BMJ Health & Care Informatics

‘Refbin’ an online platform to extract and classify large-scale information: a pilot study of COVID-19 related papers

Shania Lunna,[1] Isabelle Flinn,[1] James Prytherch,[2] Camille Torfs-Leibman,[1] Sarah Robtoy,[1] Matt Bansak,[2] David Krag [3]

To cite: Lunna S, Flinn I, Prytherch J, *et al*. ‘Refbin’ an online platform to extract and classify large-scale information: a pilot study of COVID-19 related papers. *BMJ Health Care Inform* 2022;**29**:e100452. doi:10.1136/bmjhci-2021-100452

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100452).

Received 30 August 2021
Accepted 13 February 2022

[1]University of Vermont, Burlington, Vermont, USA
[2]Plomics, Shelburne, Vermont, USA
[3]Surgery, University of Vermont, Burlington, Vermont, USA

Correspondence to
Dr David Krag;
david.krag@uvm.edu

ABSTRACT

Introduction The number of new biomedical manuscripts published on important topics exceeds the capacity of single persons to read. Integration of literature is an even more elusive task. This article describes a pilot study of a scalable online system to integrate data from 1000 articles on COVID-19.

Methods Articles were imported from PubMed using the query ‘COVID-19’. The full text of articles reporting new data was obtained and the results extracted manually. An online software system was used to enter the results. Similar results were bundled using note fields in parent–child order. Each extracted result was linked to the source article. Each new data entry comprised at least four note fields: (1) result, (2) population or sample, (3) description of the result and (4) topic. Articles underwent iterative rounds of group review over remote sessions.

Results Screening 4126 COVID-19 articles resulted in a selection of 1000 publications presenting new data. The results were extracted and manually entered in note fields. Integration from multiple publications was achieved by sharing parent note fields by child entries. The total number of extracted primary results was 12 209. The mean number of results per article was 15.1 (SD 12.0). The average number of parent note fields for each result note field was 6.8 (SD 1.4). The total number of all note fields was 28 809. Without sharing of parent note fields, there would have been a total of 94 986 note fields.

Conclusion This pilot study demonstrates the feasibility of a scalable online system to extract results from 1000 manuscripts. Using four types of notes to describe each result provided standardisation of data entry and information integration. There was substantial reduction in complexity and reduction in total note fields by sharing of parent note fields. We conclude that this system provides a method to scale up extraction of information on very large topics.

INTRODUCTION

With >1 000 000 new articles per year[1 2] it is nearly impossible to comprehensively assimilate newly published literature.[3 4] PubMed and Google Scholar are powerful search tools, but the product is only a set of citations. Systematic reviews have a short life cycle[5] and the process of establishing continuity between reviews is not well defined. Tools that facilitate reading, describing and integrating data are limited. Linking results across multiple manuscripts is difficult.[3] Tagging articles with keywords may help bundle sets of manuscripts but is a poor substitute for quality extraction of results.

We present pilot results extracting and integrating data from 1000 COVID-19 publications using a new online system and discuss the scalability and public dissemination of the information.

METHODS

A single account was used for the COVID-19 database (Refbin.com, Plomics, Shelburne, Vermont). Refbin retrieved citations from PubMed using COVID-19 as the search term. Articles that reported new data were selected (1000 of 4236). The results were manually extracted from full-text articles. Reviewers included six undergraduates, five postgraduate and three faculty. The average number of articles per reviewer was 132 (SD 243).

The extracted results were entered manually into the COVID-19 library as independent units of information. Each result was described by a set of note fields. The note fields were arranged in parent–child order and included (1) result, (2) description of the result, (3) population and (4) topic. The result note field occupied the lowest child note field. The citation was dragged to the result note field and was automatically linked to all parent note entries (see online supplemental methods for additional details).

RESULTS

From 1000 publications, 12 209 individual results were extracted. These were organised under 15 first-level topic note fields (online supplemental figure 1). The order, pattern and total number of parent notes varied for each extracted result (figure 1). The average number of note fields required to describe an

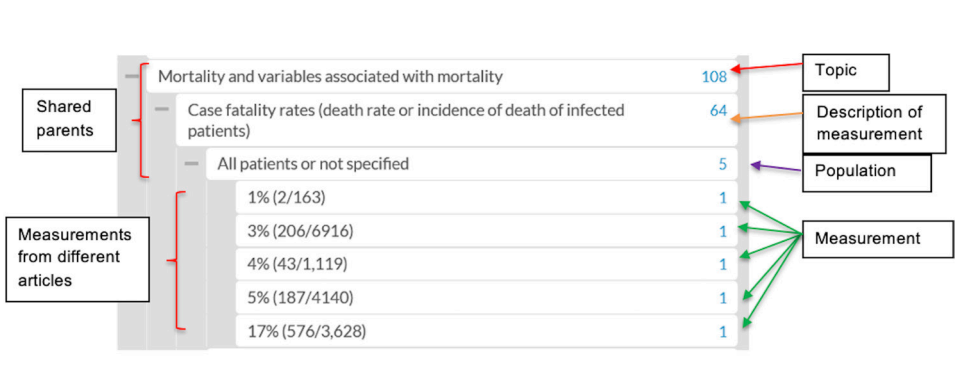


Figure 1 Screenshot of multiple entries of case fatality rates at the same hierarchical level sharing the same parents and four types of note fields. The number on the right edge of the note field is a citation counter. The parent note fields display the sum of citations affiliated with its children.

extracted result was 7.78 (SD 1.42) (online supplemental table 1).

The results from one manuscript were entered into multiple relevant locations throughout the database. The results from multiple manuscripts with similar features, for example case fatality rates, were bundled together. These similar results were entered into the same note field level and then could share parents (figure 1).

Sharing of parent note fields allowed substantial reduction in the total number of note fields and simplification of data entry. Without sharing of parents, the total number of all note fields would have been 94 986. With sharing of parents, the total number of note fields was reduced ~70% to 28 613 total note fields. This represents an absolute reduction of 66 373 note fields (online supplemental figure 2).

To assess communication tools, 719 corresponding authors from 1000 publications were invited by email to obtain a personal Refbin account that included a read-only live portal to the COVID-19 library. This was accepted by 21% (148 of 719) and represented authors from 22 different countries. A publicly available website (COVIDpublications.org) was also created that displayed a live version of the COVID-19 library.

DISCUSSION

Four types of note fields were sufficient to describe each of the wide variety of results extracted from 1000 articles. Similar results shared parent note fields. This facilitated integration of data by creating frameworks to bundle results from different articles. This resulted in 70% reduction in the total number of note fields.

Online review methods minimised the frequency of scheduled meetings. Online group discussions were reserved for more challenging issues. Treating the extracted results as independent units of information allowed the results from one article to be placed in different library locations. This also allowed similar results from multiple articles to be integrated into a single location.

This pilot study describes a system that facilitates multiple individuals to read, describe and integrate results in a scalable manner. A live portal to the COVID-19 library was initiated for 148 researchers from 22 countries. A public version of the COVID-19 library was accomplished (COVIDpublications.org). This sets the stage for researchers to work together on areas of common interest and enhanced sharing of information across international borders.

Limitations of this study include dependency on the narrative language to describe a result. This limitation is mitigated by using the four note field types to describe each result, which helps standardise descriptions. Manual extraction of information is time-consuming. However, by providing the information as in this pilot study, time investment by a larger population of users may potentially be reduced. Reproducibility is a potential limitation. This was somewhat mitigated in this pilot study by using extraction rules and oversight from multiple reviewers. Formal assessment of reproducibility will be a goal of future work.

Large-scale healthcare issues such as COVID-19 or opioid use disorder have dramatic adverse effects on personal and global health. The efforts of thousands of researchers worldwide working on a health problem will be sped up if their collective research output was better organised and extracted into a user-friendly database. This pilot study demonstrated the feasibility to accomplish this task. Treating results like units of information provides freedom to assemble and integrate information from multiple manuscripts in a logical manner. This method of extraction using parent–child relationships and automated linkages facilitates integration of information and makes entry of data easier, especially by less experienced reviewers. We conclude that this system provides a platform to scale up extraction of information on very large topics to be managed by multiple individuals residing in diverse locations.

Contributors All authors contributed equally.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests MB and DK have ownership interest in Plomics. SL, IF, CT-L, SR and JP declare no financial or competing interests.

Patient consent for publication Not required.

Ethics approval This study does not involve human participants.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iD

David Krag http://orcid.org/0000-0001-5355-5999

REFERENCES

1 Müller H-M, Van Auken KM, Li Y, *et al*. Textpresso central: a customizable platform for searching, text mining, viewing, and curating biomedical literature. *BMC Bioinformatics* 2018;19:94.
2 Simon C, Davidsen K, Hansen C, *et al*. BioReader: a text mining tool for performing classification of biomedical literature. *BMC Bioinformatics* 2019;19:57.
3 Subramanyam RV. Art of reading a journal article: Methodically and effectively. *J Oral Maxillofac Pathol* 2013;17:65–70.
4 Linzer M, Mercando A, Hupart KH. Role of a medical Journal Club in residency training. *Academic Medicine* 1986;61:471–3.
5 Eden J, Levit L, Berg A. *Finding what works in health care: standards for systematic reviews*. Washington (DC), 2011.

Supplementary Methodology

From March 2020 to May 2021, 4236 articles retrieved from PubMed responding to the search term COVID-19 were screened and were selected if the article presented new data. Review articles, articles that did not present new data, and articles that were case reports were not selected. Out of 4236 articles screened, 1000 publications reporting new data were selected for entry. Licensed full text articles were imported to the working RefBin account and automatically affiliated were the citation. Results were extracted by the reviewers who read, described, and integrated results from the 1000 selected articles.

Selection of results to be extracted was based on prioritizing primary observations and statistically and clinically relevant unplanned, univariate, and multivariate secondary calculations. Each selected result was treated as an independent unit of information. Four types of note fields were used to describe each result: 1) observation or measurement, 2) description of the measurement, 3) population or sample and 4) the topic. These four note fields were arranged in parent-child order (Figure 1A). The notes were entered manually, and the result note field was entered as the lowest child level. The citation was dragged to this lowest note field. The citation was automatically linked to all the parent note fields.

Similar results from different articles were integrated by sharing parent note fields (Figure 1B). This allowed multiple results to be listed at the same note field level for convenient comparison of multiple results from multiple manuscripts. This substantially reduced the total number of note fields necessary to describe results (Supplemental Figure 2 and Supplemental Table 1). This strategy created an expanding template of organized information. In short, things got easier as more data was entered. This process of integration of results also facilitated standardization of data extraction. The review process ensured that entries were accurate and appropriately integrated with other similar results.

Topics were not predefined but generated in response to sets of results extracted into the data base. The extracted results from 1000 publications grew into 15 topics as first level headings (Supplemental Figure 1).

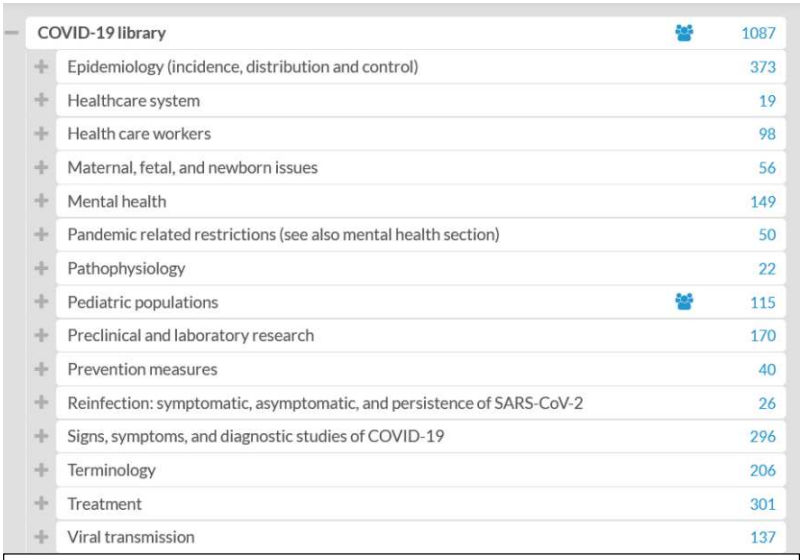
Multiple results from one article (average of 12 results) were usually entered into completely different areas of the database to allow grouping of similar results from different articles. By dragging the article to each extracted result note field, the citation remains conveniently linked to all the extracted data. This was a liberating step that freed the data from the confines of being necessarily described in one list with all the other often disparate results from one article.

Groups of entries were frequently edited to accommodate new data entries. When a note field was moved, all the children note fields moved with it and linkages to citations were automatically updated. This allowed convenient free form development of sets of data.

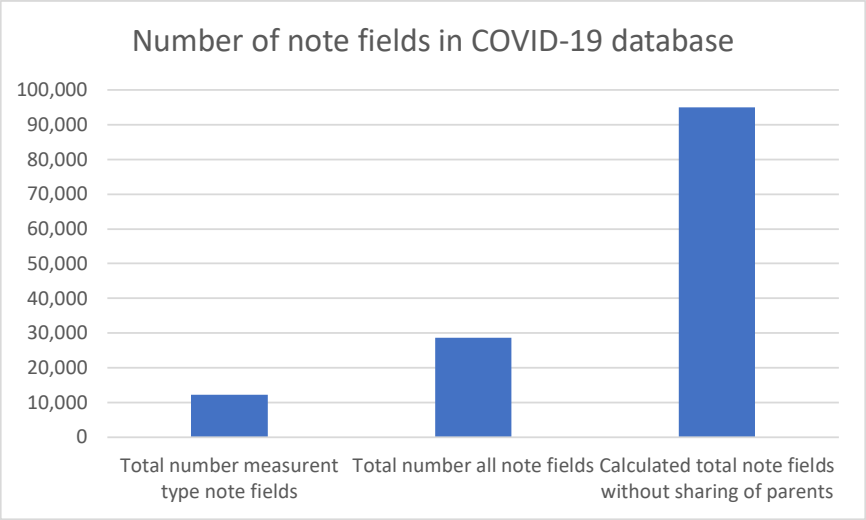
Workflow was established to manage all aspects of the review process and used the same parent child note field process. For example, in workflow, a parent entry included the name of the reviewer. Multiple children note fields described the steps of the review process such as “claimed and under review” and “ready for secondary review”. Each citation was dragged to the appropriate workflow step during the review process. Each article underwent a minimum of two reviews. The reviewers read, described, and integrated the observations presented in each manuscript into the COVID-19 library. The pilot study was considered complete when results from 1000 manuscripts were entered into the COVID-19 library.

The quality of the data was not graded in scalar manner such as 1 to 5. Results were described by entering the specific value followed by confidence intervals or other values that describe significance. For example, incidence rates included the percent followed in parentheses by the affected number of subjects

in the numerator and the total population in the denominator. This approach allowed a user to conveniently see whether 4% was 4/10 or 40/1000.



Supplemental Figure 1: Screenshot of first level headings in the COVID-19 database



Supplemental Figure 2: Number of note fields in the database: measurement type note field, total number note fields and calculated number if sharing of parents was not done.

Average number of note fields per extracted observation	7.78 (SD 1.42)
Mean number of measurements extracted per article	15 (SD 12)
Total number of all note fields in the COVID-19 library	28,613
Total number of the observation type note fields	12,209
Calculated number of note fields if no sharing of parents	94,986
Reduction in number of note fields due to sharing of parents	66,373

Supplemental Table 1: Number of note fields in the COVID-19 database