 doi:10.20944/preprints202007.0051.v1

World Health Organization (WHO) COVID-19 database: WHO needs it?

Kodvanj I[1], Homolak J[1], Virag D[1], Trkulja V[1]

[1] Department of Pharmacology, University of Zagreb School of Medicine, Zagreb, Croatia

ABSTRACT

A large number of COVID-19 publications has created a need to collect all research-related material in centralized databases. Generating and maintaining such databases regularly, while preserving the quality of the content is challenging, especially considering that the bibliometric databases rely on different data categorization strategies. In this short article, we investigate the functionality and quality of the WHO, PubMed and Scopus databases with a focus on missing values and duplicate entries related to COVID-19. Even though the WHO database is compiled from multiple sources, we conclude that using only the WHO database is not satisfactory as a lot of articles are still available exclusively in other databases. In addition to that, a more careful investigation revealed significant quality problems with all databases in terms of missing values, and many duplicate entries in the WHO database.

Keywords: COVID-19, WHO, database, systematic review, data quality, data

Corresponding author: Ivan Kodvanj (ikodvanj@gmail.com)
Funding: None.
Competing Interests: None.
Ethical Approval: Not applicable.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0051.v1

INTRODUCTION

Coronavirus disease 2019 (COVID-19), caused by the severe acute respiratory syndrome coronavirus virus 2 (SARS-CoV-2), started in Wuhan, China in December 2019. In about six months more than 10 million cases of COVID-19 were identified worldwide, and many lives were lost[1]. An opportunity to put aside differences and unite against the pandemic was seized by many in the academic community. As a result, an unprecedented surge in the number of articles published on this topic followed[2,3]. Publishers had to adapt to the situation in order not to hinder the progress and at the same time to endow themselves with a large number of articles that would be well-cited. Thus, many started publishing ahead of print to make articles available as soon as possible and most made all COVID-19 content open-access to make it accessible to everyone[4]. To illustrate the surge in publications, we point out that in the last 7 months more than 28000 articles related to COVID-19 were published and are so far available in PubMed. This resulted in the need to generate a systematic database of all articles related to COVID-19 in order to take full advantage of the research. The leader in world health, The World Health Organisation (WHO), decided to assist by creating one of the largest databases of COVID-19 research-related data described as “comprehensive multilingual source of current literature on the topic”[5], and updates it from Monday to Friday[5] to keep up with the current publishing trend. Several similar noteworthy attempts exist; for example, the Center for Disease Control and Prevention (CDC) created its database that is updated by “systematically searching various bibliographic databases and hand searching selected grey literature sources”, with WHO database-derived data also included[6]. Another example is the LitCovid database, created and maintained by the National Library of Medicine that contains articles available in the PubMed database[7]. Although the incentive to gather COVID-19-related research data in one place and provide an information platform to accelerate and foster research is praiseworthy, a systematic and thoughtful approach should be imperative. This is especially important in the context of databases organized and maintained by distinguished and respectable healthcare organizations such as WHO and CDC, as it is expected that at least some of the researchers will use their platforms without thorough questioning of the content quality. For this reason, we explored the functionality and quality of the WHO database by comparing it to the widely used PubMed and Scopus databases.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0051.v1

MATERIALS AND METHODS

Data acquisition

The whole global research database on COVID-19 maintained by WHO[5] was downloaded on May 19th and on June 26th 2020. PubMed (using pubmedR[8], search term “COVID-19”) and Scopus (search phrase "COVID-19") databases were accessed on June 26th and June 27th, and accompanying article information was stored. Next, metadata on the retrieved results on 8 specific unrelated topics chosen for demonstrative purposes (azithromycin, chloroquine, depression, diabetes, hypertension, quarantine, shock, tocilizumab) were obtained from the three databases (each topic as an individual search term in the WHO database; each term in conjunction with “AND COVID-19” as a search term in PubMed and Scopus).

Data analysis

In the first step, URLs were converted to DOIs and harmonized in the WHO database. DOIs in all databases were deduplicated and used to compare the databases in respect to a) total number of retrieved articles; b) number of overlapping and c) number of unique articles (not contained in any other database), regarding the “umbrella” term (COVID-19) and in respect to each of the 8 specific topics. In the next step, quality of the databases was assessed by evaluation of the retrieved metadata using two indicators – missing information (considered were all categories of metadata) and duplicate data in the categories in which duplicates are not expected (DOI, ID, URL, abstract). Additionally, duplicates were evaluated based on the combination of title, authors and journal categories. Missing values in abstract, title and keywords categories were hierarchically clustered on complete data. Search terms of specific topics were identified in the metadata retrieved by searching the databases with the same search terms to determine the contribution of categories in identification of articles, followed by clustering to investigate the overlap between search term matches in categories. Analysis of the Scopus metadata was limited to 4000 entries due to the limitations imposed by the Scopus website (we merged 2000 topmost articles sorted by source title alphabetically and 2000 topmost sorted in reverse alphabetical order). Analysis was conducted in R[9], and the entire code and data is available on GitHub[§].

[§] Complete raw data and R code available at: https://github.com/ikodvanj/bibliographicdb

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0051.v1

Terminology

Term “entry” refers to any article in the databases we analysed. A single data point in any column of the databases is considered a value and columns are referred to as categories. *Exclusive* articles are articles contained in only one database, while *shared* are present in more than one.

RESULTS

The WHO database contained the largest number of COVID-19-related articles (36838), followed by PubMed (25700) and Scopus (19451). Following the exclusion of duplicate entries, the total number of articles with a full overlap across the databases was 15302 (Fig 1A). Each database contained a number of *exclusive* articles not included in other databases (the largest number in the WHO database - 6865) (Fig 1A). The total number of such *exclusive* articles across all databases was 9146. However, when we searched each database on specific topics, the number of retrieved results was similar in each database (Fig 1B). Paradoxically, it appeared that the WHO database provided a modest number of articles not available in other databases (Fig 1C), despite having the highest number of total and *exclusive* articles (Fig 1A). In contrast, Scopus had the smallest number of total and *exclusive* articles (Fig 1A), but in 7 out of 8 specific queries it provided the highest number of *exclusive* articles, not available in other databases (Fig 1C).

To investigate this intriguing phenomenon, we evaluated the amount of missing values across the databases. All databases suffered from a significant number of missing values (Fig 2A). Distribution of the missing values across categories of data in the databases is displayed in figure 2B. We were particularly interested in abstracts, keywords (called descriptors in the WHO database) and titles, as information retrieval systems often depend on these categories. A substantial number of data entries were missing in these categories in all the examined databases (Fig 2B). Next, hierarchical clustering of the missing values was performed (Fig 3A) revealing that a substantial proportion of articles in the WHO database is missing both abstract and keywords (descriptors). On the other hand, the proportion of articles with missing abstract and keywords (authors’ and indexed) is much smaller in PubMed and Scopus databases. Additionally, the clustering of matching search terms in abstract, title and keywords categories of search results on the specific topic showed that a small proportion of articles are discovered as a result of matching search terms in more than one category (Fig 3B). Apparently, in WHO and PubMed,

search term was dominantly matched in abstract category, whereas in Scopus it was matched in ID category.

Since we expected to find missing values in these categories in some of the publications (letters, comments, opinions, etc.), we filtered the entries based on the document type, excluding all non-original research article types. This was performed only for Scopus and PubMed because the WHO database does not have a category that specifies the type of entry. Analysis of the filtered data indicated that the abstracts were still missing for a significant proportion of entries in PubMed and Scopus (37.998% and 19.62% respectively). In order to better understand the etiology of this problem, we further investigated the articles with missing abstract data. A closer look at the identified filtered articles revealed that a significant proportion of them were in fact wrongly classified as journal articles while they were actually letters[**], opinions, etc[††].

Finally, it appeared that a considerable number of entries in the WHO database were duplicates. Figure 2C, shows the percentage of values identified as duplicates in categories of the WHO database in which duplicate values are not expected. More than 12% of the entries in the Full text URL, Abstract and the combination of Authors, Title and Journal category were identified as duplicate entries. However, only 4.56% of the entries in the DOI category were identified as duplicates. We hypothesize that the difference between other categories and the DOI category is due to a high percentage of missing DOI values. In addition, we noticed inconsistencies in the way the values were entered. Based on the proportion of duplicate values across different categories, we approximate that at least 10% of the entries in this database are duplicates. The list of DOI-based duplicate entries is provided in **supplement 1**.

[**] 5270th row of the PubMed database with doi: 10.1001/jama.2020.10125
[††] 6591th row of the PubMed database with doi:10.1136/bmj.m2119

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0051.v1

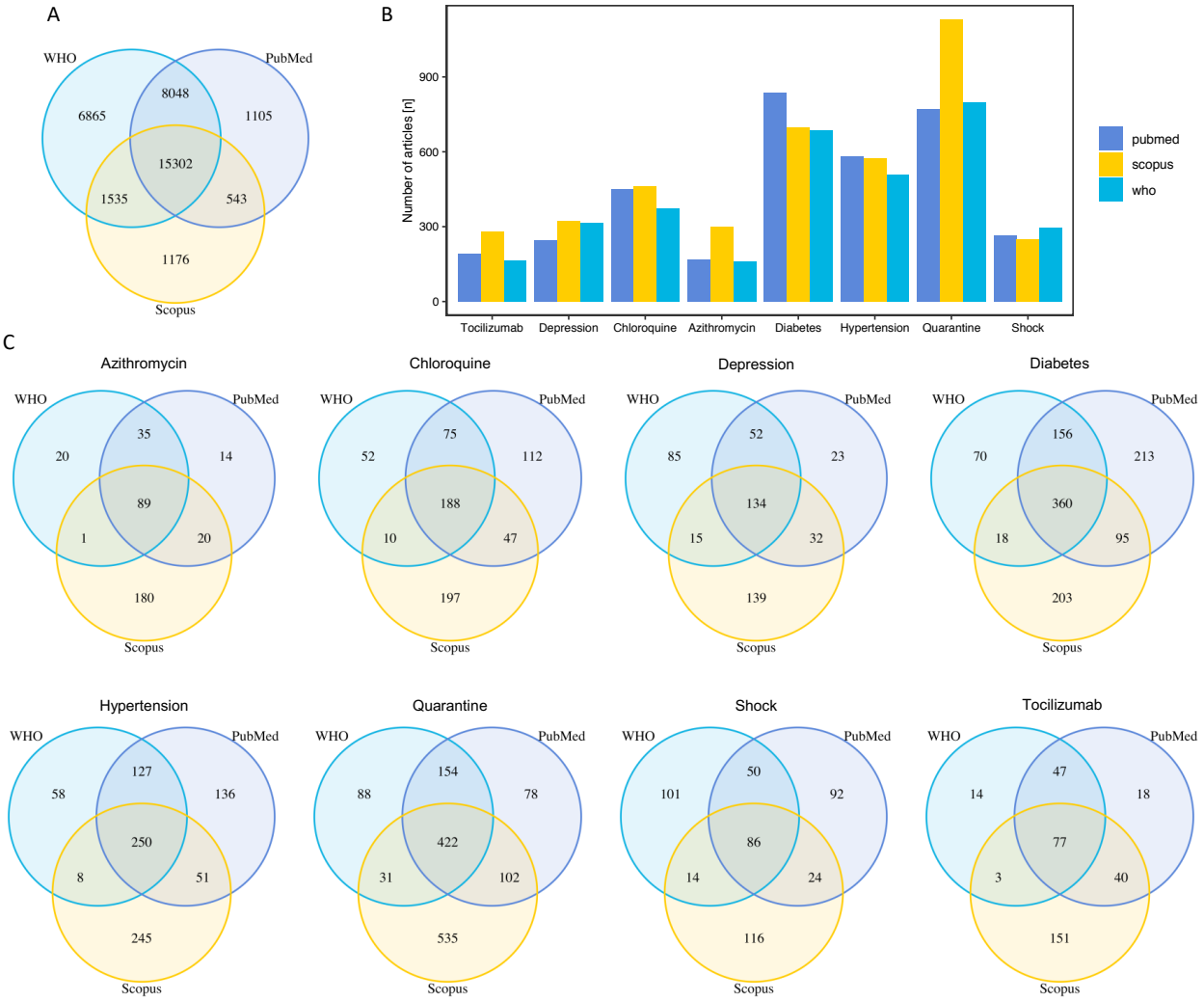

Figure 1. Number of articles indexed in different databases. A) Number of articles in different databases displayed as Venn's diagrams. Numbers at intersections depict articles present in multiple databases. **B)** Numbers of search results across the three databases for different specific queries. **C)** Venn's diagrams for specific queries.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0051.v1

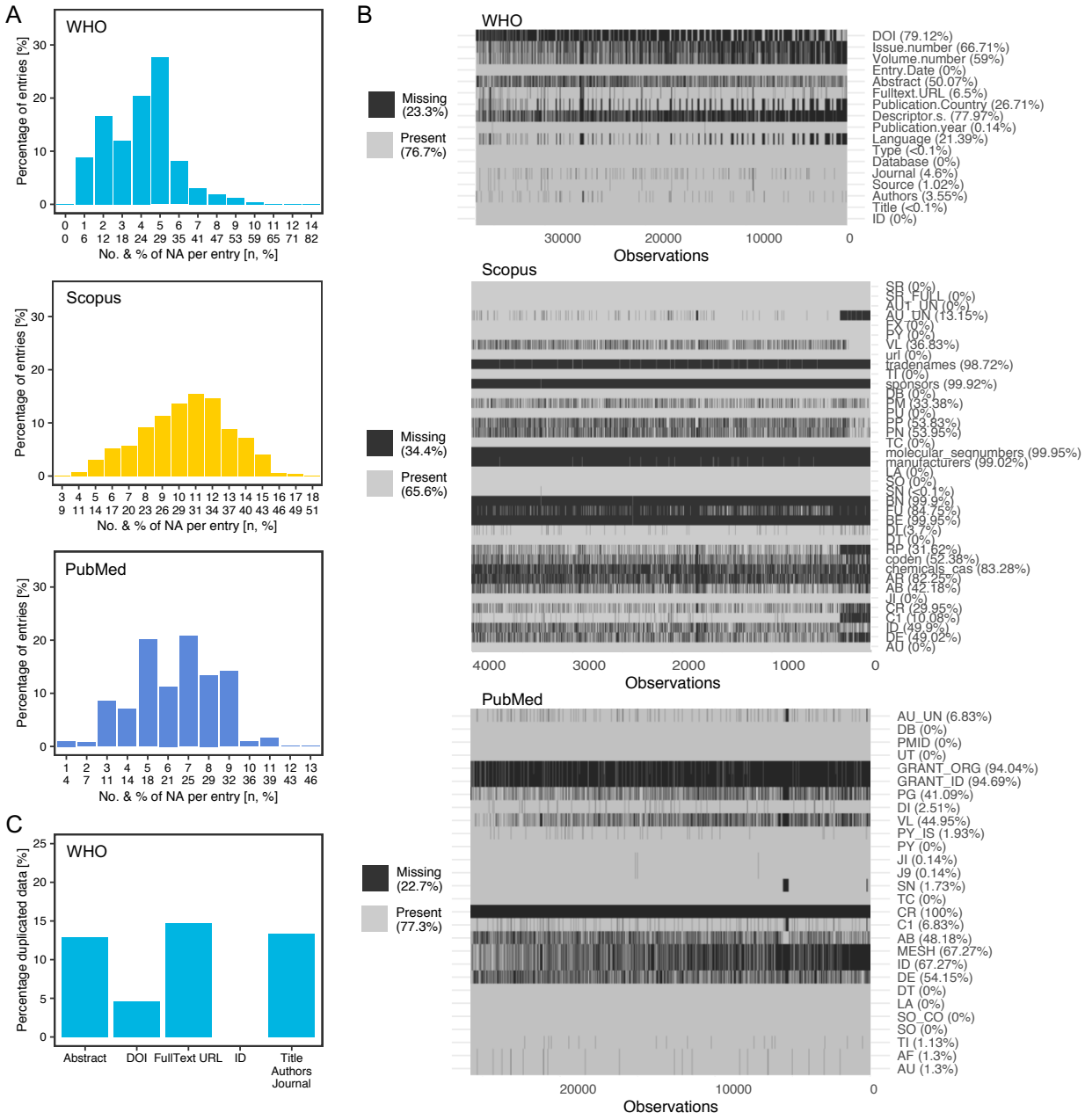

Figure 2. Missing and duplicate values. A) The percentage of entries with indicated number of missing values. **B)** Distribution across different categories in the database. Exact percentage of missing values is listed next to the title of the category on the left. **C)** The number of duplicate values in different categories of WHO database. *Abbreviations: AB - abstract, AF - author full name, AR - article number, AU - authors, AU_UN - authors affiliation, BE - editors, C1 - author address, CR - cited references, DB - bibliographic database, DE - author keywords, DI - digital object identifier, DT - document type, FU - funding agency and grant number, FX - funding text, ID - indexed keywords, JI - ISO Source Abbreviation, LA - language, PG - page count, PM/PMID - PubMed ID, PU - publisher, PY - published year, RP - reprint address, SN - international standard serial number, SO - publication name, TC - web of science core collection times cited count, TC - WoS core collection times cited count, TI - title, UT - accession number, VL - volume.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0051.v1

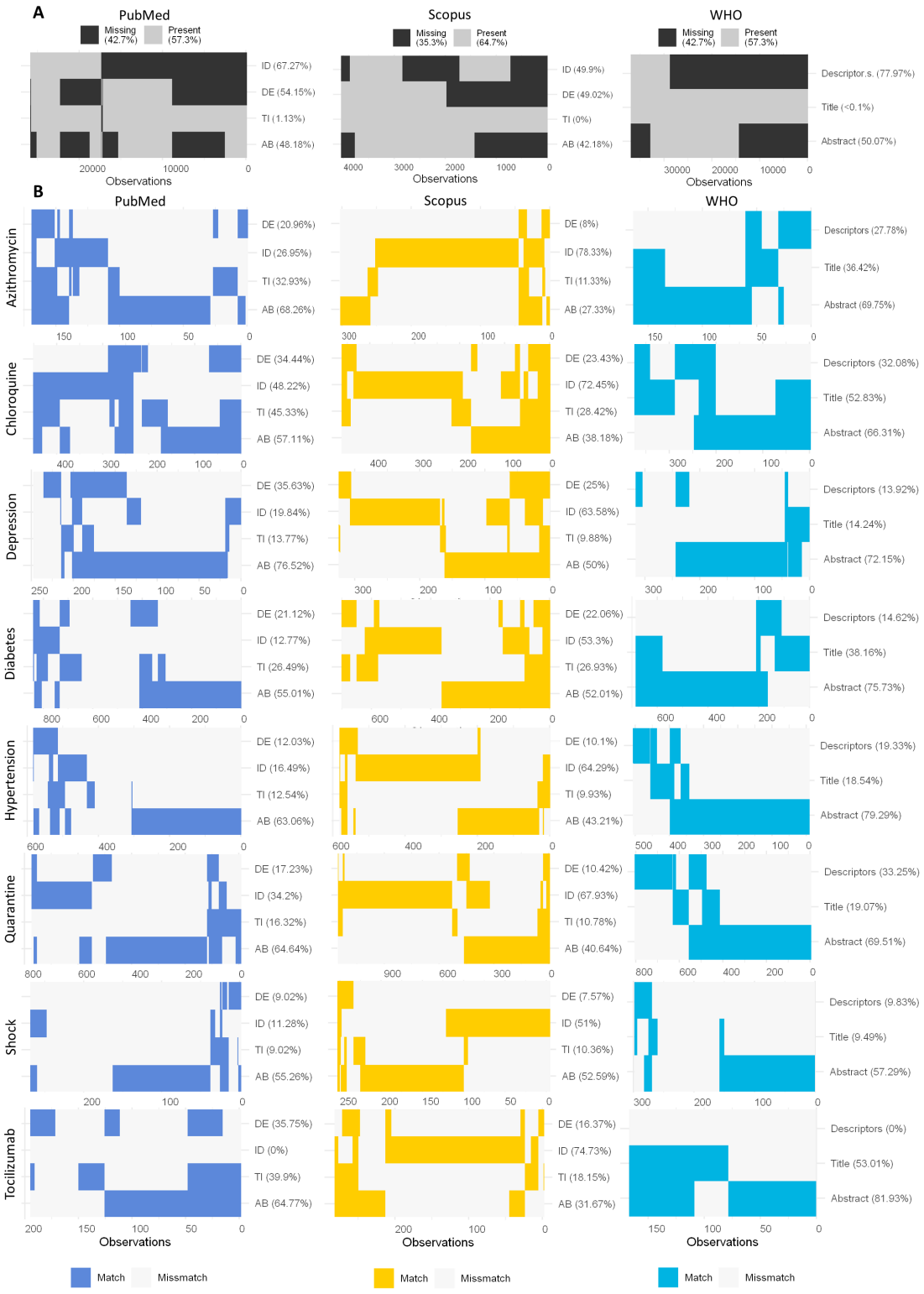

Figure 3. Hierarchical clustering (mcquitty method) of missing values and search term matching. A) Clustering of missing values for abstract, title and keywords (descriptors, DE, ID) categories. **B)** Clustering of search terms matches in abstract, title and keywords categories of results of specific topic. *Abbreviations: AB - abstract, DE - authors keywords, ID - indexed keywords, DE - authors keywords.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0051.v1

DISCUSSION

Quick, simple and reliable access to knowledge on a specific practical question relevant in daily practice, or broader general information on a topic to comprehensively evaluate the existing evidence-base or evidence gaps through, e.g. a systematic review, is a constant need in healthcare. It is particularly so under the circumstances of the COVID-19 pandemic with exponentially generated observations on its various aspects, with still growing numbers of patients worldwide and with healthcare workers and policy makers exposed to a tremendous daily workload. Traditional approach to literature search involves use of multiple databases and can be very time-consuming[10]. It is reasonable to assume that having a database set-up on a specific topic and compiled from multiple sources would simplify this process as it would contain all articles that are present in other databases. The present analysis was undertaken from a user's perspective to evaluate whether using the global research database on COVID-19, maintained by the WHO, would indeed enable one to reliably access the desired data without a need for systematic searches of other bibliographic databases. Unfortunately, the WHO database apparently suffers from significant problems with data quality, and there seems to be quite some information on the COVID-19-related topics outside of it that are accessible through "standard" bibliographic databases like PubMed and Scopus.

Expectedly, the WHO database contained the largest number of publications since it is compiled from different sources; however, we were somewhat surprised that there are 2281 articles found exclusively in PubMed or Scopus. Equally unexpected (and for the same reason) was the finding of (only) 15302 publications *shared* by all three databases, while a total of 9146 were found exclusively in individual databases. In contrast, the searches on the 8 specific topics retrieved similar numbers of articles in all three databases – another finding that did not seem logical, since the WHO database contained a considerably larger total number of papers and the number of *exclusive* papers than the other two. Interestingly, the number of COVID-19 related articles in PubMed is larger than Scopus even though the number of journals indexed in Scopus (41,154) is larger than in PubMed (about 30,000)[11,12]. Further analysis suggested data missingness is a likely explanation of disproportionate number of retrieved results on specific topics for the following reasons: Scopus and Pubmed have lower proportion of articles with missing both authors and indexed keywords (about 30%), and provided higher number of *exclusive* articles (on specific topics); while 77% of articles the WHO database had missing values in the keywords (descriptors) category, and returned fewer number of articles on specific topics.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 July 2020 doi:10.20944/preprints202007.0051.v1

Additionally, a higher proportion of missing values in the indexed keywords and abstract categories of PubMed database (Fig 3A) explains why Scopus retrieved more articles than PubMed, relative to its size, as many articles in Scopus were retrieved as a result of matching the search term and indexed keywords.

At this point, we need to emphasize that all three databases differed in the missingness of values in other shared categories, as well. However, it is difficult to contextualize the meaning of missing values in some categories and draw conclusions about the quality of databases, as the importance of categories is in the eye of the beholder: a researcher performing bibliometric analysis might find a list of the authors and journal names very important, while categories that describe the content of an article (title, keywords and abstract) might be more important to a researcher doing a systematic review as these categories are searched by information retrieval systems. On the other hand, some categories serve multiple purposes and likely affect all users of the database. DOI and other unique identifiers are such categories as they can be used for finding full texts of articles and for tidying-up databases (i.e. removal of duplicate entries). Other categories, such as pages, volume and issue are arguably not as important or needed in compiled databases as this information can be found elsewhere and do not directly affect the functionality of the database. Thus, in our opinion, curators of these databases should prioritize while collecting information on the articles, depending on the intended use of the database, to avoid missing values in the database. This is especially important in response to the current global emergency when limited resources might affect the number of missing values. In older versions of the WHO database, there were more categories and more missing values. By removing non-essential categories more effort could be put into collecting article information and missing values could be reduced. Nevertheless, only small improvements can be seen despite removing some of the categories. Of note, we were unable (despite a thorough and enthusiastic search) to find a (concise) statement about the intended use of the WHO database. We believe that this is not a good practice as using databases for non-matching purposes might result in erroneous conclusions. If one is to assume that the final goal is to create a searchable database intended for clinicians and researchers, then more effort should be invested in reducing the number of missing values of article metadata (abstract, title, keywords), as missing values in these categories might impair the functionality of the database if traditional information retrieval systems are used. Unique identifiers must be provided for all articles because of their already mentioned uses – identification of duplicates and finding full text. Latter is not only important for the users of the database but also for retrieval of full text for the purposes of text mining. This is especially

important as it has been hypothesised that full text mining might facilitate and simplify the identification of topic-relevant articles in bibliographic databases when used as an alternative or in conjunction with classic Boolean search strategies[13,14]. Several such noteworthy attempts exist[15,16], and some of them rely on other databases (e.g. the WHO database). Thus, the needs of such projects should be considered when planning the structure of new databases, as redesigning the structure of established databases is an enormous task. Furthermore, inclusion of a special category dedicated to the compounds (drugs) used in the studies could be beneficial and result in a more precise identification of articles of interest. We deem this category to be very important, as it provides the user with a way to identify the studies of interest based on the compounds, even if these compounds are not mentioned in the title, abstract or keywords. Also, it is a fail-safe in case one of these categories is missing. This is where the Scopus database stands-out as it has a similar category. However, it is important to acknowledge that the curators of the WHO database had a similar idea with the descriptors category. Filtering articles based on this category provides an attractive approach to the identification of articles of interest; however, the high prevalence of missing values in these categories does not only reflect poor quality of the database but can easily result in forming biased conclusions. Finally, we want to draw attention to the inconsistencies and duplicate entries in the database and emphasize the need for caution when using this database as a source of articles for bibliometric analysis. In older versions of the database, we noticed a lot of inconsistencies in the use of delimiters and the ways of writing of the authors', journals' names, and DOIs. Nevertheless, WHO actively works on enhancing the quality of the database and improvements are obvious; however, sparse inconsistencies are still present, and at least 10% of entries are duplicates. Additionally, when comparing the older and newer version of the WHO database, changes in the database structure are apparent. Although changes might have been necessary, at this moment several other databases rely on the WHO COVID-19 global research database, and we are concerned that unannounced changes in the database structure could easily result in significant impairment of data quality and structure in other databases. For example, the CDC database[7] and CORD-19[15] are compiled from multiple databases, among them the WHO database, allowing the propagation of errors from one database to another if the content is not checked.

The present work suffers from several limitations: a) we did not manually check whether each of the entries in the provided supplement was a duplicate or not, still, most of the entries obviously were. Supplement is provided for the demonstrative purposes only and their number is just an estimation; b) we identified duplicates based on DOIs that were extracted from the

 doi:10.20944/preprints202007.0051.v1

“FullText URL” in the WHO database, since DOIs in the DOI category are missing in WHO database. This is not ideal as approximately 6.95% of full text URLs are missing and about 0.7% of URLs are not derived from DOIs; c) Another obvious limitation is the time point of the analysis. For this reason, we performed an analysis of the WHO database at two time points (results for older version are provided in supplement 2); d) analysis displayed in Figure 3B revealed that some retrieved articles do not have search term mentioned in any categories, suggesting that (i) the exported article metadata is partial, or (ii) information retrieval system is searching other fields (despite specifically selecting abstract, title and keywords) that are not exported from the WHO database.

In conclusion, under the circumstances of the COVID-19 pandemic, centralization of all pertinent research-related material would be beneficial as it would facilitate dispersion of information. The attempt to accomplish such a task is first and foremost brave and admirable. Still, it stands to reason that the real challenge is not to merge the data from different sources, but to keep the newly created database clean as this affects the functionality. From the standpoint of a researcher interested in using WHO database as a bibliographic database, it is worrisome that we retrieved more results with queries on PubMed and Scopus, than from the WHO global research database. Thus, we conclude that the WHO database alone is not sufficient as a source of information, even though it is compiled from multiple sources by a very respected and trustworthy organization.

REFERENCES

1. Roser, M., Ritchie, H., Ortiz-Ospina, E. & Hasell, J. Coronavirus Pandemic (COVID-19). https://ourworldindata.org/coronavirus (2020).
2. Chahrour, M. *et al.* A Bibliometric Analysis of COVID-19 Research Activity: A Call for Increased Output. *Cureus* **12**, e7357 (2020).
3. Tao, Z. *et al.* COVID-19 will stimulate a new coronavirus research breakthrough: a 20-year bibliometric analysis. *Ann Transl Med* **8**, 528 (2020).
4. Homolak, J., Kodvanj, I. & Virag, D. Preliminary Analysis of COVID-19 Academic Information Patterns: A Call for Open Science in the Times of Closed Borders. *Scientometrics* (2020) doi:10.1007/s11192-020-03587-2.
5. Global research on coronavirus disease (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov.
6. COVID-19 Research Articles Downloadable Database. https://www.cdc.gov/library/researchguides/2019novelcoronavirus/researcharticles.html (2020).
7. LitCovid - NCBI - NLM - NIH. https://www.ncbi.nlm.nih.gov/research/coronavirus/.
8. Aria, M. *Gathering Metadata About Publications, Grants, Clinical Trials from 'PubMed' Database [R package pubmedR version 0.0.2]*. (Comprehensive R Archive Network (CRAN), 2020).
9. R Core Team. *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. (2020).
10. Bramer, W. M., Rethlefsen, M. L., Kleijnen, J. & Franco, O. H. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Syst. Rev.* **6**, 245 (2017).

 doi:10.20944/preprints202007.0051.v1

11. Sources. https://www.scopus.com/sources.uri?zone=TopNavBar&origin=searchbasic.
12. List of All Journals Cited in PubMed®. https://www.nlm.nih.gov/bsd/serfile_addedinfo.html.
13. Lefebvre, C., Glanville, J., Wieland, L. S., Coles, B. & Weightman, A. L. Methodological developments in searching for studies for systematic reviews: past, present and future? *Syst. Rev.* **2**, 78 (2013).
14. Thomas, J., McNaught, J. & Ananiadou, S. Applications of text mining within systematic reviews. *Res Synth Methods* **2**, 1–14 (2011).
15. Wang, L. L. *et al.* CORD-19: The Covid-19 Open Research Dataset. (2020).
16. Allen Institute For AI. COVID-19 Open Research Dataset Challenge (CORD-19). https://kaggle.com/allen-institute-for-ai/CORD-19-research-challenge.