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

Introduction: A large number of COVID-19 publications has created a need to collect all research-related material in practical and reliable centralized databases. The aim of this study was to evaluate the functionality and quality of the compiled World Health Organisation COVID-19 database and compare it to Pubmed and Scopus.

Methods: Article metadata for COVID-19 articles and articles on 8 specific topics related to COVID-19 was exported from the WHO global research database, Scopus and Pubmed. The analysis was conducted in R to investigate the number and overlapping of the articles between the databases and the missingness of values in the metadata.

Results: The WHO database contains the largest number of COVID-19 related articles overall but retrieved the same number of articles on 8 specific topics as Scopus and Pubmed. Despite having the smallest number of exclusive articles overall, the highest number of exclusive articles on specific COVID-19 related topics was retrieved from the Scopus database. Further investigation revealed that PubMed and Scopus have more comprehensive structure than the WHO database, and less missing values in the categories searched by the information retrieval systems.

Discussion: This study suggests that the WHO COVID-19 database, even though it is compiled from multiple databases, has a very simple and limited structure, and significant problems with data quality. As a consequence, relying on this database as a source of articles for systematic reviews or bibliometric analyses is undesirable.

Key words: COVID-19, WHO, database, systematic review, data quality

Correspondance: Ivan Kodvanj, ikodvanj@gmail.com

INTRODUCTION

In response to the Coronavirus disease (COVID-19) an unprecedented number of articles were published[1,2]. Publishers adapted to the situation not to hinder the progress and to endow themselves with a large number of articles that would be well-cited. Thus, many started publishing open-access, *submission-to-publication* time reduced and a lot of articles were published ahead-of-print to make them available sooner[3]. The surge in publications resulted in the need to generate a systematic database of all COVID-19 related articles to full advantage of the research. The leader in world health, the World Health Organisation (WHO), created one of the largest databases of COVID-19 research-related databases described as "comprehensive multilingual source of current literature on the topic"[4]. Several similar noteworthy attempts exist including the database maintained by the Center for Disease Control and Prevention (CDC)[5] and the LitCovid database created by the National Library of Medicine[6]. 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 (WHOdb) by comparing it to the widely used PubMed and Scopus databases.

METHODS

Data acquisition

The whole global research database on COVID-19 maintained by WHO[4] was downloaded on May 19th and on June 26th 2020. PubMed (using pubmedR[7], search term "COVID-19") and Scopus (search phrase "COVID-19") databases were accessed on June 26th and June 27th, and accompanying article metadata was stored. The procedure was repeated for the retrieved results on 8 specific unrelated topics (terms azithromycin, chloroquine, depression, diabetes, hypertension, quarantine, shock, tocilizumab were used as an individual search terms in the WHOdb; 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 WHdb. DOIs in all databases were deduplicated and used to compare the databases in respect to a) the total number of retrieved articles; b) number of overlapping and c) the 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, and 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 identify the contribution of categories in the identification of articles, followed by clustering to investigate the overlap between search term matches in categories. The analysis was conducted in R[8], and the entire code and data are available on GitHub[9].

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. The terms keywords and descriptors are used interchangeably in the context of the WHOdb.

RESULTS

The WHOdb 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 WHOdb - 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 WHOdb 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).

Further investigation of this phenomenon revealed that all databases suffer 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. Special attention was directed at abstracts, keywords (called descriptors in the WHOdb) and title categories, as information retrieval systems (IRS) often depend on them. 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 WHOdb 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 were discovered as a result of matching search terms in more than one category (Fig 3B). Apparently, in the WHO and PubMed databases, the search terms were dominantly matched in the abstract category, whereas in Scopus it was matched in the ID category (indexed keywords).

Since the missing values were expected in certain 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 WHOdb does not have a category that specifies the type of entry. This analysis indicated that the abstracts were still missing for a significant proportion of entries in PubMed and Scopus (37.998% and 19.62% respectively). However, 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 WHOdb were duplicates. Figure 2C, shows the percentage of values identified as duplicates in categories of the WHOdb 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 DOI-based duplicate list is provided in **supplement 1)**.

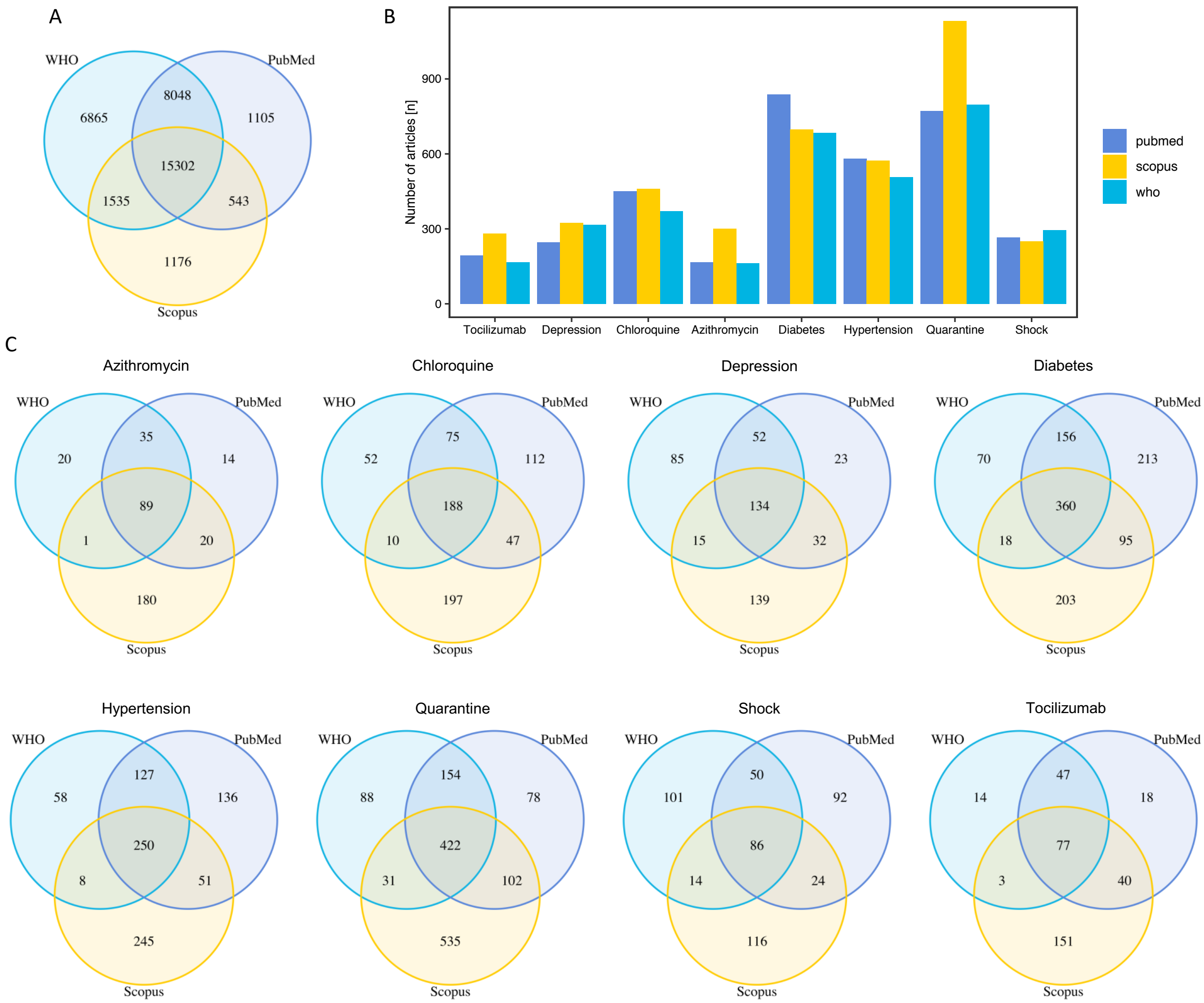

Figure 1. The number of articles indexed in different databases. A) The number of articles in different databases displayed as Venn's diagrams. Numbers at intersections depict articles present in multiple databases. **B)** Numbers of search results across the three databases for different specific queries.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

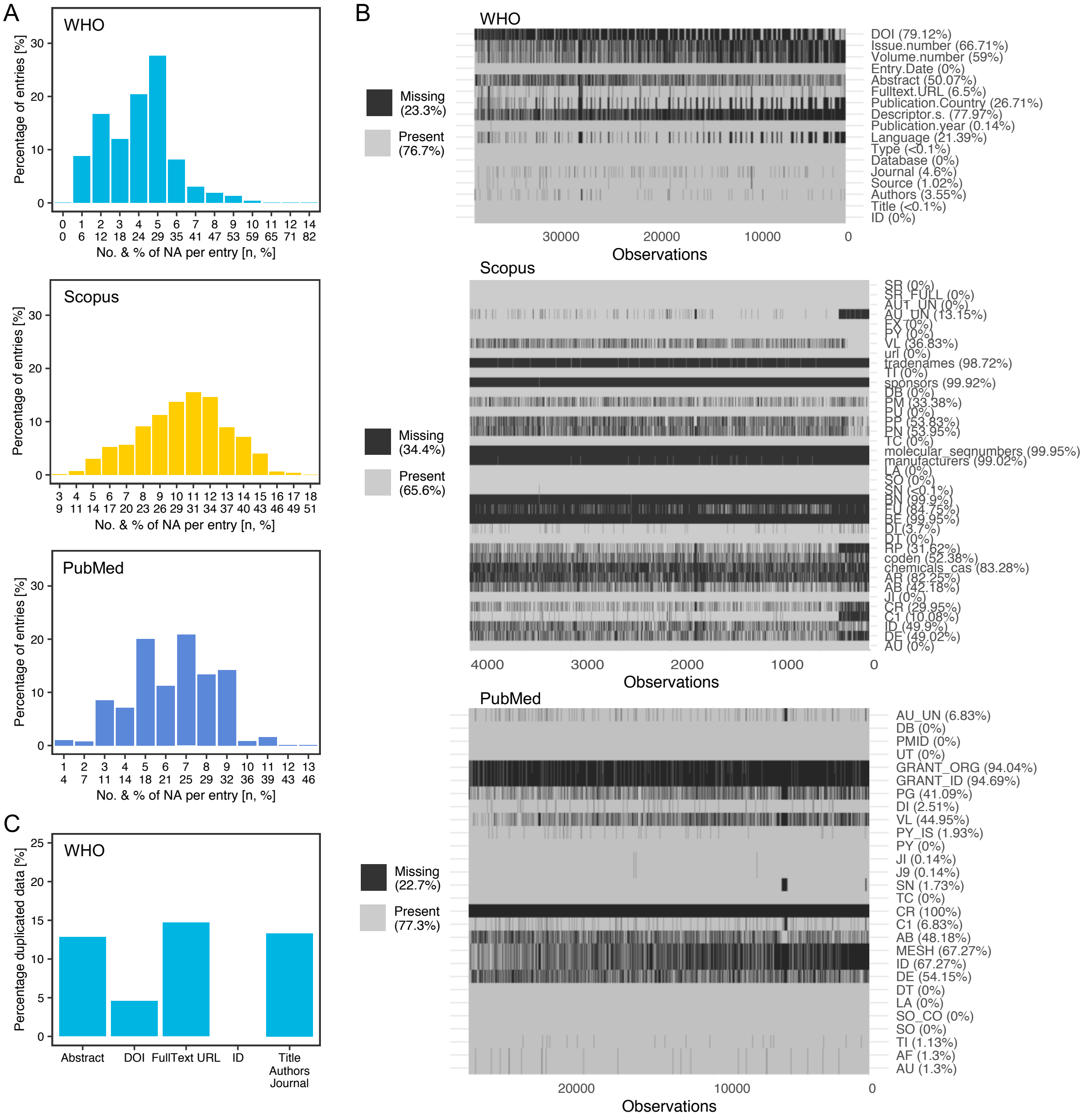

Figure 2. Missing and duplicate values. A) The percentage of entries with the indicated number of missing values. **B)** Distribution across different categories in the database. The exact percentage of missing values is listed next to the title of the category on the left. **C)** The number of duplicate values in different categories of the 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.*

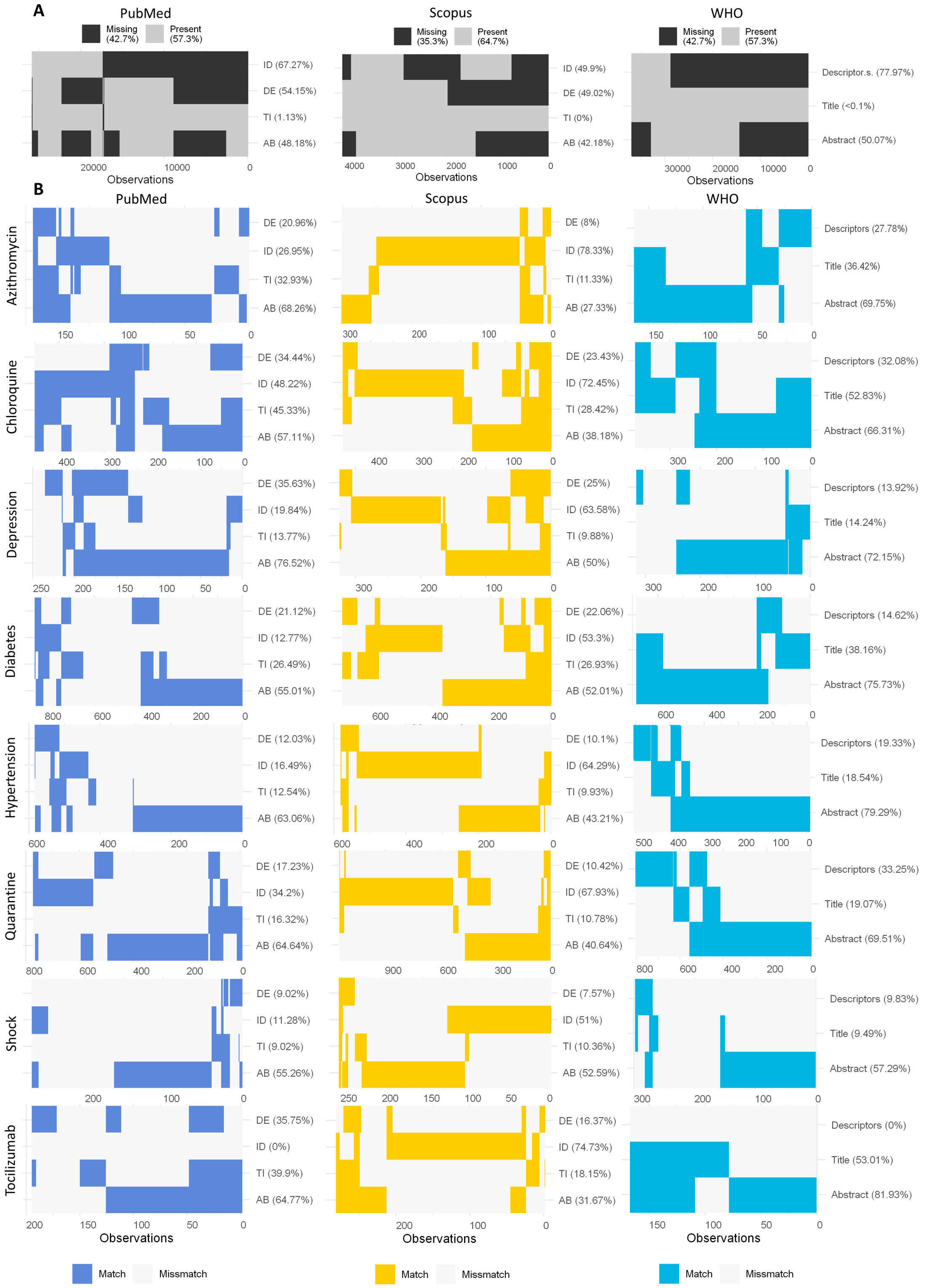

Figure 3. Hierarchical clustering (mcquitty method) of missing values and search term matching. A) Clustering of missing values for abstract, title and keywords (descriptors, DE, ID) categories. **B)** Clustering of search terms matches in abstract, title and keywords categories of results of the specific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

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. The traditional approach to literature search involves the 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 WHOdb 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. Thus using only the WHOdb as a source of articles for systematic reviews is undesirable. This problem was already encountered by other authors, for example, Viner et al.[11] reported that the WHOdb retrieved only one article that was excluded because it did not match the topic of their research.

Expectedly, the WHOdb contained the largest number of publications since it is compiled from different sources; however, we were 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. Conversely, the search on the 8 specific topics retrieved similar numbers of articles in all three databases – another unexpected finding, since the WHOdb contained a considerably larger number of total and *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)[12,13]. Further analysis suggested data missingness and having two keywords categories in the Scopus and PubMed databases are a likely explanation of a disproportionate number of retrieved results on specific topics relative to the size of the databases for the following reasons: Scopus and Pubmed have a lower proportion of articles with missing both authors and indexed keywords (about 30%, i.e. about 70% of articles have keywords), and provided a higher number of *exclusive* articles (on specific topics); while 77% of articles the WHOdb had missing values in the keywords (descriptors) category, and returned a fewer number of articles on specific topics. 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.

The missingness of data should be contextualized based on the intended use of the database. If it is assumed that the WHOdb is intended to be used by clinicians and researchers as a bibliographic source, then a priority should be reducing the missing values in categories searched by IRS as this directly affects the functionality of the database. Following the evolution of the WHOdb, it is clear that some categories were removed from the database rendering the maintenance of the database easier. However, no significant improvement in the data missingness can be seen. On the other hand, adding additional article-descriptive categories to the database should increase its functionality by providing a way to filter the articles and to act as a failsafe in case there are missing values in other categories (e.g. Scopus has a special category dedicated to the compounds (drugs) used in the studies). Filtering articles based similar categories provides an attractive approach to the identification of articles of interest as it increases the probability of matching an article with search terms or provides an additional way to filter the search results; however, one should be aware that the high prevalence of missing values in these categories does not only reflect poor quality but can easily result in forming biased conclusions.

Additionally, unique identifiers must be provided for all articles to allow for the identification of duplicates and deriving the full-text URL. Latter is not only important for the users of the database but also for retrieval of full-text for 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[14,15]. Several such noteworthy attempts exist[16,17], and some on the WHOdb. Thus changing the structure of the database might affect the other databases that depend on it.

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. This is especially prominent in older versions of the WHOdb database, where a lot of inconsistencies were noticed in the use of delimiters and the way of writing of the authors' and journals' names, and DOIs.

Limitations

The present work suffers from several limitations: a) since DOIs in the DOI category are missing, DOIs were extracted from the "FullText URL" in the WHOdb to identify the duplicates and investigate overlapping (this is not ideal as approximately 6.95% of full-text URLs are missing and 0.7% are not derived from DOIs); b) analysis displayed in Figure 3B shows that some articles do not have search term mentioned in any categories, suggesting that (i) the exported article metadata is partial, or (ii) IRS is searching other fields (despite specifically selecting abstract, title and keywords) that are not exported from the WHOdb; c) limitations imposed by the Scopus website limited the dataset used for assessing the quality to 4000 articles (2000 topmost articles sorted by source title alphabetically and in reverse order were merged).

Conclusion

In conclusion, under the circumstances of the COVID-19 pandemic, centralization of all pertinent research-related material would be beneficial as it would facilitate the dispersion of information and simplify its access. 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 design a good structure of the database and keep it clean as this affects the functionality. From the standpoint of a researcher interested in using the WHOdb as a bibliographic database, it is worrisome that more results were retrieved with queries on PubMed and Scopus than from the WHO global research database. Thus, we conclude that the WHOdb alone is not sufficient as a source of information, even though it is compiled from multiple sources by a very respected and trustworthy organization.

ACKNOWLEDGEMENTS

Contributions: All authors conceptualized the design of the study. IK and DV conducted the data analysis. All authors discussed the results. IK and JH wrote the manuscript and all DV and VT provided valuable suggestions. All authors approved the final version of the manuscript.

Funding: None.

Competing interests: None.

Ethical approval: Not applicable.

REFERENCES

1. Chahrour M, Assi S, Bejjani M, et al. A Bibliometric Analysis of COVID-19 Research Activity: A Call for Increased Output. *Cureus*. 2020;12(3):e7357. doi:10.7759/cureus.7357

2. Tao Z, Zhou S, Yao R, et al. COVID-19 will stimulate a new coronavirus research breakthrough: a 20-year bibliometric analysis. *Ann Transl Med*. 2020;8(8):528. doi:10.21037/atm.2020.04.26

3. 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*. June 2020. doi:10.1007/s11192-020-03587-2

4. Global research on coronavirus disease (COVID-19). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov. Accessed June 26, 2020.

5. COVID-19 Research Articles Downloadable Database. https://www.cdc.gov/library/researchguides/2019novelcoronavirus/researcharticles.html. Published 25 June 2020. Accessed June 26, 2020.

6. LitCovid - NCBI - NLM - NIH. https://www.ncbi.nlm.nih.gov/research/coronavirus/. Accessed June 26, 2020.

7. 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. https://CRAN.R-project.org/package=pubmedR. Accessed June 26, 2020.

8. R Core Team. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*.; 2020. https://www.r-project.org/. Accessed June 26, 2020.

9. ikodvanj. ikodvanj/bibliographicdb. GitHub. https://github.com/ikodvanj/bibliographicdb. Accessed July 12, 2020.

10. Bramer WM, Rethlefsen ML, Kleijnen J, Franco OH. Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study. *Syst Rev*. 2017;6(1):245. doi:10.1186/s13643-017-0644-y

11. Viner RM, Russell SJ, Croker H, et al. School closure and management practices during coronavirus outbreaks including COVID-19: a rapid systematic review. *The*

Lancet Child & Adolescent Health. 2020;4(5):397-404. doi:10.1016/S2352-4642(20)30095-X

12. Sources. https://www.scopus.com/sources.uri?zone=TopNavBar&origin=searchbasic. Accessed June 26, 2020.

13. List of All Journals Cited in PubMed®. https://www.nlm.nih.gov/bsd/serfile_addedinfo.html. Accessed June 26, 2020.

14. Lefebvre C, Glanville J, Wieland LS, Coles B, Weightman AL. Methodological developments in searching for studies for systematic reviews: past, present and future? *Syst Rev*. 2013;2:78. doi:10.1186/2046-4053-2-78

15. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. *Res Synth Methods*. 2011;2(1):1-14. doi:10.1002/jrsm.27

16. Wang LL, Lo K, Chandrasekhar Y, et al. CORD-19: The Covid-19 Open Research Dataset. April 2020. http://arxiv.org/abs/2004.10706. Accessed June 16, 2020.

17. Allen Institute For AI. COVID-19 Open Research Dataset Challenge (CORD-19). https://kaggle.com/allen-institute-for-ai/CORD-19-research-challenge. Accessed June 16, 2020.