and Quality Directorate, Geneva University Hospital, Geneva, Switzerland

Published online: 02 March 2021

Introduction

Writing scientific papers is not always an easy task, especially for inexperienced authors. However, publication of research is crucial, because it allows to disseminate scientific knowledge (Mabe 2010; Lafrenière et al. 2013) and increase the recognition of researchers (Vale 2015; Bavdekar and Tullu 2016; Post et al. 2012).

It may be important to track researchers' publications over time, for example in order to explore various issues linked to research policy, such as the role of gender, cross-institutional collaboration or funding on publication trajectories (Lerchenmueller and Sorenson 2016; Akers et al. 2016; Gasparyan et al. 2014, 2016). In biomedical and health sciences fields, this type of search is usually done with *PubMed*, a free search engine launched in 1996 that provides full access to MEDLINE® (Fiorini et al. 2018). Today, it has about 2.5 million daily users worldwide (Fiorini et al. 2018).

With the spread of scientific literature, author name ambiguity has been recognized as a growing issue and has become crucial when dealing with information on individual researchers (Akers et al. 2016; Gasparyan et al. 2016). Indeed, significant difficulties are often encountered when database searches are compiled and researchers' publications are tracked due to shared or ambiguous author names. There are two aspects of name ambiguity: name homonymy (i.e. several authors share the same name) and name synonymy (i.e. an author has several names) (Liu et al. 2014). Queries with a homonymous name often lead to irrelevant results. However, name synonymy is also a frequent problem in the scientific literature; for example, a Spanish study found that about half of Spanish authors were listed under more than one name (Ruiz-Pérez et al. 2002).

The *ORCID* (Open Researcher and Contributor ID) registry was launched in October 2012 in order to address author name ambiguity by providing researchers with a unique 16-character author identifier (Akers et al. 2016; Gasparyan et al. 2014, 2016, 2017); attaching a unique identifier to their publications may mitigate the risk of misattribution and provide disambiguated search results. However, with inevitable indexing errors, this risk is not completely eliminated. Researchers can maintain the same identifier throughout their career, even if their institutional affiliation or their name changes over time, e.g. due to marriage (Akers et al. 2016). *ORCID* is free for individual researchers and relatively easy to use; researchers can manually add their publications to their profile or use external databases to import them. In August 2018, the number of registered *ORCID* accounts was more than 5 million. *ResearcherID* is another author identifying system that was introduced in January 2008 by Thomson Reuters (Gasparyan et al. 2017). It has been criticized for being commercial and proprietary, in contrast to *ORCID*. *ResearcherID* and *ORCID* enable data exchange between their databases.

Given the existing limitations in retrieving authors' lists of publications from databases and the recent launch of author identifying systems, we aimed to assess the accuracy of *PubMed*-based author lists of publications. To do this, we asked hospital-based senior physicians to generate a list of their own publications and then report any errors. We also aimed to assess the use by these physicians of author identifiers (*ORCID* and *ResearcherID*) to address author name ambiguity.

Methods

Study site and study population

This study relied on a self-completed online survey. It was conducted in 2019 in the French-speaking part of Switzerland, where there is a large 2000-bed teaching hospital (Geneva University Hospital). The Geneva university Hospital is the largest hospital in the country (around 14,000 employees, 17% of whom are physicians). We selected a random sample of 300 physicians from among the 1173 hospital-based senior physicians (heads of division, staff physicians or senior registrars). They were invited to participate by e-mail. This e-mail included information on the aim of our study and practical procedures for completing the online survey. Reminder messages (one per physician) were sent to all physicians who did not answer to our questionnaire within 1 month. These reminder messages were personalised e-mails sent by the investigator who knew each of these physicians best. Those who have not published at least one peer-reviewed article were excluded from the survey (they were instructed to record this information in the first question of the survey and the survey was completed).

Data collection

Participating physicians were asked to generate a list of their publications in *PubMed*. To do so, they were asked to proceed as they usually do when conducting this type of search for other authors in the context of their clinical or research activities. In our study, we planned to ask the participants themselves to generate the lists of publications because our aim was to estimate in a pragmatic way the accuracy of search carried out by senior physicians who are not librarians and whose search strategies are not necessarily optimal. If we had asked librarians to generate these lists and then asked participants to check them, we would probably have underestimated the number of errors *in real life*.

The physicians were also asked to complete a questionnaire about socio-demographic characteristics (gender, age group, department, nationality, and average number of hours worked per week for clinical and research activity), the number of peer-reviewed publications and the type of author name query used in *PubMed* to retrieve the list of their publications (PubMed Help 2019). The response options for this question were: by entering their first and last name in the search box with/without quotation marks, by entering their last name plus initials in the search box with/without marks, by using the author search box (advanced search), or other. They were then asked to report the number of errors in their list of publications (i.e. the number of articles in the list that were not co-authored by themselves and the number of articles co-authored by themselves that were not included in the list of their publications) and the reason for these errors. The responses options for this question were respectively: name homonymy or other; and name synonymy, name change, spelling error of the first name/last name, journal not referenced in *PubMed*, or other. Finally, they were asked about their knowledge (yes/no) and use (yes/no) of and their views on *ORCID* and *ResearcherID* (response options: very useful, rather useful, useful, not very useful, not useful).

Two librarians and five physician-scientists reviewed the questionnaire to identify any difficulty responders might have met in responding to the questions.

Ethical considerations

All data were collected in an anonymous manner. Tacit consent was presumed from the physicians if they completed the survey. Since this study did not involve the collection of personal health-related data it did not require ethical review, according to current Swiss law.

Statistical analyses and sample size

We used frequency tables to describe categorical variables and means and standard deviations (SD) or medians and interquartile range (IQR) to summarize continuous or count variables. We computed the average number of errors per physician and the proportion of physicians with zero, one, two, three, four and ≥ five errors. We compared the average number of publications and errors by gender, age group (using the median of the distribution: < and ≥ 45 years) and type of author name query used (advanced query vs. other) (PubMed Help 2019). We used univariate negative binomial regressions (count data with over-dispersion) to examine the difference in the distribution of the variables between the groups (Stata 16 Help for Nbreg 2019; Negative Binomial Regression|Stata Data Analysis Examples 2019; Negative Binomial Regression|Stata Annotated Output 2019). For statistically significant differences (p value < 0.05), we performed multivariate analyses, incorporating gender, age group and number of publications into the model. Finally, we assessed the proportion of physicians knowing and using *ORCID* and *ResearcherID*, and the proportion of physicians finding these tools useful (i.e. "very useful" or "rather useful"). We presented the results for the entire sample and compared them by gender and age group (< and ≥ 45 years) using Fischer's exact tests (several cells with small frequencies).

Using the sample size determination for negative binomial regression, we estimated that by comparing two groups (those who used the author search box and those who did not), a sample of 120 would be sufficient to detect a difference of five publications with a type I and II error both set at 5%. We expected the average number of errors to be one in the first group and five in the second group, the ratio of the number of participants to be 0.2 and the dispersion parameter to be five (Zhu and Lakkis 2014). We performed all statistical analyses with STATA version 15.1 (College Station, USA).

Results

Of the 300 senior physicians contacted by e-mail, 156 (52%) agreed to participate in the study. Table 1 summarizes the main characteristics of the 145 physicians who published at least one peer-reviewed article (= physician-scientists). Their median age was 45 years and 71% were men. The majority of these physicians were Swiss (78%). Most of them practiced in the following Departments: Internal Medicine (31%), Pediatrics, Gynecology and Obstetrics (14%), and Community Health and Medicine (13%). On average, 8 h of their weekly working time was spent on research activities.

The total number of peer-reviewed articles published by the 145 physicians was 8,673. Table 2 shows the average number of publications, the average number of errors in authors' lists of publications and the reasons for these errors. On average, physicians published 60 articles. To retrieve their list of publications, the majority of them entered their first

Table 1 Physician-scientists' characteristics ($N = 145$)

Characteristics[a]	*N* (%)	Median (IQR)
Gender ($N = 130$)		
Male	92 (70.8)	
Female	38 (29.2)	
Age group ($N = 130$)		
< 45 years	65 (50.0)	
≥ 45 years	65 (50.0)	
Medical specialty ($N = 124$)		
Internal medicine	38 (30.7)	
Paediatrics, gynaecology and obstetrics	17 (13.7)	
Community health and medicine	16 (12.9)	
Surgery	15 (12.1)	
Anaesthesiology, pharmacology, intensive care and emergencies	9 (7.3)	
Oncology	8 (6.5)	
Psychiatry	7 (5.7)	
Rehabilitation and geriatrics	6 (4.8)	
Clinical neurosciences	5 (4.0)	
Radiology and medical informatics	3 (2.4)	
Nationality ($N = 129$)		
Swiss	101 (78.3)	
Non swiss[b]	28 (21.7)	
Average number of hours worked per week ($N = 128$)		
For clinical activity		55 (10)
For research activity		8 (11)

[a]Numbers do not add to 145 because of missing data

[b]Nationality: 14 French, 3 Greek, 2 Italian, 1 German, 1 Belgian, 1 American, 1 Dutch, 1 Tunisian, 1 Angolan and 1 Rwandan (2 missing data)

and last names in the search box (56%). Only 17% used the author search box (advanced search). On average, there were 5 articles (SD 13) in the lists of publications that were not co-authored by study participants. The median number of these errors was 0 (IQR 2) with more than two-thirds of participants observing no errors. The majority of the errors leading to a number of publications exceeding the actual figure were due to name homonymy (78%). In contrast, the number of articles co-authored by study participants that did not appear in the lists of publications averaged 3 (SD 8). The median number of these errors was 1 (IQR 3) with just under 50% of participants observing no errors. The majority of errors were due either to journals that were not referenced by *PubMed* (33%) or to name synonymy (30%).

As shown in Table 3 and Figs. 1 and 2, the number of errors was higher among older physicians and those who did not use the advanced search option. The results were similar by limiting our analyses to non-Swiss physicians only. The mean number of articles not co-authored by the physician was 0.5 (SD 0.8) for those who used the advanced search option vs. 5.3 (SD 14.1), *p*-value 0.18, for those who did not use this option. The mean number of missing articles co-authored by the physician was 1.7 (SD 2.3) for the advanced search option vs. 2.6 (SD 6.0), *p*-value 0.62. These differences were not statistically significant

Table 2 Average number of peer-reviewed publications, average number of errors and reasons for these errors in *PubMed*-based authors' lists of publications ($N=145$ physician-scientists)

	N (%)	Mean (SD)	Median (IQR)
Actual number of peer-reviewed publications ($N=145$ physicians)		59.8 (73.1)	34 (77)
Type of author name query used in *PubMed* to retrieve a list of his/her publications ($N=140$ physicians)			
By entering his/her first and last name in the search box	79 (56.4)		
By using the author search box (search builder)	24 (17.1)		
By entering his/her last name + initials	18 (12.9)		
Other	19 (13.6)		
Number of articles in the list of peer-reviewed publications that are not co-authored by the physician-scientist ($N=140$ physicians)			
0	96 (68.6)		
1	9 (6.4)		
2	8 (5.7)		
3	2 (1.4)		
4	1 (0.7)		
≥ 5	24 (17.2)		
Average number of articles		5.0 (12.8)	0 (1.5)
Reasons for these errors ($N=702$ articles)			
Name homonymy	549 (78.2)		
Physician-scientist listed in a group of co-investigators[a]	136 (19.4)		
Unknown reason	17 (2.4)		
Number of articles co-authored by the physician-scientist that are not in the list of his/her publications ($N=137$ physicians)			
0	63 (46.0)		
1	22 (16.1)		
2	16 (11.7)		
3	10 (7.3)		
4	4 (2.9)		
≥ 5	22 (16.0)		
Average number of articles		3.2 (7.8)	1 (3)
Reasons for these errors ($N=444$ articles)			
Journal not referenced in *PubMed*	147 (33.1)		
Name synonymy	135 (30.4)		
Incomplete name	63 (14.2)		
Spelling error of the name	39 (8.8)		
Name change	28 (6.3)		
Article not yet indexed	14 (3.2)		
Unknown reason	18 (4.0)		

[a]The physician-scientist is part of a research team. The research team added his/her name to the list of co-authors of the article even though he/she had not participated in the study in question

but the sample was small ($N=28$). In multivariate analysis (Table 4), the adjusted ratio of means between those who did not use and those who used the advanced search option was 14.3 (95% CI 2.5–80.2, p-value 0.003) for articles in the lists of publications that were not

Table 3 Average number of peer-reviewed publications and errors in *PubMed*-based authors' lists of publications, stratified by gender, age group and type of author name queries ($N = 145$ physician-scientists)

	Men	Women	p-value[a]	< 45 years	≥ 45 years	p-value[a]	Using the author search box	Other	p-value[a]
	Mean (SD)	Mean (SD)		Mean (SD)	Mean (SD)		Mean (SD)	Mean (SD)	
	Median (IQR)	*Median (IQR)*		*Median (IQR)*	*Median (IQR)*		*Median (IQR)*	*Median (IQR)*	
Number of peer-reviewed publications ($N = 145$ physicians)	61.6 (74.8)	46.6 (57.3)	0.21	23.1 (28.7)	91.3 (82.2)	< 0.001	59.8 (86.3)	59.9 (69.7)	0.99
	35 (77.5)	*23.5 (61)*		*13 (31)*	*70 (87)*		*26.5 (33.5)*	*36.5 (79)*	
Number of articles in the list of peer-reviewed publications that are not co-authored by the physician ($N = 140$ physicians)	5.9 (14.3)	2.4 (7.0)	0.15	2.8 (9.1)	6.9 (15.2)	0.11	1.0 (2.6)	5.9 (13.9)	0.02
	0 (2)	*0 (0)*		*0 (0)*	*0 (2)*		*0 (0)*	*0 (2)*	
Number of articles co-authored by the physician that are not in the list of his/her publications ($N = 137$ physicians)	2.5 (6.2)	4.7 (10.7)	0.09	1.1 (1.6)	5.2 (10.6)	< 0.001	1.5 (2.0)	3.6 (8.4)	0.05
	1 (2)	*1.5 (3)*		*1 (2)*	*1 (4)*		*1 (2)*	*1 (3)*	

[a]Negative binomial regressions

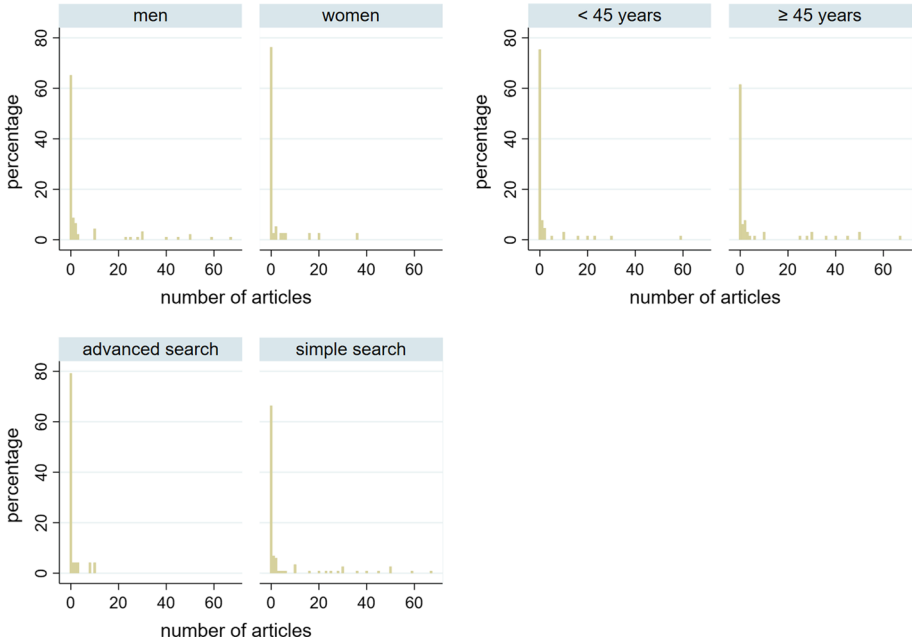

Fig. 1 Number of articles not co-authored by the physician, by gender, age group and type of author name query (advanced search option [=advanced search] versus other type of author name query used [=simple search])

co-authored by study participants. This figure was 1.3 (95% CI 0.5–3.7, *p*-value 0.58) for articles co-authored by study participants that were not in the lists of publications.

Finally, Table 5 shows physicians' knowledge and use of *ORCID* and *Researcher ID*. The majority were aware of *ORCID* (82%), but only 21 physicians (16%) added all their articles into the system. By contrast, these figures were 39% and 6% for *ResearcherID*. The proportions of physicians who knew and used *ORCID* and *ResearcherID* tended to be higher among men and older physicians, but the differences were not statistically significant.

Discussion

Main findings

In summary, in this cross-sectional study conducted in the French-speaking part of Switzerland, we showed that only 17% of participants used the advanced search option in *PubMed* although it seems to be more accurate than the other types of author name query for retrieving authors' publications. On average, for the advanced search group, there was only one article in the lists that was not co-authored by participants (vs. 6) and 1.5 article co-authored by participants that did not appear in the lists (vs. 3.5). The majority of these errors were due to name homonymy (81%), respectively to publications in journals not referenced by *PubMed* (33%) and name synonymy (30%). We also showed that, although

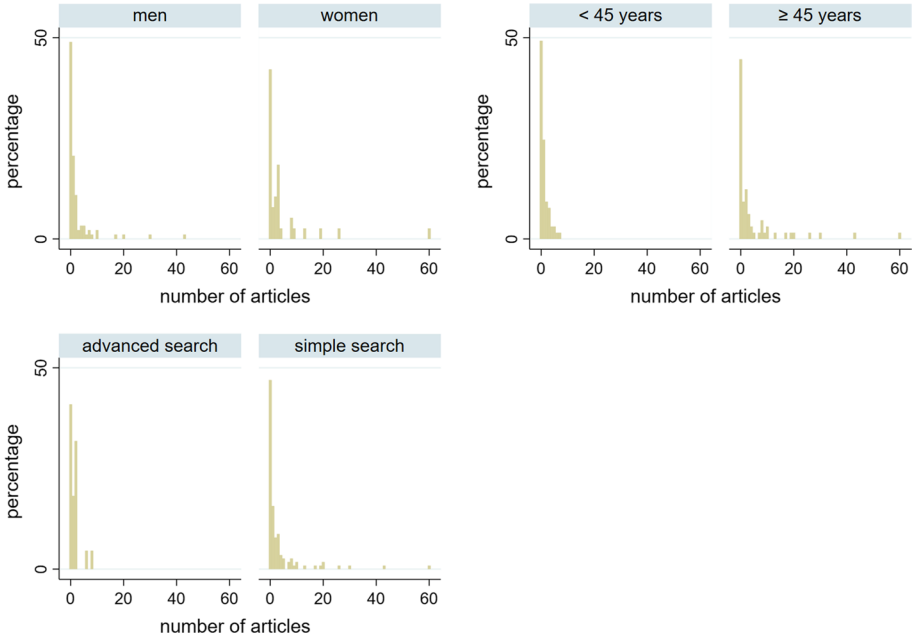

Fig. 2 Number of missing articles co-authored by the physician, by gender, age group and type of author name query (advanced search option [=advanced search] versus other type of author name query used [=simple search])

82% of participants were aware of *ORCID*, only 16% added all their articles to the system. These figures were 39% and 6% respectively for *ResearcherID*.

Comparison with existing literature

There are three main ways to retrieve author-based lists of publications or to check the accuracy of pre-established lists of publications (Lerchenmueller and Sorenson 2016; D'Angelo and van Eck 2020). The first is to directly ask authors whose publication lists have been compiled whether they contain errors. This type of approach can be considered the gold standard since no one is more qualified than the authors themselves to assess whether their publication lists are correct. However, this approach is particularly time-consuming. This is probably the reason our study is the first, to our knowledge, to have chosen this evaluation method. A second method is to rely on unique author identifiers, such as *ORCID* and *ResearcherID* (see below for further discussion). For these identifiers to be usable in practice, the active participation of the majority of authors is required, which is unfortunately hard to achieve. The last method consists of using author name disambiguation algorithms. These tools are also available for *PubMed*. The advantage is to be able to proceed relatively quickly and autonomously, without having to resort to authors. According to several studies, the risk of errors is relatively low, estimated at a few percent in general (Lerchenmueller and Sorenson 2016; Liu et al. 2014; Sanyal et al. 2019; Torvik and Smalheiser 2009; Kawashima and Tomizawa 2015). Unfortunately, these estimates are not based on a comparison analysis against the gold standard (evaluation by the

Table 4 Unadjusted and adjusted association between the number of peer-reviewed publications and errors in *PubMed*-based authors' lists of publications, and gender, age group and type of author name queries ($N = 145$ physician-scientists)

Characteristics	Number of peer-reviewed publications				Number of articles not co-authored by the physician†				Number of articles not in the list‡			
	Ratio of means (95% CI)	*p*-value§	Adj. ratio (95% CI)	*p*-value¶	Ratio of means (95% CI)	*p*-value§	Adj. ratio (95% CI)	*p*-value¶	Ratio of means (95% CI)	*p*-value§	Adj. ratio (95% CI)	*p*-value¶
Women (ref. group: men)	0.8 (0.5–1.2)	0.21			0.4 (0.1–1.4)	0.15			1.9 (0.9–3.9)	0.09		
≥ 45 years (ref. group: < 45 years)	4.0 (2.8–5.6)	< 0.001	1.5 (1.2–2.0)	< 0.001	2.5 (0.8–7.4)	0.11			4.5 (2.4–8.5)	< 0.001	3.2 (1.5–6.6)	0.002
Not using the author search box (ref group: using the author search box)	1.0 (0.6–1.7)	0.99			5.9 (1.4–24.7)	0.02	14.3 (2.5–80.2)	0.003	2.5 (1.0–6.1)	0.05	1.3 (0.5–3.7)	0.58

†Number of articles in the list of peer-reviewed publications that are not co-authored by the physician

‡Number of articles co-authored by the physician that are not in the list of his/her publications

§Univariate negative binomial regressions

¶Multivariate negative binomial regressions (adjusted for gender, age group and number of publications

Table 5 Physician-scientists' knowledge and use of *ORCID* and *Researcher ID*, overall and stratified by gender and age group ($N=145$ physician-scientists)

	Overall *N* (%)	Men *N* (%)	Women *N* (%)	*p*-value[a]	< 45 years *N* (%)	≥ 45 years *N* (%)	*p*-value[a]
ORCID							
Knows *ORCID* ($N=135$)	110 (81.5)	77 (83.7)	29 (76.3)	0.33	52 (80.0)	54 (83.1)	0.82
Has an *ORCID* account ($N=135$)	96 (71.1)	67 (72.8)	24 (63.2)	0.30	44 (67.7)	47 (72.3)	0.70
Has added at least some of his/her articles to *ORCID* ($N=135$)	58 (43.0)	43 (46.7)	13 (34.2)	0.24	28 (43.1)	28 (43.1)	1.00
Has added all his/her articles to *ORCID* ($N=135$)	21 (15.6)	18 (19.6)	2 (5.3)	0.06	11 (16.9)	9 (13.9)	0.81
Considers *ORCID* useful or very useful ($N=135$)	65 (48.2)	49 (53.3)	13 (34.2)	0.06	26 (40.0)	36 (55.4)	0.11
Researcher ID							
Knows *Researcher ID* ($N=135$)	52 (38.5)	34 (37.0)	13 (34.2)	0.84	18 (27.7)	29 (44.6)	0.07
Has a *Researcher ID* account ($N=135$)	25 (18.5)	17 (18.5)	6 (15.8)	0.81	9 (13.9)	14 (21.5)	0.36
Has added at least some of his/her articles to *Researcher ID* ($N=135$)	17 (12.6)	13 (14.1)	3 (7.9)	0.39	6 (9.2)	10 (15.4)	0.42
Has added all his/her articles to *Researcher ID* ($N=135$)	8 (5.9)	8 (8.7)	0	0.10	1 (1.5)	7 (10.8)	0.06
Considers *Researcher ID* useful or very useful ($N=135$)	22 (16.3)	18 (19.6)	2 (5.3)	0.06	6 (9.2)	14 (21.5)	0.09

[a]Fisher's exact tests

authors themselves). For example, the *author-ity* model was evaluated using automatically generated gold standard datasets (based on email addresses, grant numbers or Principal Investigator IDs assigned by the NIH to the scientists applying for grants, self-citations, ISI highly cited researchers databases and Community of Science profiles) (Lerchenmueller and Sorenson 2016; Torvik and Smalheiser 2009).

In addition, there are two main ways to submit an author-based query to *PubMed*. You can enter the author's surname plus or minus first name in free text in the *PubMed* search box (for the first name, you can enter either the first letter or the whole first name). You can also use the advanced search tool and choose a name plus or minus first name from the drop-down list (for the first name, you have to choose again between the first letter or the whole first name). Searching with the surname and the first letter of the first name obviously increases the risk of "false positive" results (articles in the list not co-authored by the researcher) but decreases the risk of "false negative" results (missing articles in the list). The reverse is true if the search is made with the whole surname and first name.

We showed that, when used by physician scientists, the advanced search option in *PubMed* reduces the risk of errors. However, it would also be interesting in the future to assess how physician scientists use the advanced search tool and to evaluate the intrinsic performance (accuracy of search) of the different advanced strategies listed above. Probably the most effective author search strategy would be to choose from the *PubMed* drop-down list the author's name with the first letter of his/her first name, and then to reduce the "noise" ("false positive" results) by completing the query, using for example the author's affiliation or e-mail address. Queries concerning authors' affiliation are already available in *PubMed*, those concerning their e-mail address unfortunately not yet. It should also be borne in mind that authors' affiliation and/or e-mail address may vary over time. As already stated, there are also specific disambiguation tools, some of which are freely available (Lerchenmueller and Sorenson 2016; Liu et al. 2014; Sanyal et al. 2019; Torvik and Smalheiser 2009; Kawashima and Tomizawa 2015). However, their use by physician scientists is probably not imaginable on a large scale.

Several studies have shown that name ambiguity (homonyms or synonyms) could be a major issue when retrieving the author lists of publications from databases (Akers et al. 2016; Gasparyan et al. 2016; Liu et al. 2014; Ruiz-Pérez et al. 2002). Despite these concerns, our study suggests that the advanced author search option in *PubMed* could significantly reduce the risk of errors. The average number of errors obtained with the advanced search option was one for the first type of errors and 1.5 for the second type, i.e. 1.7% and 2.5% respectively if the number of errors is expressed as a percentage of errors relative to the average number of peer-reviewed publications ($N=60$). In addition, we showed that the reduction of the risk of errors mainly concerned articles in the lists of publications that were not co-authored by physicians (the number of errors decreased from six with the standard search option to one with the advanced search option). The difference remained statistically significant in multivariate analysis. For the second type of errors, the difference between the two types of query was lower (from 3.5 for standard search to 1.5 for advanced search).

Unfortunately, there seems to be only a minority of physicians using the advanced search option in *PubMed* (17% in our study) with the risk of obtaining inaccurate data. However, the study participants were all senior physicians (heads of division, senior staff physicians or senior residents) who are generally accustomed to conducting *PubMed* research in the context of their clinical or academic activities (e.g., writing scientific articles, teaching). We believe that a short training by librarians or peers would probably be useful to improve *PubMed* queries because the advanced search option

is easy to use and would allow to refine and improve the search, not only for author searches (PubMed Help 2019). Pregraduate education in this field should also be encouraged.

Somewhat surprisingly, 19% of the articles mistakenly found in publication lists ($N=136$) were related to the fact that the physicians were part of a research team that added their names to the list of co-authors when they had not participated in these studies. This is not a lack of performance of the search engine as such, because the names in these articles actually refer to the right authors. However, the fact remains that these articles were not co-authored by these physicians and were therefore "errors" in these publication lists. As this issue apparently occurs relatively often, it should be addressed in the future.

Of the articles in the publication lists that were not co-authored by the study physicians, 17 (2.4%) were of unknown cause according to the participants' responses. It can be hypothesized that some of these errors were nevertheless related to name homonomy, assuming that not all respondents were aware of the exact definition of the concept of homonymy, even though it was explained in the questionnaire. In addition, some of the 18 missing articles "of unknown cause" (4%) could be linked to articles whose author(s) are in fact a group (i.e. non-author entities). Some articles are indeed sometimes published under the name of organizations. In this case, although these organizations are not individual authors, *PubMed* still considers them as such, and these articles may not be retrieved by writing the name of the author(s) in the search box (Liu et al. 2014).

Of course, the results obtained in this study would not necessarily have been the same if the lists of publications had been generated by librarians who use optimal search strategies to obtain these lists probably more often than do hospital-based senior physicians. However, the aim of this study was not to assess the accuracy of *PubMed*-based author lists of publications as such, but their accuracy when searches are performed by hospital-based senior physicians.

To evaluate the accuracy of *PubMed* searches by author, we asked senior physicians to generate their own list of publications and then to check if their list was correct. We could also have generated these lists ourselves and then asked the physicians participating in the study to verify them. We preferred to follow the first approach because our aim was to assess the accuracy of *PubMed* searches when they were performed by senior physicians. However, we do not think that the type of author name query used, and therefore the results of the study, would have been very different if they had performed a search for the publications of other authors. By the way, they were asked to proceed as they normally do when conducting this type of research for other authors as part of their clinical or research activities.

To reduce the number of errors even further it would be interesting to use author identifiers, such as *ORCID* and *ResearcherID*. These author-identifying systems provide researchers with a unique identifier and lead to disambiguated search results (Akers et al. 2016; Gasparyan et al. 2014, 2016, 2017). However, as shown in our study, very few physicians add their publications to *ORCID* and even fewer to *researcherID*. In these circumstances, these identifiers cannot be considered as reliable for most authors' publication searches. An alternative solution (for the future) would be to develop an automatic referencing system that would automatically include any new publication in the author identifying systems. Then, by using the author identifier option [auid] in *PubMed* it would be possible to retrieve a full list of publications associated with this unique identifier (PubMed Help 2019). This type of referencing would have the advantage of being automatic but the disadvantage of only covering new publications, which would often significantly underestimate the number of researchers' publications.

Limitations

This study has several limitations. First, the number of participants was relatively small, particularly in the advanced search group, although it exceeded the minimum sample size we had calculated. Second, the study was only conducted in the French-speaking part of Switzerland. However, we do not believe that this reduces the generalizability of this study. Indeed, 22% of the participants were non-Swiss (mostly physicians from Western countries) and there is no evident reason to think that Swiss physicians are unique in the *PubMed* research strategies they use. Moreover, we showed that the results concerning the number of errors in publication lists were similar by limiting our analyses to non-Swiss physicians only ($N=28$). Therefore, our results can probably be generalised to other hospitals in Switzerland or in many Western countries. Of note, the issue of incorrectly abbreviated, misspelled and misplaced first, middle and surnames is patent and striking for authors from some countries or regions, particularly in Asia and the Middle East (Gasparyan et al. 2016). This means that our results are probably not generalizable to these regions. Third, as stated above, to evaluate the accuracy of *PubMed* searches by author, we asked senior physicians to generate their own list of publications and then to check if their list was correct. However, we do not think that the type of author name query used, and therefore the results of the study, would have been very different if they had performed a search for the publications of other authors. Fourth, the study design did not allow us to directly compare, for a given participant, the two types of *PubMed* search used (advanced search option [Y/N]). Finally, since the participants themselves had to count the errors and document the reasons for them, we cannot rule out information bias. Indeed, some physicians may have miscounted the errors and/or assessed the reasons incorrectly.

Conclusion

When used by hospital-based senior physicians, the advanced search option in *PubMed* is accurate for tracking researchers' publications. A short training by librarians or peers to learn how to employ this option would be desirable as few physicians use it. Author identifiers are only used by a minority of senior physicians and are therefore not recommended, as they would lead to inaccurate results.

Acknowledgements We would like to warmly thank all the physicians who participated in the study.

Funding Open Access funding provided by Université de Genève.

Data availability Data associated with this article are available in the Open Science Framework at https://osf.io/zwuaq/.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Akers, K. G., Sarkozy, A., Wu, W., & Slyman, A. (2016). ORCID author identifiers: A primer for librarians. *Medical Reference Services Quarterly, 35*(2), 135–144.

Bavdekar, S. B., & Tullu, M. S. (2016). Research publications for academic career advancement: An idea whose time has come. But is this the right way? *Journal of Postgraduate Medicine, 62*(1), 1–3.

D'Angelo, C. A., & van Eck, N. J. (2020). Collecting large-scale publication data at the level of individual researchers: a practical proposal for author name disambiguation. *Scientometrics, 123*(2), 883–907.

Fiorini, N., Canese, K., Bryzgunov, R., Radetska, I., Gindulyte, A., & Latterner, M., et al. (2018). PubMed Labs: an experimental system for improving biomedical literature search. Database: The Journal of Biological Databases and Curation (Internet). 2018 Sep 18 [cited 2019 Dec 15]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6152140/.

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.

Gasparyan, A. Y., Nurmashev, B., Yessirkepov, M., Endovitskiy, D. A., Voronov, A. A., & Kitas, G. D. (2017). Researcher and Author Profiles: Opportunities, Advantages, and Limitations. *Journal of Korean Medical Science, 32*(11), 1749–1756.

Gasparyan, A. Y., Yessirkepov, M., Gerasimov, A. N., Kostyukova, E. I., & Kitas, G. D. (2016). Scientific author names: Errors, corrections, and identity profiles. *Biochemia Medica, 26*(2), 169–173.

Kawashima, H., & Tomizawa, H. (2015). Accuracy evaluation of Scopus Author ID based on the largest funding database in Japan. *Scientometrics, 103*(3), 1061–1071.

Lafrenière, D., Menuz, V., Hurlimann, T., & Godard, B. (2013). Knowledge dissemination interventions: A literature review. *SAGE Open, 3*(3), 2158244013498242.

Lerchenmueller, M. J., & Sorenson, O. (2016). Author disambiguation in PubMed: Evidence on the precision and recall of authority among NIH-funded scientists. *PLoS ONE, 11*(7), e0158731.

Liu, W., Islamaj Doğan, R., Kim, S., Comeau, D. C., Kim, W., Yeganova, L., et al. (2014). Author name disambiguation for PubMed. *Journal of the Association for Information Science and Technology, 65*(4), 765–781.

Mabe, M. A. (2010). Scholarly communication: A long view. *New Review of Academic Librarianship, 16*(sup1), 132–144.

Negative Binomial Regression|Stata Annotated Output [Internet] (2019). Retrieved 2, Nov 2019, from https://stats.idre.ucla.edu/stata/output/negative-binomial-regression/.

Negative Binomial Regression|Stata Data Analysis Examples [Internet] (2019). Retrieved 2, Nov 2019, from https://stats.idre.ucla.edu/stata/dae/negative-binomial-regression/.

Post, R. E., Weese, T. J., Mainous, A. G., & Weiss, B. D. (2012). Publication productivity by family medicine faculty: 1999 to 2009. *Family Medicine, 44*(5), 312–317.

PubMed Help [Internet] (2019). Bethesda (MD): National Center for Biotechnology Information (US). Retrieved 2, Nov 2019, from https://www.ncbi.nlm.nih.gov/books/NBK3827/.

Ruiz-Pérez, R., Delgado López-Cózar, E., & Jiménez-Contreras, E. (2002). Spanish personal name variations in national and international biomedical databases: Implications for information retrieval and bibliometric studies. *Journal of the Medical Library Association, 90*(4), 411–430.

Sanyal, D., Bhowmick, P., & Das, P. (2019). A review of author name disambiguation techniques for the PubMed bibliographic database. *Journal of Information Science*, 016555151988860.

Stata 16 Help for Nbreg [Internet] (2019). Retrieved 2, Nov 2019, from https://www.stata.com/help.cgi?nbreg.

Torvik, V. I., & Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data, 3*(3), 1–29.

Vale, R. D. (2015). Accelerating scientific publication in biology. *Proceedings of the National Academy of Sciences USA, 112*(44), 13439–13446.

Zhu, H., & Lakkis, H. (2014). Sample size calculation for comparing two negative binomial rates. *Statistics in Medicine, 33*(3), 376–387.