Research note. Open letter to the users of the new *PubMed*: a critical appraisal

María García-Puente; Elena Pastor-Ramon; Oskia Agirre; José-María Morán; Iván Herrera-Peco

How to cite this note:

García-Puente, María; **Pastor-Ramon, Elena**; **Agirre, Oskia**; **Morán, José-María**; **Herrera-Peco, Iván** (2020). "Research note. Open letter to the users of the new *PubMed*: a critical appraisal". *Profesional de la información*, v. 29, n. 3, e290336.

https://doi.org/10.3145/epi.2020.may.36

Manuscript received on June 17th 2020
Accepted on June 22th 2020

María García-Puente ✉
https://orcid.org/0000-0002-6521-665X

Fundación Jiménez Díaz
Avda. Reyes Católicos, 2. 28040 Madrid
AlterBiblio. Madrid, Spain.
maria@alterbiblio.com

Elena Pastor-Ramon
https://orcid.org/0000-0003-2609-6541

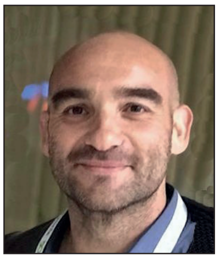
Virtual Health Sciences Library of the Balearic Islands (Bibliosalut)
Ctra. de Valldemossa, 79, módulo L+1.
07120 Palma (Illes Balears), Spain
epastor@bibliosalut.com

Oskia Agirre
https://orcid.org/0000-0001-8991-691X

Universidad del País Vasco (UPV/EHU) Biblioteca
Campus de Gipuzkoa, Plaza Elhuyar, 2.
20018 Donostia-San Sebastián, Spain
oskia.aguirre@ehu.eus

José-María Morán
https://orcid.org/0000-0002-5538-182X

University of Extremadura, Nursing Department, Nursing and Occupational Therapy College
Cáceres, Spain.
jmmorang@unex.es



Iván Herrera-Peco
https://orcid.org/0000-0002-5183-5679

Universidad Alfonso X el Sabio, Facultad de Ciencias de la Salud
Avda. Universidad, 1. 28621 Villanueva de la Cañada (Madrid), Spain
iherrpec@uax.es

Abstract

PubMed is a free database used daily by about 2.5 million people to search and retrieve scientific documents related to Health Sciences. In May 2020, certain changes were made to its search algorithm, which at first sight improves the location of scientific articles, but upon analyzing its operation in more depth, we detected some changes that make the reproducibility of bibliographic searches difficult. In order to safeguard the reproducibility and replicability of the searches carried out for systematic reviews, narratives and meta-analyzes, we suggest accompanying these strategies with a file in a format compatible with reference managers, to facilitate comparison and verification of the strategy to be replicated in a future.

Keywords

PubMed; Information storage and retrieval; Search engines; Databases; Search algorithms; User-computer interface; Reproducibility of results.

1. Introduction

PubMed is the most widely used search tool for biomedical and life sciences literature. Each day it is accessed by approximately 2.5 million users worldwide, processing more than 3 million search requests (**Fiorini** *et al.*, 2018). In 2019 alone, over 3.3 billion searches were performed on *PubMed* (*National Library of Medicine*, 2020a).

Until May 18, 2020 (**Canese** *et al.*, 2020), *PubMed* used an information search algorithm that offered documents searches by relevance, which was calculated thanks to the TF-IDF (Term Frequency-Inverse Document Frequency) system. This method, which is based on manual analyses, became untenable given the volume of information that *PubMed* currently handles (more than 30 million records) (*National Library of Medicine*, 2020a). Because of this, the new version of *PubMed* uses a search algorithm based on machine learning known as Best Match. This algorithm incorporates a system that allows retrieval and reordering of the displayed articles by relevance. Relevance is calculated according to the frequency with which some factors appear related to previous searches, these factors being as follows: i) article use, ii) publication date, iii) relevance score (number of times the document matches a search), and iv) article type (**Fiorini** *et al.*, 2018b). It will then return an ordered list of existing documents in *PubMed* that match the query terms.

> *PubMed* is the most widely used database by biomedical researchers, healthcare professionals, and health science librarians

2. Problem with the reproducibility of bibliographic searches

However, despite these new features, which initially allow the *PubMed* search algorithm to work better, we wish to highlight certain changes in the functioning of this algorithm that may alter one of the pillars on which scientific documents are based, such as systematic reviews and meta-analyses, that is, the reproducibility of bibliographic searches (**Lee**, 2017; **Moher** *et al.*, 2015).

PubMed performs searches based on the terms entered by the user by applying Automatic Term Mapping (ATM). ATM translates the user's strategy by performing calculations with the use of three different tables:

- Subject Translation Table,
- Journal Translation Table, and
- Author Translation Table.

In each of these tables, *PubMed* takes into account various parameters which, moreover, differ considerably from the translation or processing performed in legacy *PubMed*. For example, in the first table, Subject Translation Table, the system searches for *MeSH* descriptors, entry terms (synonyms), subheadings, supplementary concepts, publication types, the singular and plural form of the word, American and British spellings, the translation into the generic name if the entry matches a drug trade name, and finally, the UMLS matches. A new feature is the search for synonyms and related words, in plural and singular form. This change alone adds results that were not retrieved with the former algorithm.

> We highlight certain changes in the functioning of the new algorithm that may alter one of the pillars on which scientific documents are based: the reproducibility of bibliographic searches

3. Comparing searches. Example

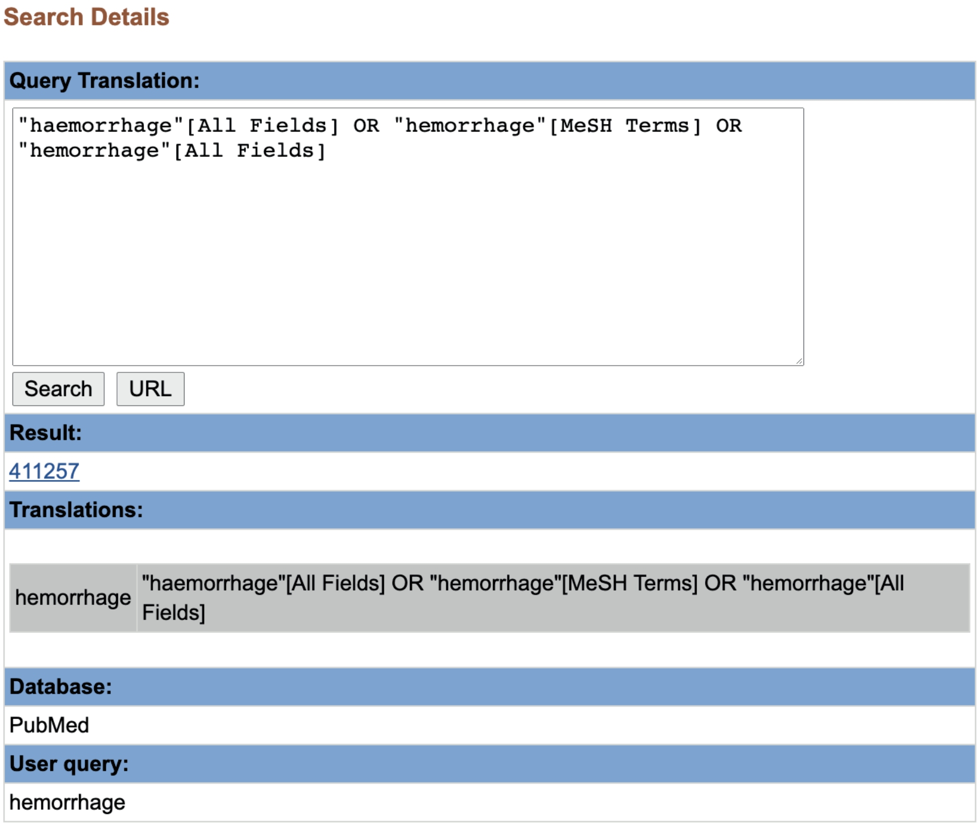
Firstly, we would like to focus on the search strategies that could have been executed with the legacy *PubMed* search algorithm, prior to the updating of the search and sorting algorithm. Let's take the search for the term hemorrhage as an example. Figure 1 shows the translation performed by legacy *PubMed*, which returned 411,257 results (as of June 17, 2020), while Figure 2 shows the translation performed by *PubMed*'s new search algorithm, which returned 4,880,066 results.

Another major change that substantially modifies the number of results obtained in *PubMed* is the use of truncation, which is represented by a "*". In legacy *PubMed*, searches using truncation only took into account the first 600 variants of the term, while the new algorithm does not limit the number of variants retrieved.

4. Importance of the dates

Finally, we come across the biggest obstacle we have encountered when trying to replicate a search strategy. In order to replicate the strategy, we have taken into account not only the terms and syntax used, but also the date on which it was carried out. The first problem arises with the different dates that a record contains in its indexation in *PubMed* (*PubMed*, 2020). We identified a few:

- Create Date (CRDT): date when the record of the appointment was first created;
- Date Completed (DCOM): date when the process of the record in the database is finished, and those records that are "In Process" do not have this field;
- Date Created (DA): date when the process of the record starts;



Search Details

Query Translation:

```
"haemorrhage"[All Fields] OR "hemorrhage"[MeSH Terms] OR
"hemorrhage"[All Fields]
```

Result:

411257

Translations:

hemorrhage	"haemorrhage"[All Fields] OR "hemorrhage"[MeSH Terms] OR "hemorrhage"[All Fields]

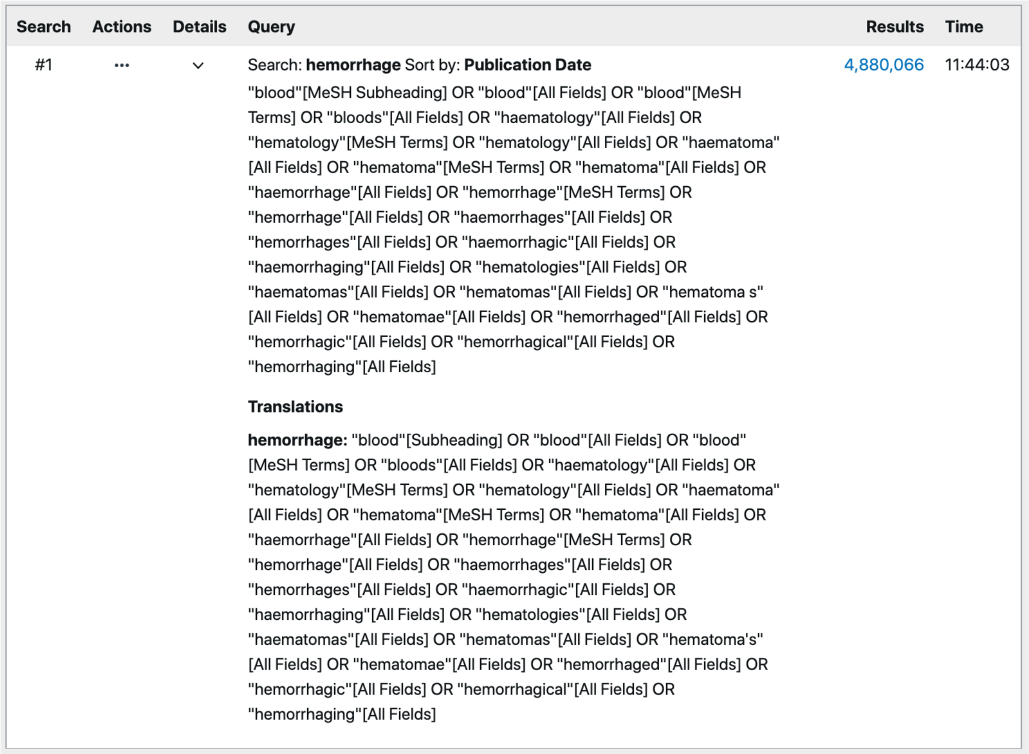
Database:

PubMed

User query:

hemorrhage

Figure 1. Translation of the user search in legacy *PubMed*



Search	Actions	Details	Query	Results	Time
#1	···	⌄	Search: **hemorrhage** Sort by: **Publication Date** "blood"[MeSH Subheading] OR "blood"[All Fields] OR "blood"[MeSH Terms] OR "bloods"[All Fields] OR "haematology"[All Fields] OR "hematology"[MeSH Terms] OR "hematology"[All Fields] OR "haematoma"[All Fields] OR "hematoma"[MeSH Terms] OR "hematoma"[All Fields] OR "haemorrhage"[All Fields] OR "hemorrhage"[MeSH Terms] OR "hemorrhage"[All Fields] OR "haemorrhages"[All Fields] OR "hemorrhages"[All Fields] OR "haemorrhagic"[All Fields] OR "haemorrhaging"[All Fields] OR "hematologies"[All Fields] OR "haematomas"[All Fields] OR "hematomas"[All Fields] OR "hematoma s"[All Fields] OR "hematomae"[All Fields] OR "hemorrhaged"[All Fields] OR "hemorrhagic"[All Fields] OR "hemorrhagical"[All Fields] OR "hemorrhaging"[All Fields] **Translations** **hemorrhage:** "blood"[Subheading] OR "blood"[All Fields] OR "blood"[MeSH Terms] OR "bloods"[All Fields] OR "haematology"[All Fields] OR "hematology"[MeSH Terms] OR "hematology"[All Fields] OR "haematoma"[All Fields] OR "hematoma"[MeSH Terms] OR "hematoma"[All Fields] OR "haemorrhage"[All Fields] OR "hemorrhage"[MeSH Terms] OR "hemorrhage"[All Fields] OR "haemorrhages"[All Fields] OR "hemorrhages"[All Fields] OR "haemorrhagic"[All Fields] OR "haemorrhaging"[All Fields] OR "hematologies"[All Fields] OR "haematomas"[All Fields] OR "hematomas"[All Fields] OR "hematoma's"[All Fields] OR "hematomae"[All Fields] OR "hemorrhaged"[All Fields] OR "hemorrhagic"[All Fields] OR "hemorrhagical"[All Fields] OR "hemorrhaging"[All Fields]	4,880,066	11:44:03

Figure 2. Translation of hemorrhage by *PubMed*'s new search algorithm

- Date of Electronic Publication (DEP): date on which the editor makes the electronic version of the article available to the public;
- Date of Publication (DP): contains the full date on which the issue of the journal was published, and may contain the year, month and day, although only the year is mandatory and this data is collected directly from the journal. In addition, it includes both printed and electronic publication dates.
- Entrez Date (EDAT): date when the quotation was added to *PubMed*.

> We alert researchers and health science librarians about the problems detected with the indexing of records in *PubMed* according to the dates registered

When replicating a search, we must establish the time frame for retrieving the same number of results and, ideally, the same results. However, we face the problem of selecting the right date filter. Normally, the first search strategy does not have a temporary filter since it tries to collect all the scientific output indexed until that moment and it is when trying to replicate it that we must add the temporary filter to try to reproduce the search in the same context.

The usual filter is the "DP" publication date, which we have seen includes both the printed and electronic versions. These two versions may be several months apart, so when the second search was performed we may have retrieved articles that were not yet in the database when the first search was performed with the strategy defined below.

An example of this situation is given by trying to replicate the following search equation, whose search time limit was December 30, 2019 to January 5, 2020. This equation was executed during the 4th week of January 2020, and was replicated in the 3rd week of June 2020 (June 16, 2020).

The search strategy is as follows:

> SARS-CoV-2 OR SARS-CoV2 OR COVID2019 OR 2019-nCoV OR COVID- 19 OR COVID19 OR 2019-nCoV OR (("novel coronavirus" OR coronavirus OR "new coronavirus") AND (wuhan[tiab])) OR SARS-coronavirus 2[tw] OR "coronavirus 2"[tw] OR "coronavirus disease 2019" OR "2019-novel coronavirus" OR "new coronavirus" OR "COVID-19" [Supplementary Concept] OR "COVID-19 diagnostic testing" [Supplementary Concept] OR "spike glycoprotein, COVID-19 virus" [Supplementary Concept] OR "COVID-19 drug treatment" [Supplementary Concept] OR "LAMP assay" [Supplementary Concept] OR "severe acute respiratory syndrome coronavirus 2" [Supplementary Concept] OR "COVID-19 serotherapy" [Supplementary Concept] OR "COVID-19 vaccine" [Supplementary Concept] NOT ("animals"[MeSH Terms] NOT "humans"[MeSH Terms]) AND ("2019/12/30"[PDat] : "2020/01/05"[PDat])

The result obtained during the fourth week of January returned a total of 50 results, while replicating the search in the third week of June gave 416 results (**García-Puente**, 2020).

When analyzing the *PubMed* (former *MedLine*) output format of one of these new records, with pmid 32529948, we note that the publication date (PD) 2020 Jan 1 has been assigned, while the other dates provided indicate, in all cases, 2020/06/13. When consulting the data directly on the journal's website, we note that the date of online publication is June 12, 2020, with no date of print publication yet.

> We focus on the search strategies executed with legacy *PubMed* search algorithm, prior to the updating of the search and sorting algorithm

To ensure that this is not an isolated case, we checked several other records. The item with pmid 32528206 has as PD 2020, without indicating month or day, and the rest of the dates match the previous registration: June 13, 2020. However, when consulting the PDF of the article we see that it was received on April 25, so it is impossible that any of its dates can comply with our strategy in which we limit the PD from December 30, 2019 to January 5, 2020. The rest of the articles analyzed have results that resemble these two.

5. New descriptors also cause problems, but less important

Finally, we would like to emphasize that we are aware that the results of a search may vary slightly if, for example, a new descriptor is entered, as occurred in 2019 when Systematic Review was introduced as a *MeSH* descriptor for article type and a retrospective cataloging was performed: Systematic Review [Publication Type]. However, this type of modification would not substantially alter the search to the degree seen with the case described.

6. Final remark

As stated above, and with the aim of ensuring the reproducibility and replicability of the bibliographic searches carried out in the narrative, systematic and meta-analysis reviews, we wish to alert researchers and health science librarians to the problems detected with the indexing of records in *PubMed* according to the dates registered.

> We need to create solid, well-documented search strategies to ensure the reproducibility and replicability of the bibliographic searches

In this context, the need to create solid, well-documented search strategies should be highlighted and accompanied whenever possible by a file in a format compatible with reference managers, which will facilitate the comparison and verification of the search strategy to be replicated in the future.

7. References

Canese, Kathi; **Chan, Jessica**; **Collins, Marie**; **Trawick, Bart**; **Weis, Sarah** (2020). "The new and improved PubMed is here". *NLM Technical bulletin*. 2020 May 19.
https://www.nlm.nih.gov/pubs/techbull/mj20/mj20_PubMed_new.html

Fiorini, Nicolas; **Canese, Kathi**; **Bryzgunov, Rostyslav**; **Radetska, Ievgeniia**; **Gindulyte, Asta**; **Latterner, Martin**; **Miller, Vadim**; **Osipov, Maxim**; **Kholodov, Michael**; **Starchenko, Grisha**; **Kirreev, Evgeny**; **Lu, Zhiyong** (2018a). "PubMed Labs: An Experimental system for improving biomedical literature search". *Database (Oxford)*, v. 2018, bay094.
https://doi.org/10.1093/database/bay094

Fiorini, Nicolas; **Canese, Kathi**; **Starchenko, Grisha**; **Kireev, Evgeny**; **Kim, Won**; **Miller, Vadim**; **Osipov, Maxim**; **Kholodov, Michael**; **Ismagilov, Rafis**; **Mohan, Sunil**; **Ostell, James**; **Lu, Zhiyong.** (2018b) "Best Match: New relevance search for PubMed". *PLoS biology*, v. 16, n. 8, pp. e2005343. *https://doi.org/10.1371/journal.pbio.2005343*

García-Puente, María (2020). Search results of the same search strategy perform in 2 different dates. figshare. Dataset.
https://doi.org/10.6084/m9.figshare.12546104.v1

Lee, Young-Ho (2018). "An overview of meta-analysis for clinicians". *Korean journal of internal medicine*, v. 33, n. 2, pp. 277-283.
https://doi.org/10.3904/kjim.2016.195

Moher, David; **Shamseer, Larissa**; **Clarke, Mike**; **Ghersi, Davina**; **Liberati, Alessandro**; **Petticrew, Mark**; **Shekelle, Paul**; **Stewart, Lesley A.**; **Prisma-P Group** (2015). "Preferred reporting items for systematic review and meta-analysis protocols (Prisma-P) 2015 statement". *Systematic reviews*, v. 4, n. 1.
https://doi.org/10.1186/2046-4053-4-1

National Library of Medicine (2020a). *Medline PubMed production statistics*
https://www.nlm.nih.gov/bsd/medline_pubmed_production_stats.html

National Library of Medicine (2020b). *Medline /PubMed data element (Field) descriptions*.
https://www.nlm.nih.gov/bsd/mms/medlineelements.html