Using automation to produce a "living map" of the COVID-19 research literature

Ian Shemilt (a)*, Anneliese Arno (a)*, James Thomas (a)*, Theo Lorenc (b), Claire C Khouja (b), Gary Raine (b), Katy Sutcliffe (a), Preethy D'Souza (a), Kath Wright (b), Amanda Sowden (b)
(a) EPPI-Centre, UCL Social Research Institute, University College London, London, UK
(b) Centre for Reviews and Dissemination, Alcuin College, University of York, York

* Joint first authors

Abstract
The COVID-19 pandemic has disrupted life worldwide and presented unique challenges in the health evidence synthesis space. The urgent nature of the pandemic required extreme rapidity for keeping track of research, and this presented a unique opportunity for long-proposed automation systems to be deployed and evaluated. We compared the use of novel automation technologies with conventional manual screening; and Microsoft Academic Graph (MAG) with the MEDLINE and Embase databases locating the emerging research evidence. We found that a new workflow involving machine learning to identify relevant research in MAG achieved a much higher recall with lower manual effort than using conventional approaches.

Key words: *evidence synthesis; literature mapping; COVID-19; automation; machine learning.*

Introduction

The COVID-19 pandemic disrupted life worldwide, and also presented unique challenges in the health evidence synthesis space. As previous papers have observed, COVID-19 evidence has been published at an unprecedented rate: by June 2020, the United States National Institute of Health (NIH) had indexed more than 28,000 articles (1). A thorough, though non-systematic and non-exhaustive, list compiled by the NIHR Policy Research Programme Reviews Facility identified more than 250 COVID-19 maps, auto-searches, and databases as of 19th June 2020 (2). The urgent nature of the pandemic required extreme rapidity for keeping track of research, and this presented a unique opportunity for long-proposed automation systems to be deployed and evaluated.

Observing the range of different semi-automation approaches being adopted across many databases, we initially proposed to conduct an analysis of the strengths and weaknesses of each technology. However, despite appearing similar, many tools had quite different objectives, and so in order to provide a robust evaluation, we decided to conduct a formal cost-effectiveness analysis, where the costs and effects of adopting specific automation tools could be assessed in detail. We selected the COVID-19 living evidence map (3), produced by the Reviews Facility as a case study (illustrated in Figure 1).

About the "living map"

The NIHR Policy Research Programme Reviews Facility[1] is a collaboration between the EPPI Centre at University College London, the Centre for Reviews and Dissemination at the University of York, and the Public Health, Environments and Society at the London School of Hygiene and Tropical Medicine. The facility uses the methods of evidence synthesis to inform policy development and implementation.

In February 2020, a few weeks after the WHO declared a global pandemic, it became clear that there was a need to keep on top of the emerging research evidence. After discussion with DHSC and the office of the Chief Medical Officer, the first evidence map was published in mid-March.

[1] https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=73

Address for correspondence: James Thomas, UCL Institute of Education, University College, London, 20 Bedford Way, London WC1H 0AL, UK. E-mail: james.thomas@ucl.ac.uk.

Searches were run on the MEDLINE and Embase database platforms each week (to begin with) and, out of the 1,049 records found in the first search after duplicates had been removed, 271 met the inclusion criteria. Records were assigned to one of eleven descriptive categories, which captured the key characteristic of the record (for example "treatment development" and "transmission").

The workflow was established as a mostly manual process. Records were downloaded in the form of text files and imported into EndNote. Deduplication took place in EndNote before the records were uploaded into EPPI-Reviewer (4) and the deduplication process run again. Records were then manually screened and assigned to the aforementioned categories with difficult to assign records discussed within the team. The map itself was published using the "EPPI-Mapper" application (5), which is a self-contained HTML5 application, containing the data and the code necessary to produce an interactive visualisation (*Figure 1*).

By the beginning of June 2020, the scale of both the pandemic, and the work involved in maintaining the map, was becoming apparent. After an initial peak in the first search (which was effectively "catching up" on publications up until that point), search yields steadily rose from a few hundred each week to between two and three thousand records per week (*Figure 2*). Following developments in search strategies for COVID-19 literature, the search itself developed over this period too, but it seems likely that most of the increase was simply due to the volume of research being produced.

The map itself had been accessed more than 10,000 times by this point, and the team was receiving frequent requests for copies of the data. This prompted development of the mapping software to enable users to download all, or subsets, of the data in RIS format. This new feature proved popular and accessible; very few requests for data were received after it was deployed in September 2020.

The challenge of addressing the increasing workload of screening the records was addressed in several ways.

First, as the time required for deduplication across tens of thousands of records was increasing every week, eventually taking more than a day of work in EndNote, we adopted a new deduplication algorithm in EPPI-Reviewer (which had been co-incidentally under development and was not implemented simply for this project). This has proved to be both more accurate and efficient than the original de-duplication method.

Second, we evaluated options for the semi-automation of the workflow, and the searching of a single source of

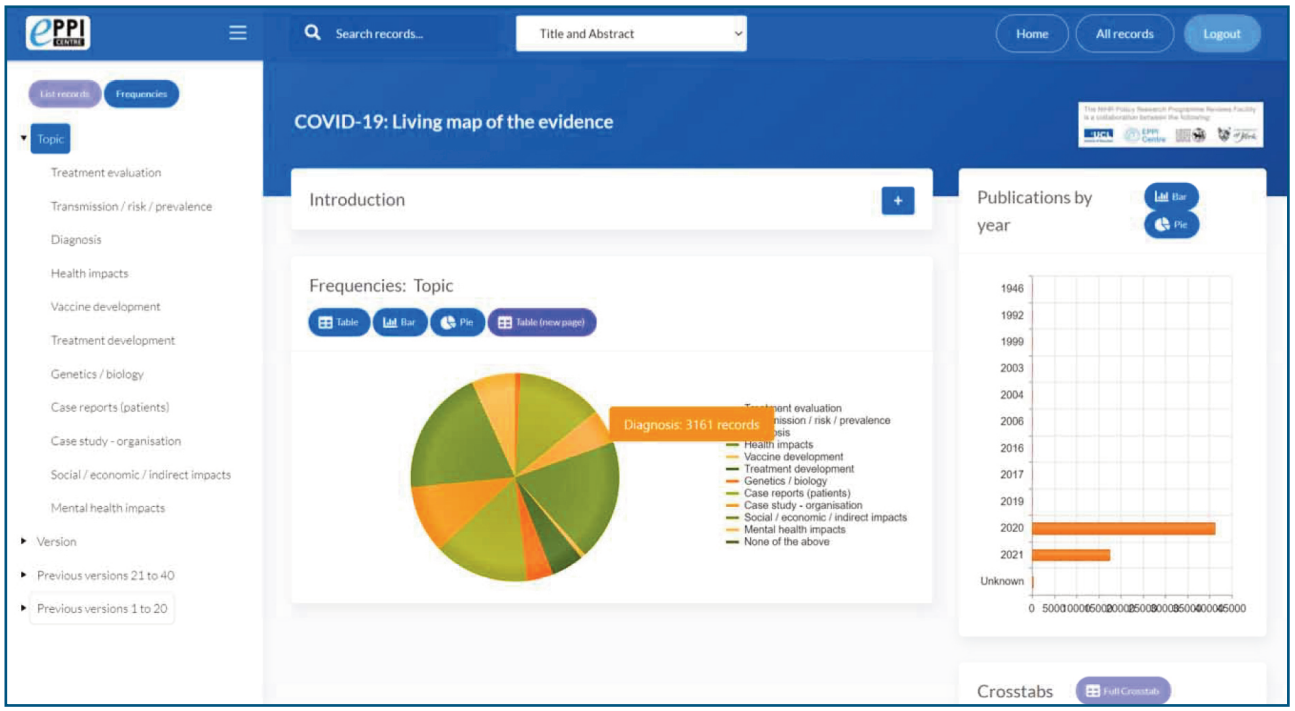

Fig. 1. *Living COVID-19 evidence map.*

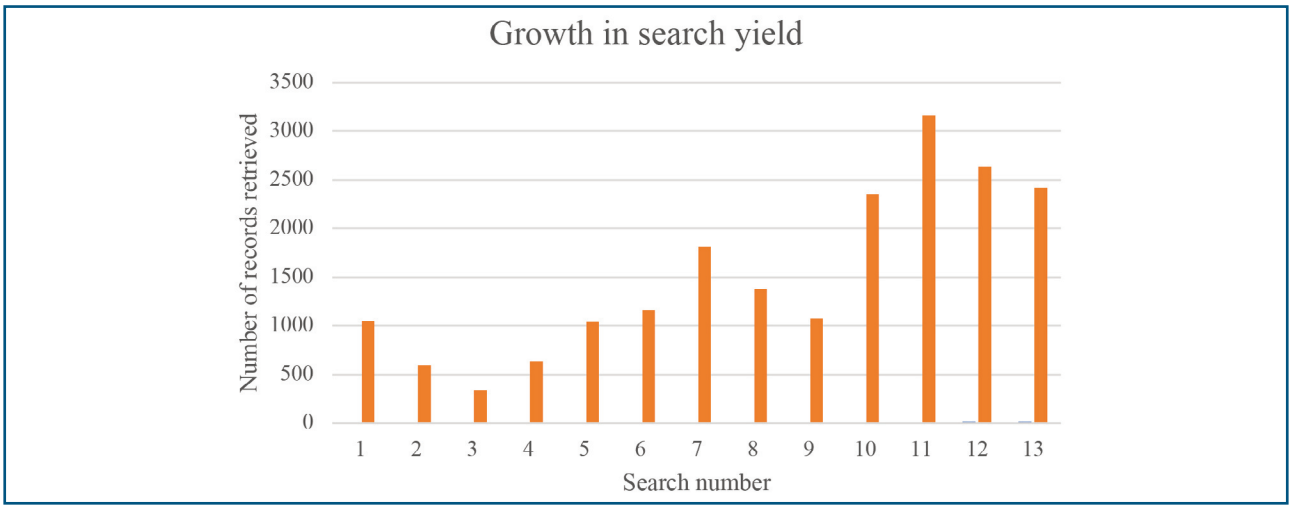

Fig. 2. *The growth in the number of records retrieved in searches 1-13.*

bibliographic records (Microsoft Academic Graph (6)), as opposed to the combination of MEDLINE and Embase.

Methods

Objective

Our objective was to investigate the acceptability, efficiency, and effectiveness of using semi-automated, versus manual, study identification methods to identify eligible study reports for our living map of COVID-19 research; and of using Microsoft Academic Graph as a single source for identification of research.

Acceptability

Adopting the use of semi-automation requires clarity about the process within which it will be introduced. In this case, it was agreed that recall was a key issue: was the team aiming to achieve 100% recall, or was a lower percentage acceptable? If a lower percentage was acceptable, what figure was this? Resource was also an important issue: what was the maximum resource that could be devoted to the task, (this was regardless of whether this was sufficient to assess all records)? These questions informed the adoption decisions made about semi-automation.

Efficiency

Three options for increasing efficiency were evaluated and rolled out in the live workflow:

1. the use of a machine-learning classifier to automatically exclude irrelevant records;
2. the use of the above classifier PLUS prioritised screening with a fixed weekly screening target;
3. the use of Microsoft Academic Graph as a single source of records.

Use of a machine-learning classifier

EPPI-Reviewer contains a feature that uses logistic regression to distinguish between two classes of records (relevant or irrelevant). The classifier requires "training data", i.e. examples of the two classes of records, from which to learn. In this use case, we had thousands of examples of relevant and irrelevant records from which to build the classifier. When built, the classifier can be applied to unseen records, returning a probability score that the record is, or is not, the class of interest. This score can be used to "calibrate" the classifier when used in practice, to determine a pre-specified level of recall. There is usually a trade-off to be made between precision and recall, where higher levels of recall are associated with lower levels of precision. Team deliberations (see "acceptability", above) determined the level of recall that was used in practice.

Use of a machine-learning classifier, plus prioritised screening with a fixed screening target

Prioritised (or "priority") screening uses a machine-learning model to rank the records according to their likely relevance. It uses the same model as described above to score records according to relevance, but the key addition here is that the records are then screened

in order of relevance, and so those records most likely to be included are found at the top of the list. As screeners begin to record their decisions, the priority screening mode observes these decisions and periodically updates the order of the record list such that studies more likely to be included according to previous decisions are now listed towards the top. When using such a workflow, the question for reviewers is whether they should screen the whole list, or whether they should stop after assessing a given proportion, or fixed number. In our use case, a fixed screening target was adopted.

Microsoft Academic Graph as a single source of records

The final change to the workflow was a switch to using Microsoft Academic Graph (MAG) instead of the more conventional sources of MEDLINE and Embase. MAG is an open-access dataset comprising more than 250 million bibliographic records in a network graph map, constructed with the aim of creating a comprehensive single source for citation information. In the "MAG-enabled" workflows, a novel machine-learning recommender model automatically searches each update of the MAG dataset and imports the resulting records into EPPI-Reviewer. The rationale for using this source was to eliminate the need for manual searching of MEDLINE/Embase, and to reduce duplicate checking to a minimum. The team first evaluated the recall of MAG compared with MEDLINE/Embase, by checking whether all the records retrieved by the conventional searches for June 2020, were present in MAG. The "reverse" recall was also checked to see how many papers published during this period (according to MAG) were present in MEDLINE/Embase.

Results

Acceptability

The team discussed the trade-offs involved in maximising recall when using machine learning to increase precision and reduce unnecessary manual work. An issue of concern was performance for each inclusion category – does the classifier or MAG perform especially well for some categories, while not as well for others? There was a similar concern regarding study designs retrieved using semi-automation – might semi-automation perform well for randomised controlled trials (RCTs) for example, but less well for cohort studies? The team decided that a recall of 95% would be acceptable when using the binary machine-learning classifier. The team also decided that the maximum resource available each week was sufficient to screen 1,500 records, so the "fixed screening target" was set at this level.

Efficiency

During the first 19 weeks of operation, the team screened 34,193 records retrieved from MEDLINE/Embase at an average precision of 36%. This fully manual period is used as a baseline.

Use of a machine-learning classifier

The machine-learning classifier, calibrated to achieve 95% recall, was used during weeks 20-29 to automatically eliminate records that were unlikely to be relevant. During this period, 19,891 records were screened from MEDLINE/Embase with an average precision of 61%.

Use of a machine-learning classifier, plus prioritised screening with a fixed screening target

The use of prioritised screening was introduced during weeks 30-34, along with a fixed screening target of 1,500 records per week. During this period 7,685 records were screened from searches of MEDLINE/Embase with an average precision of 79%.

Microsoft Academic Graph as a single source of records

Figure 3 shows the number of unique records found in each source during our evaluation period and the overlap between them. We found that while MAG had a 99% recall overall, MEDLINE/Embase only had a recall of up to 83% due to the large number of additional records found in MAG that were not in our conventional searches.

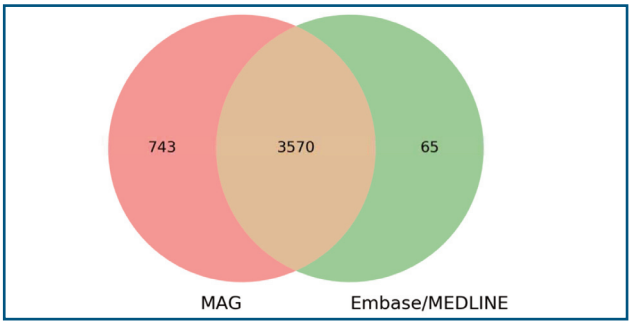

Fig. 3. *Number of records found in each source.*

We, therefore, moved over to use MAG as a single source from weeks 35 onwards, maintaining the use of the machine-learning classifier, prioritised screening, and the fixed screening target. During this period, 32,100 records were screened at an average precision of 69%.

Discussion

This analysis showed that the semi-automated MAG-enabled workflow achieved a higher recall and higher precision than the fully manual workflow and the workflow using the machine-learning classifier alone. It did not achieve the levels of precision obtained using the same automation tools used in the MEDLINE/Embase workflow. However, as it has a higher "baseline" recall (99% compared with 83% for MEDLINE/Embase) and has other efficiencies linked to removing the need to carry out manual searches and deduplicate results, the MAG-enabled workflow was more efficient than the other options. In addition, MAG appears to be more language-inclusive in its study identification, potentially improving our ability to identify non-English-language studies (i.e. we observed, but did not systematically assess, more non-English language records appearing in the workflow when evaluating the possibility of switching to using MAG as a single source of records).

Conclusions

Using MAG in the maintenance of a COVID-19 living evidence map resulted in a higher recall compared with manual searches of MEDLINE and Embase. When combined with other automation tools, namely a binary machine-learning classifier and active learning screening prioritisation, use of MAG had a higher recall and a lower cost, making it more effective and more efficient.

Acknowledgements

This article summarises a presentation given at the 6th meeting of the International Collaboration for the Automation of Systematic Reviews (ICASR) in April 2021; a longer report on the cost-effectiveness analysis will be published shortly.

The substantive and methodological work on the map was commissioned by the National Institute for Health Research (NIHR) Policy Research Programme (PRP) for the Department of Health and Social Care (DHSC). It was funded through the NIHR PRP contract, PR-R6-0113-11003. The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the NIHR or the DHSC. The methodological work, and automated workflow makes use of technologies developed: within the Human Behaviour-Change Project (HBCP), which was funded by the Wellcome Trust; and by UCL and Microsoft.

Submitted on invitation.
Accepted on 13 June 2021.

REFERENCE

1. Hutson M. Artificial-intelligence tools aim to tame the coronavirus literature. Nature. 2020:d41586-020.
2. Resources relating to Covid-19: EPPI-Centre, Social Science Research Unit, UCL Social Research Institute, University College London; 2021. Available from: http://eppi.ioe.ac.uk/cms/Projects/DepartmentofHealthandSocialCare/Publishedreviews/COVID-19Livingsystematicmapoftheevidence/COVID-19Resources/tabid/3767/Default.aspx.
3. Lorenc T, Khouja C, Raine G, Shemilt I, Sutcliffe K, D'Souza P, et al. COVID-19: living map of the evidence London: EPPI-Centre, Social Science Research Unit, UCL Social Research Institute, University College London; 2020. Available from: http://eppi.ioe.ac.uk/COVID19_MAP/covid_map_v50.html.
4. Thomas J, Graziosi S, Brunton J, Ghouze Z, O'-Driscoll P, Bond M. EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis. London: EPPI-Centre, UCL Social Research Institute, University College London; 2020.
5. Digital Solution Foundry, EPPI-Centre. EPPI-Mapper, Version 1.2.5. London: EPPI-Centre, UCL Social Research Institute, University College London; 2020.
6. Sinha A, Shen Z, Song Y, Ma H, Eide D, Hsu B-JP, et al., editors. An overview of Microsoft Academic service (MA) and applications. Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion) ACM; 2015; New York, NY, USA.
DOI: https://doi.org/10.1145/2740908.2742839