BMJ Health & Care Informatics

Pilot trial comparing COVID-19 publication database to conventional online search methods

Camille Torfs-Leibman,[1] Takamaru Ashikaga,[2] David Krag ,[3] Shania Lunna,[3] Sarah Robtoy,[1] Rachel Bombardier[4]

To cite: Torfs-Leibman C, Ashikaga T, Krag D, *et al*. Pilot trial comparing COVID-19 publication database to conventional online search methods. *BMJ Health Care Inform* 2022;**29**:e100616. doi:10.1136/bmjhci-2022-100616

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2022-100616).

Received 27 June 2022
Accepted 15 October 2022

[1]University of Vermont, Burlington, Vermont, USA
[2]Department of Biomedical Statistics, University of Vermont, Burlington, Vermont, USA
[3]Department of Surgery, University of Vermont, Burlington, Vermont, USA
[4]University of Vermont Larner College of Medicine, Burlington, Vermont, USA

Correspondence to
Camille Torfs-Leibman; ctorfsleibman@gmail.com

ABSTRACT

Background and objectives Literature review using search engines results in a list of manuscripts but does not provide the content contained in the manuscripts. Our goal was to evaluate user performance-based criteria of concept retrieval accuracy and efficiency using a new database system that contained information extracted from 1000 COVID-19 articles.

Methods A sample of 17 students from the University of Vermont were randomly assigned to use the COVID-19 publication database or their usual preferred search methods to research eight prompts about COVID-19. The relevance and accuracy of the evidence found for each prompt were graded. A Cox proportional hazards' model with a sandwich estimator and Kaplan-Meier plots were used to analyse these data in a time-to-correct answer context.

Results Our findings indicate that students using the new information management system answered significantly more prompts correctly and, in less time, than students using conventional research methods. Bivariate models for demographic factors indicated that previous research experience conferred an advantage in study performance, though it was found to be independent from the assigned research method.

Conclusions The results from this pilot randomised trial present a potential tool for more quickly and thoroughly navigating the literature on expansive topics such as COVID-19.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ The output of biomedical publication search tools is a set of manuscripts but access to content is restricted.

WHAT THIS STUDY ADDS

⇒ This study provides evaluation of a new method to extract, repurpose and disseminate information derived from published manuscripts.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Despite massive advancements in scientific methods, the science and practice of extraction and integration of information in manuscripts have lagged. We describe a method that advances a new method to improve efficiency of extraction and integration of information.

INTRODUCTION

PubMed contains over 32 000 000 publications and this collection grows by approximately 1 000 000 articles each year.[1 2] The first step in accessing this wealth of knowledge is to acquire a list of publications through search engines such as PubMed and Google Scholar. The user must then painstakingly comb through full-text articles to find the information they seek.

Community curation platforms such as Wikipedia allow rule-based descriptions of virtually any topic. However, Wikipedia is not designed to provide a comprehensive summary of the information contained within a set of publications. Community curation has been used on databases such as UniProt (model organism genome database) but these function more as annotation events than a system to extract information contained within full-text articles.[3] Artificial intelligence (AI) tools for information integration have sought to overcome these obstacles while simultaneously decreasing the time input required from the researcher. Many of the currently existing AI models are limited to generating detailed groupings of articles based on their contents or extracting and comparing information only within narrow categories.[4] These systems have yet to reach the point of providing users with thoroughly researched, discrete answers to their questions.

An online information management system (Refbin.com) was used to manually extract and integrate newly reported data from 1000 COVID-19 articles.[5] As reported in our recent publication,[5] extracted information was described using a minimum of four types of note fields (topic, population, description of the type of measurement and the actual reported measurement). Extracted results

in the same topic were merged so that parent topic note fields were shared. The full text of each article was read and individual observations, such as the incidence rate of COVID-19 infection in college students, were manually entered into the database. That piece of information was grouped with similar observations from other publications. Rather than organising and filing information based on the publication, this new COVID-19 publication database was organised into logical groups of data such as mental health issues related to lockdown or observations on maternal to fetal transmission of COVID-19. A user of this database can readily see specific sets of information and navigate rapidly to increasing levels of detail. In this database, the user does not need to necessarily know what they are looking for and perform a search. The user navigates through topics covering all of the data to quickly find the information they need.

The aim of this pilot trial was to determine whether this new type of database would provide an operational improvement over conventional methods to more rapidly and more accurately find evidence to support statements about COVID-19. A sample of students from the University of Vermont was randomly assigned to use the COVID-19 publication database or their usual preferred search methods to research eight prompts about COVID-19. If assigned to use their preferred research methods, students were allowed to use any means they knew of to find and read primary research articles. We report here the results showing an improved outcome using the new system.

METHODS

Study design

Eighteen participants were recruited from biomedically related courses and student organisations at the University of Vermont to participate in this randomised pilot trial. Prior to participating in the trial, the students completed a participant information survey including questions about their major, level of education, professional goals, research experience, usual speed of task completion and comfort with technology. The final sample included 17 students.

The participants were stratified based on self-rated questions about their previous research experience and usual task completion time requirements. Once stratified, the participants were randomly assigned to use either the new COVID-19 publication database (group A) or their preferred research methods (group B).

The trial was conducted over Microsoft Teams. Prior to taking the test, all participants received a 10 min step-by-step lesson on how to use the COVID-19 publication database prior to being assigned to group A or group B. The participants were informed that only evidence from primary literature would be counted as correct and that only 4 min were allowed per question. The tasks included finding specific pieces of information and were presented in survey form.

The relevance and accuracy of the evidence found for each prompt were individually graded by two members of our research team and then reviewed. Using parameters of specificity established during prompt development, each answer was graded in a binary of correct or incorrect. A correct answer submitted after 4 min (240 s) had passed was marked as incorrect. The time participants needed to answer each prompt was recorded by Qualtrics as the time elapsed between opening the question and clicking the submit button.

This pilot study qualified for an exemption from ethics review by the Institutional Review Board at the University of Vermont and the University of Vermont Medical Center. According to the definition of activities constituting research at these institutions, this pilot trial met the criteria for operational improvement activities and was, therefore, exempt from review.

Query prompt development

The prompts used in this investigation were written using a question stem stating a fact or observation about COVID-19, followed by a request for supporting evidence that the participant had to find using their assigned research method. At the time of this study, the COVID-19 database contained information on 1000 COVID-19 articles. Prompts were written by the study team that confirmed each prompt was answerable using both the COVID-19 database and Google (table 1). A preliminary investigation and refinement of the prompts were conducted with four students. Participants accessed the prompts through a Qualtrics survey, which randomly assigned them 8 of 18 total prompts.

Analysis of prompt and language difficulty was conducted using Microsoft Word's readability tool. The Flesch-Kincaid Grade Level test yielded a score of 13.4, meaning that the language used in these prompts was best suited for those with some college education.

Statistical analysis

Statistical analyses were conducted using SAS Version 9.4 and SYSTAT Version 11 software. Contingency tables were created using the number of correctly answered prompts and the variables: research method (COVID-19 Database or other methods), day of study participation,[1–17] academic background (major 1=biochemistry/biology; 2=dietetics, nutrition and food science (DNFS)), level of education (1=undergraduate; 2=graduate), previous research experience (1=uncomfortable; 2=neutral; 3=comfortable) and gender (0=male; 1=female).

A Cox proportional hazards' model was used to analyse these data in a time-to-correct answer context. The Cox model and survival analyses in general are used widely throughout medical literature but have had limited use in the computer science literature.[6–10] The event of interest in this study is a correctly answered question, enabling us to simultaneously evaluate completion time and accuracy. We examined the effect of several covariates on the time until a participant answered a prompt correctly or until their response was

Table 1 The complete set of prompts administered at random to participants

	Prompt
1	Find two primary research articles demonstrating that Black Americans and predominantly Black communities have suffered from higher incidences of SARS-CoV-2 than White Americans and predominantly White communities. One of these articles should present the data as a rate of infection, while the other should present evidence of this statement as an OR.
2	Find an article in which <3% of healthcare workers without symptoms of COVID-19 tested positive for SARS-CoV-2 through antibody testing. Find another article in which more than 18% of asymptomatic healthcare workers tested positive for SARS-CoV-2 through antibody testing.
3	Find two articles in which the amount of C reactive protein measured in critically severe COVID-19 patients was at least 1.5 times greater than that of severe COVID-19 patients.
4	Find one article showing that essential workers, people receiving treatment for PTSD, and young adults are among groups with the greatest risk of developing suicidal thoughts as a result of the pandemic. Report their risk in terms of ORs.
5	Reports of the percentage of newborns born to COVID-19 positive mothers who demonstrate fever symptoms are variable across the literature. Find one article in which none of the newborns had a fever and find another article in which more than 20% of the newborns had a fever. These articles cannot be case reports of only a single patient.
6	Several articles have published descriptions of new-onset psychotic disorder-like symptoms in patients with COVID-19 (meaning they had no prior history of these symptoms). Find an article that reports on the success of treatment for psychotic-like symptoms in patients with COVID-19. Your answer should include the percentage of patients for whom these symptoms were resolved.
7	Find an article that describes a new-onset of demyelinating lesions in the central nervous systems of patients with severe or critical COVID-19. The article you select must have a study population of at least 20 patients with COVID-19 positive.
8	Peak infectiousness typically occurs around 2 days before symptom onset. Find an article with a sample size of at least 1000 confirmed COVID-19 cases that describe this information and also describes the number of transmission events attributable to pre-symptomatic individuals.
9	Find one article that reports a range of recorded COVID-19 incubation periods with a maximum value of more than 18 days.
10	Find one article that reports the presence of SARS-CoV-2 in the sweat of symptomatic patients with COVID-19. The study should have a sample size of at least 25 individuals. Report the number of patients with COVID-19 with positive sweat samples.
11	Individuals with recent or lifetime substance abuse disorder (SUD) diagnoses are at higher risk of requiring hospitalisation and ventilation treatments after contracting COVID-19. They are also at a greater risk of dying due to the disease. Find an article that supports these three points using data presented as ORs. Note: these data should reflect increased risk as a result of SUD independently of any other comorbid conditions.
12	Find an article that describes a 70% or greater decrease in regional ICU admissions as a result of enacting physical distancing measures.
13	During the COVID-19 pandemic, a greater proportion of obese individuals have reported having a difficult time eating healthfully or are eating less healthy foods compared with the general population. Find two articles that support this statement. One article should report the proportion of obese individuals eating less healthfully and the other should report on the same measure in the general population.
14	Chest CT scans of hospitalised patients with COVID-19 show predominantly bilateral pulmonary lesions as opposed to only unilateral involvement. This trend is even more pronounced in patients admitted to the ICU. Find an article whose findings support this statement. Report the incidence of bilateral and unilateral pulmonary lesions in these two patient groups (hospitalised patients and patients admitted to the ICU).
15	There is evidence that SARS-CoV-2 can cross the blood-brain barrier to invade the brain. Find an article that reports on histochemical analysis of post-mortem brain tissue from patients with COVID-19 in which more than 95% of the samples studied showed evidence of astrogliosis (variably or in all brain regions).
16	The COVID-19 pandemic has had pronounced effects on mental health and substance abuse rates. Find an article in which 28% or more of the general population reported increased alcohol intake. These findings must also show a positive correlation with a history of mental illness.
17	Find two articles providing evidence that transmission of SARS-CoV-2 from mother to baby can occur through the placenta. One of these articles should provide evidence of transmission through immunostaining techniques, while the other should provide evidence using transmission electron microscopy.
18	Find an article showing that study participants who received the Moderna vaccine had a 0% incidence of severe COVID-19 infection post-vaccination.

ICU, intensive care unit; PTSD, post traumatic stress disorder.

censored for being incorrect or over time.[11] A sandwich estimator was used in conjunction with the Cox proportional hazards' model to account for the clustering of responses by each participant.[12] Kaplan-Meier plots were created to visualise the effect of the variables previously mentioned on the rate at which prompts were answered correctly and how many were answered correctly. The Kaplan-Meier estimator was selected due to its ability to handle right-censored data which took the form of wrong or incomplete answers in this pilot study.[13] A Cox proportional hazard model for the search method as well as bivariate models for demographic factors were examined.

RESULTS

Demographics and study groups

A total of 136 responses were collected, divided randomly, and nearly equally among 18 prompts. 72 responses were gathered from group A and 64 of these responses came from participants in group B (this discrepancy is due to incomplete data from the 18th participant) (table 2). One hundred and twenty of the total 136 responses were submitted by female participants reflecting the majority-female classes from which these students were recruited (online supplemental table 1). The study participants were stratified according to major and previous research experience and then randomly assigned to either group. This resulted in an equal number of the responses from group A (24) and group B (24) by students who identified as biology or biochemistry majors (online supplemental table 2). A similarly even distribution among the research methods was attained with students studying DNFS (group B=40, group A=48) (online supplemental table 2). The majority of the responses (104/136) were completed by participants who considered themselves

Table 2 Number of responses by research method group

Prompt number	Group A	Group B	Number of responses
1	4	2	6
2	3	5	8
3	2	5	7
4	5	4	9
5	4	2	6
6	5	3	8
7	7	2	9
8	3	3	6
9	4	4	8
10	4	3	7
11	2	6	8
12	4	4	8
13	4	4	8
14	4	3	7
15	6	2	8
16	4	4	8
17	3	4	7
18	4	4	8
Total	72	64	136

neither uncomfortable nor comfortable in working with biomedical literature (online supplemental table 3). This trial spanned 17 days, from April 26 to May 12, 2021, with a relatively even distribution in the date of study completion among the research method groups. Pearson Chi-square tests showed no association between any of the previously mentioned demographic factors and the research method groups to which the participants were assigned.

No significant associations were found using Pearson Chi-square when considering the day of participation, major and previous research experience in relation to the total number of correct responses across both research method groups. Within group A, biochemistry and biology majors answered 62.5% of their prompts correctly and DNFS students answered 45.8% correctly (p=0.18) (table 3). When looking at these same responses through the lens of previous research experience, it was found that group A students who rated themselves in the highest available category for comfort reading primary literature had a correct response rate of 75%, while the eight students who rated themselves in the middle category had a 48.4% accuracy rate (p=0.15) (table 4). Among

Table 3 Percentage of correct responses from group A participants by major

	Incorrect	Correct	N
Biology or biochemistry	37.5	62.5	24
Dietetics, nutrition and food sciences	54.2	45.8	48
Total	48.6	51.4	
N	35	37	72

Table 4 Percentage of correct responses from group A participants by comfort using and reading research literature (research experience)

	Incorrect	Correct	N
Neutral	51.6	48.4	64
Comfortable	25	75	8
Total	48.6	51.4	
N	35	37	72

the correct and incorrect responses obtained by group B, none of the evaluated demographic factors approached significance.

Kaplan-Meier plots: combined and by research method

Kaplan-Meier plots were used to simultaneously examine the accuracy of each response and the time required to obtain it. Using the survival analysis terminology frequently found in biomedical literature, exact failures or deaths correspond to prompts answered correctly and survival time represents the number of seconds required for the correct response. Right censored responses are those that were either obtained in more than 240 s or were incorrect. The survivor function in the Kaplan-Meier plots below, therefore, demonstrates the proportion of responses that were correct (died) within the time spent on each response.

The Kaplan-Meier plot in figure 1 shows the prompts answered by group A were more frequently answered correctly and more rapidly than group B. A log-rank test yielded a p value <0.001. The mean survival times for prompts at risk of failing (being answered correctly) were 226.8 s and 365.5 s for group A and group B, respectively. The overlap in the survival curves in figure 1B indicates a degree of homogeneity among the group A participants (p=0.259). Figure 1c contains the survival plot of group B. The overlap of curves here also indicates a degree of homogeneity among group B (p=0.466). Due to there only being 12 correct responses out of 64 at risk in group B, both the similarity of the survival curves and the mean survival time could be slightly distorted. This was not an issue in group A where there was a total of 37 correct responses out of 72 prompts answered.

Cox proportional hazards' models with Sandwich Estimator

A Cox proportional hazard model using a Sandwich Estimator was conducted as described by Lin and Wei to account for the clustering of responses by participant.[12] The analysis of the research method resulted in a parameter estimate for group A of 1.388 (SE of 0.254, p<0.001). The HR of obtaining a correct response in the context of the method parameter was 4.01 (95% CI 2.433 to 6.579).

Bivariate Cox models of analyses for the research method and gender, day of study participation or major demonstrated a consistently more significant influence of the research method on response accuracy than the alternate variable. The bivariate model parameter estimate for

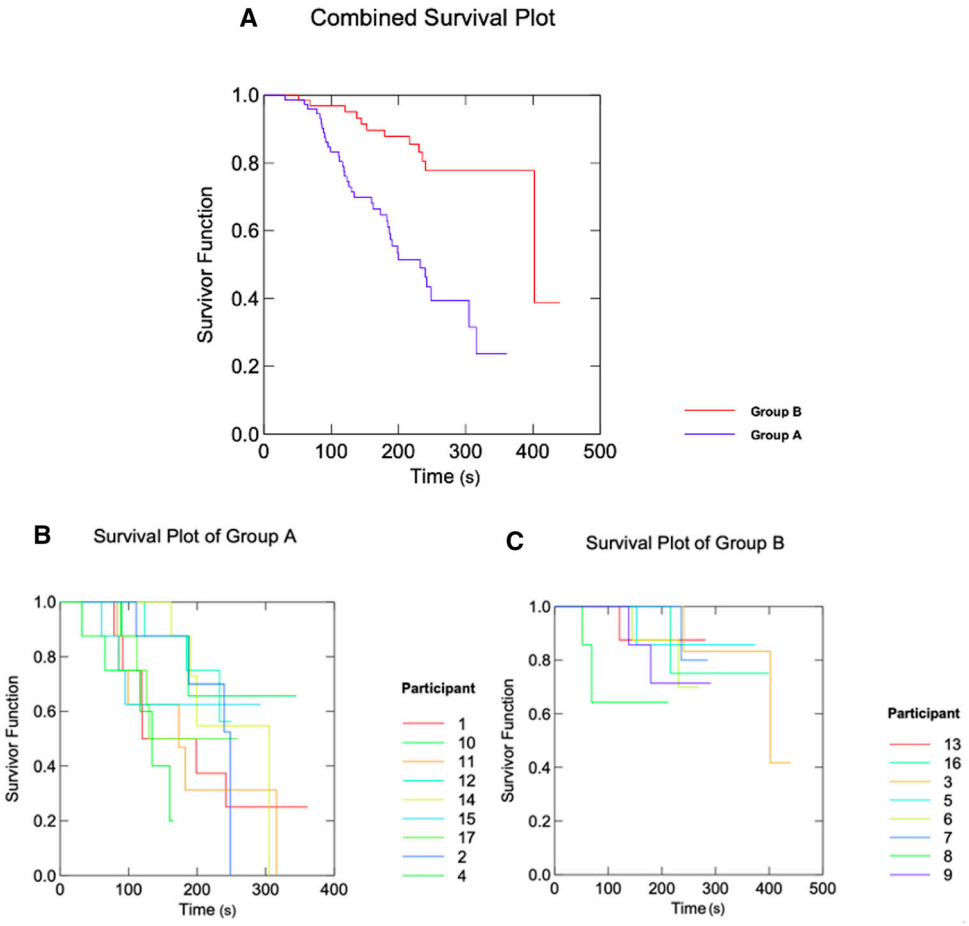

Figure 1 Kaplan-Meier plots showing the survivor function for group A compared with group B (A), as well as for the participants within group A (B) and group B (C) independently.

women was 0.144 (SE of 0.168, p=0.39), for day of study participation −0.008 (SE of 0.014, p=0.579) and for major 0.392 (SE of 0.235, p=0.096). The bivariate hazard ratios were as follows: woman, 1.154 (95% CI 0.830 to 1.606); study day, 0.993 (95% CI 0.67 to 1.019); and major, 0.676 (95% CI 0.426 to 1.072). The level of research experience, in contrast to the previous parameters, resulted in a more significant parameter estimate of 0.542 (SE of 0.118, p<0.0001) and a HR of 1.719 (95% CI 1.363 to 2.167). A correlation matrix was used to show that despite the larger parameter estimate of previous research experience, research method and experience were independent (r=0.022).

DISCUSSION

This randomised pilot trial was performed to determine whether a new information platform containing information from 1000 COVID-19 publications enabled faster and more accurate answers to prompts than conventional methods of accessing biomedical information. Kaplan-Meier plots and Cox Proportional Hazard Models confirmed that the new method of information integration (group A) enabled participants to answer prompts about COVID-19 more quickly and accurately.

Tests for respondent heterogeneity within groups A and B were negative. The paucity of correct answers collected by group B may in part be attributable to the difference between the volume of sources retrieved by a Google search compared with a search in the group A COVID-19 database. After completing the prompts, many of the members in group B commented that 4 min was simply not enough time to find answers. This rush for time coupled with the many blank responses seen from this group could reflect the negative correlation between time per question and user satisfaction described by Xu and Mease.[14] The discrepancy in timed-out prompts and blank responses between the two groups could simply be the result of the pace at which research was completed using the two methods, but it could also provide insight into frustration among participants using traditional research methods. Allowing each participant to be exposed to the full range of questions was part of the design for comparability among students. A limit of 4 min allowed each participant to be exposed to the full range of questions. This was a study design tradeoff between comparable exposure within a specific time frame versus unbounded time to complete the search and having a non-comparable opportunity to answer all specific questions. The choices of the questions were also designed to be not that sophisticated

as to make questions too difficult to be able to abstract from each of the differing approaches. There are differences when conducting real-world research and this pilot evaluation study of the interface and database.

A significant difference was seen between the number of correct answers gathered by group A and group B. This observation contrasts with other examples from the literature.[15 16] For example, when using a question answering system compared with a document retrieval system, Smucker *et al* found that participants answered questions correctly at similar rates.[15] The question answering system described in the Smucker study resembles the new COVID-19 database in that they both aim to present condensed, succinct information independently of its source articles. Therefore, we suspect that the marked improvement in answer accuracy observed in our study may be related to the format in which information is presented. The parent–child organisation structure in the new database system may enable users to more easily adjust the information they are viewing without repeatedly editing their query.

A set of bivariate Cox proportional hazard models showed that among the demographic factors considered in this pilot study previous research experience was the only one to confer an advantage in study performance. A correlation matrix, however, indicated that the prestratification of participants had evenly distributed their degree of prior research experience between the two groups. When examining responses from just group A, major (biology and biochemistry v. dietetics, nutrition and food sciences) and previous research experience did approach significance. This indicates that familiarity with research literature and biomedical language is potentially advantageous in using the database, but ultimately not to a significant extent. In general, studies evaluating information retrieval systems with user-oriented methods have had similar trouble identifying significant extraneous influences on user performance. Whether collecting data on age, sex, computer experience, online search engine familiarity or career objectives these examples from the literature show no significant differences between the groups assigned to the different search methods nor accuracy rates.[16–18]

In this pilot randomised trial, statistically significant differences were observed between groups A and B. This emphasised the marked differences between the methods and warrant expansion of the trial to validate the results. In addition to increasing the sample size, it will be helpful to source queries from a separate set of participants, experts on the topics addressed by the database or search history data from commercial engines.[14 16 19] It will also be useful to expand the diversity of participants' education status. Determining the range of students and users that could benefit from this new platform for information will help guide integration into educational systems. A limitation of the Refbin system is that, unlike Google Scholar and PubMed, methods to continuously update the database are not yet implemented.

CONCLUSION

We demonstrated that a new method of extracting information from published biomedical literature allows users to more quickly and more accurately answer questions related to COVID-19. Topics such as COVID-19 present so much data that a next-generation system is required to more rapidly allow users to answer questions related to the published data. This pilot study with 1000 COVID-19 manuscripts points to a possible solution to speed up research.

Contributors CTL is guarantor.

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 David Krag has a significant financial interest in Plomics Inc, the developer of RefBin.com. The investigator disclosed his personal financial interest to the IRB of the University of Vermont. The IRB at the University of Vermont and the University of Vermont Medical Center determined that this pilot study qualified for an exemption from ethics review. Any potential conflicts of interest were managed. All other authors declare that they have no conflicts of interest.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants but all methods were carried out in accordance with relevant guidelines and regulations as defined by the University of Vermont Institutional Review Boards (IRBs) serving the University of Vermont and the University of Vermont Medical Center. The present pilot study was classified as operational improvement activities by the IRBs. According to the definition of activities constituting research at these institutions, the methods of this pilot did not qualify as research and were exempt from ethics review. The IRBs thereby approved all methods carried out in this pilot study to proceed without further review. Informed consent was obtained from all individual participants included in this pilot study, exempted this study. Participants gave informed consent to participate in the study before taking part.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data sharing not applicable as no datasets generated and/or analysed for this study.

ORCID iD

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

REFERENCES

1 Simon C, Davidsen K, Hansen C, *et al*. BioReader: a text mining tool for performing classification of biomedical literature. *BMC Bioinformatics* 2019;19:57.
2 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.
3 LudÃ¤scher B, Lin K, Bowers S. Managing scientific data: From data integration to scientific workflows. In: *Gsa today (special issue on Geoinformatics*. 109, 2006.
4 Extance A. How AI technology can tame the scientific literature. *Nature* 2018;561:273–4.

5 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.
6 Ortega F, Convertino G, Zancanaro M. Assessing the Performance of Question-and-Answer Communities Using Survival Analysis. ArXiv14075903 Cs [Internet], 2014. Available: http://arxiv.org/abs/1407.5903 [Accessed 08 Jun 2021].
7 Ortega F, Izquierdo-Cortazar D. Survival analysis in open development projects. In: *2009 ICSE workshop on emerging trends in Free/Libre/Open source software research and development*, 2009: 7–12.
8 Zhang D, Prior K, Levene M. How long do Wikipedia editors keep active? In: Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration - WikiSym '12 [Internet]. Linz, Austria: ACM Press, 2012. Available: http://dl.acm.org/citation.cfm?doid=2462932.2462938 [Accessed 08 Jun 2021].
9 Lam SK, Uduwage A, Dong Z. WP:clubhouse? An exploration of Wikipedia's gender imbalance, 2011. Available: https://experts.umn.edu/en/publications/wpclubhouse-an-exploration-of-wikipedias-gender-imbalance [Accessed 08 Jun 2021].
10 Samoladas I, Angelis L, Stamelos I. Survival analysis on the duration of open source projects. *Inf Softw Technol* 2010;52:902–22.
11 Finch H, Lapsley D, Baker-Boudissa M. A survival analysis of student mobility and retention in Indiana charter schools. *Educ Policy Anal Arch* 2009;17:18.
12 Lin DY, Wei LJ. The robust inference for the COX proportional hazards model. *J Am Stat Assoc* 1989;84:1074–8.
13 McNeish D. Applying Kaplan-Meier to item response data. *The Journal of Experimental Education* 2018;86:308–24.
14 Xu Y, Mease D. Evaluating web search using task completion time. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval - SIGIR '09 Internet*. Boston, MA, USA: ACM Press, 2009: 676. http://portal.acm.org/citation.cfm?doid=1571941.1572073
15 Smucker MD, Allan J, Dachev B. Human question answering performance using an interactive information retrieval system 2012;9.
16 Hersh W, Pentecost J, Hickam D. A task-oriented approach to information retrieval evaluation. *J Am Soc Inf Sci* 1996;47:50–6.
17 Al-Maskari A, Sanderson M, Clough P. The Good and the Bad System: Does the Test Collection Predict Users' Effectiveness? 2008:59–66.
18 Egan DE, Remde JR, Gomez LM, *et al*. Formative design evaluation of superbook. *ACM Trans Inf Syst* 1989;7:30–57.
19 Lewandowski D. Evaluating the retrieval effectiveness of web search Engines using a representative query sample. *J Assoc Inf Sci Technol* 2014;66.

Supplemental Table 1 Number of respondents collected in each group by sex

	Group A	Group B	Total
Male	1	1	2
Female	8	7	15
Total	9	8	17

Supplemental Table 2 Number of respondents collected from each group by major

	Group A	Group B	Total
Biology or Biochemistry	3	3	6
Dietetics, Nutrition, and Food Sciences	6	5	11
Total	9	8	17

Supplemental Table 3 Number of respondents collected from each group by comfort using and reading research literature (research experience)

	Group A	Group B	Total
Uncomfortable	0	1	8
Neutral	8	5	13
Comfortable	1	2	3
Total	9	8	17