

## A User-Centered Evaluation of a COVID-19 Intelligent Query System (COVID-IQS)

Danny T.Y. Wu<sup>a,b</sup>, Fangyu Zhou<sup>a,c</sup>, Wu-Chen Su<sup>a</sup>, Hoang Vu<sup>a,d</sup>, Piyush Sahu<sup>a,d</sup>,

Brett Harnett<sup>a</sup>, Tseng-Ping Chiu<sup>f</sup>, Craig Vogel<sup>c</sup>, James J Lee<sup>e</sup>

<sup>a</sup>Department of Biomedical Informatics, <sup>b</sup>Department of Pediatrics, <sup>c</sup>School of Design, <sup>d</sup>Department of Electrical Engineering and Computer Science, <sup>e</sup>Digital Scholarship Center, University of Cincinnati, Cincinnati, USA,

<sup>f</sup>Department of Industrial Design, National Cheng Kung University, Taiwan

### Abstract

*As the fight against COVID-19 continues, it is critical to discover and accumulate knowledge in scientific literature to combat the pandemic. In this work, we shared the experience in developing an intelligent query system on COVID-19 literature. We conducted a user-centered evaluation with 12 researchers in our institution and identified usability issues in four categories: distinct user needs, functionality errors, suboptimal information display, and implementation errors. Furthermore, we shared two lessons for building such a COVID-19 literature search engine. We will deploy the system and continue refining it through multiple phases of evaluation to aid in redesigning the system to accommodate different user roles as well as enhancing repository features to support collaborative information seeking. The successful implementation of the COVID-IQS can support knowledge discovery and hypothesis generation in our institution and can be shared with other institutions to make a broader impact.*

### Keywords:

User-Centered Design; Information Storage and Retrieval; COVID-19

### Introduction

In response to the COVID-19 pandemic, government agencies and leading research institutions in the United States created a COVID-19 Open Research Dataset (CORD-19) and released it on March 16, 2020 [1]. CORD-19 is a fast growing data repository, expanding from 29,000 scientific papers to more than 130,000 papers as of March 2021. This freely available dataset allows the global research community to learn the frontier scientific research and discover new knowledge from an aggregated, central source.

Researchers have primarily been interested in applying advanced informatics approaches, such as natural language processing (NLP) and artificial intelligence (AI) on CORD-19 to propel knowledge discovery and bolster the fight against COVID-19 [2]–[4]. These informatics approaches are needed because of the rapid growth of COVID-19 related publications and datasets, encouraging the global research community to keep up with the situation, and for policy makers to enact the necessary safety measures. In our previous research, we also used NLP and network analysis to identify convergent patterns of seven recent historical viral outbreaks [5].

Several web-based tools have been released to facilitate knowledge discovery in the COVID-19 literature. For example, LitCovid collects more than 121,000 published work on PubMed and annotates the biomedical concepts in them using

PubTator [6], [7]. COVIDScholar uses NLP algorithms to synthesize information across multiple COVID-19 databases [8]. COVID-19 Primer is another website (<https://covid19primer.com/dashboard>) featuring dashboards and rankings to show emerging topics in COVID-19 scientific papers and trends of COVID-19 research in social media and news. However, none of these projects focused on researcher behavior when seeking information or system usability.

Our long-term research objective is to use human-centered AI approaches and create a usable and generalizable information hub for scientific literature. In this study, we aimed to develop an intelligent query system (COVID-IQS) with a well-performed algorithm to retrieve relevant documents from CORD-19 with highly usability. We specifically focused on the user-centered evaluation of the system in this paper. The successful development of COVID-IQS would assist researchers in learning from the COVID-19 literature and generate new hypotheses.

### Methods

#### System Development

##### Retrieval Algorithms

We participated in the COVID-19 special track in the Text Retrieval Conference (TREC-COVID) between April and July of 2020. TREC-COVID contained shared information retrieval tasks organized by the National Institute of Standards and Technology (NIST) of the United States with a goal to develop information retrieval algorithms to fulfill literature search needs of biomedical researchers during the pandemic [9], [10]. TREC-COVID created an infrastructure on top of the CORD-19 dataset, allowing more than 50 teams around the world to participate in this informatics challenge. Our team (CincyMedIR), led by the corresponding author of the present paper (Wu), participated in all five rounds of submissions and made it to the leaderboard in Round 4. Our team used ElasticSearch as the backbone [11] and experimented with various combinations of ranking algorithms and indices (e.g. term or concept-based). The medical concepts of a paper were extracted by MetaMap Lite [12]. Our successful submissions were turned into search modes in COVID-IQS. Specifically, users can perform keyword searches or pick a document in the system as the search term. Users can also choose either term-based or concept-based retrieval to capture and rank relevant documents. These settings result in four different search modes in our system: 1) keywords, term-based search, 2) document, term-based, 3) document, concept-based, and 4) keywords,

concept-based. The detailed algorithm training is outside of the scope of this paper and will be published elsewhere.

### Interface Design

Our design process began with user needs assessment. Rather than conducting interviews or surveys, we summarized the user needs based on the design and functionality of popular biomedical literature repositories (e.g., PubMed) and current COVID-19 literature search tools (e.g., LitCovid). Then, we turned the user needs into system requirements as summarized in Figure 1.

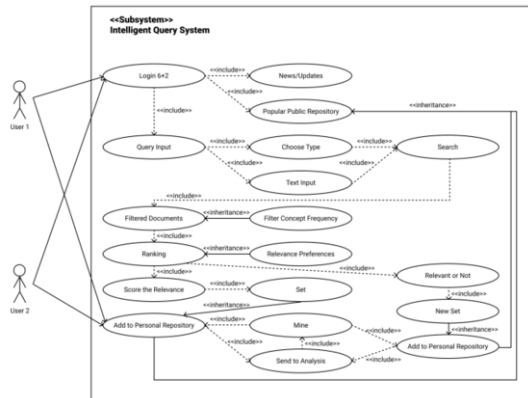


Figure 1 – Summary of user needs and system requirements.

A working prototype was developed based on system requirements and reviewed by an expert team of three master-level designersexperienced in literature review in their thesis work. The team used heuristic evaluation to capture usability issues of this prototype [13]. Many of the usability issues were related to the suboptimal information architecture and display. The prototype was further refined for the evaluation in the present study. Figure 2 shows the interface of the refined prototype with a keyword “risk factor” and its search results.

## User-Centered Evaluation

### Study Design

The study was conducted in a lab setting and virtually due to the pandemic. The study design contained four phases: background survey, usability test, usability rating survey, and semi-structured interview. The evaluation began with the background survey collecting the following demographic information: age group, highest education level, academic title, gender, search engine using habits, and teamwork experiences. Next, in the usability test, the study participants were asked to choose one COVID-19 research topic and generated their own keywords as search terms. Then, the participants performed four search tasks based on the same set of keywords. The first task used keyword searches on PubMed, which was an observation on participants’ natural searching behaviors. The second task used keyword searches with term-based matching in our system. The third task asked the participants to choose the best document on the list and used it as the query (document, term-based search). The fourth task used the same document but switched to the concept-based search mode. The usability test did not include one search mode (keywords, concept-based) because of the longer system response time and the limited time with the participants. In each task, the participants reviewed at least the first 10 relevant documents and indicated the relevance

of each document in the system. The participants filled out the Systems Usability Scale (SUS) consisting of 10 questions in a 5-point Likert scale with 5 being the highest [14]. Lastly, the participants provided open-ended feedback on the perceived usability issues. The study was reviewed and approved by our institutional review board (#2020-0618).

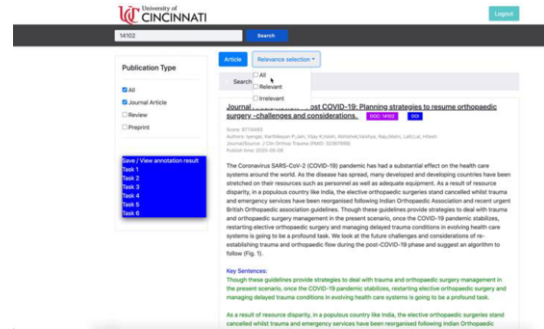


Figure 2 – Search interface of the prototype system being evaluated in the present study.

### Participant Recruitment

A total of 12 researchers were recruited via convenience sampling and snowball sampling at the University of Cincinnati College of Medicine (UCCoM) since the first rollout of COVID-IQS targeted researchers at UCCoM. We first targeted the principal investigators at UCCoM who received an internal COVID-19 research pilot award and invited them to join the study. Then, we asked the participants to recommend other researchers who have contributed to a COVID-19 research and have experience in a team-based research.

### Data Collection and Analysis

The data collection took place from November 2020 to February 2021. All the study sessions were conducted online. In the background survey, a Google Form was used to record the responses. In the usability test and the semi-structured interviews, WebEx screen and audio recording was used to capture the behavioral data and user feedback. In the usability rating survey, an Excel spreadsheet was used to record the SUS scores of each participant. The data were stored at the UC OneDrive and can only accessed by the research team members.

The background survey and the SUS scores were summarized statistically. The composite SUS scores were calculated according to the guideline, ranging from 0 to 100. A SUS score above 68 indicates an above average usability and is considered acceptable. The data from the usability test and the semi-structured interviews were coded qualitatively within and across the three search tasks and analyzed thematically [15]. In addition, the frequency of the codes was calculated as references to prioritize the identified issues. The results were reviewed and discussed within the research team to generate design considerations and recommendations.

## Results

### Participant Demographics

Table 1 shows the demographics of the 12 participants. The majority of participants were within the middle-age group with either a MD or a PhD. Most of our participants were male (N=10) and in their mid career (N=9, associate professor or above).

Table 1 – Participant demographics

ID	Rank	Gender	Age Grp.	Degree
1	Full Prof.	Male	46-55	PhD
2	Assoc. Prof.	Male	36-45	MD
3	Asst. Prof.	Female	26-35	PhD
4	Full Prof.	Male	46-55	PhD
5	Asst. Prof.	Male	46-55	PhD
6	Full Prof.	Male	56-65	MD
7	Assoc. Prof.	Male	36-45	MD
8	Assoc. Prof.	Male	36-45	MD
9	Rsch. Asst.	Male	26-35	Master
10	Librarian	Female	26-35	Master
11	Full Prof.	Male	56-65	MD
12	Assoc. Prof.	Male	36-45	PhD

The background survey revealed the participants' literature search behaviors and preferences. PubMed was the most popular literature search engine (91.7%, N=11). Journal (100%, N=12), Review (83.3%, N=10), and Pre-print (75%, N=9) were the top three types of articles that the participants focused on in their literature search. Moreover, the participants had different preferences to save the relevant documents as follows: on their local personal computer (90%, N=9), in a cloud drive (50%, N=5), in a folder of the search engine (30%, N=3), and printed out as paper copies (30%, N=3). These literature search behaviors and preferences provided context for interpreting the usability issues and recommended references for final design decisions.

Table 2 – SUS Scores by Group

Group (N)	Mean	Acceptability	SD
All participants (12)	69.38	Marginally	21.46
Group A (3)	36.67	Not acceptable	10.41
Group B (9)	80.28	Acceptable	8.43

### Usability Rating

The average composite SUS score of the usability test was 69.38 (standard deviation or SD: 21.46), which is slightly above the average score (68) and considered marginally acceptable. However, the examination of the SUS scores shows a bipolar distribution. As shown in Table 2, three data points (Group A) are significantly below average with a SUS score of 36.67 (SD: 10.41), while the other nine data points (Group B) are higher than the average at 80.28 (SD: 8.43). In other words, the usability of COVID-19 was acceptable for three quarters of the participants but not acceptable for a quarter of the participants, leading to an overall marginally acceptable usability (SUS=69.38).

### Usability Issues

Usability issues were grouped based on the thematic analysis into four categories, with regards to user mental model, system functions, information display, and implementations.

#### Category 1: Distinct User Needs

Researchers have different roles in literature search with their own strategies, aims, and processes and therefore have distinct needs. The prototype system did not fully consider the needs for specific user groups and was focused more on one type of design for general researchers versus others, leading to seemingly conflicting usability issues in this category. For example, the analysis shows that some participants aimed to identify mainstream articles, others aimed to seek for innovative ideas.

These two aims can be distinguished by searching with or without pre-prints, and the system should allow both. This category of usability issues helped us redesign the system to accommodate various user needs and ultimately provide customized information based on known user roles and behaviors.

#### Category 2: Functionality Errors

This category of usability issues resulted from the mismatched mental models between the designers and the users [16]. Here the mental model refers to the perceived use of the system. When the perceived use between the designers and the users did not match, the system provided wrong functionality to the users. These usability issues were covered heuristics including the visibility of system status, the matching between system and the real world, and the user control and freedom. An example in this category was related to the repository where users can save the selected documents. The user would like to further group the selected documents in their own ways and annotate them. However, the system did not provide such functionality to support these grouping and annotations. These usability issues helped the research team to bridge the mental models and refine the system functionality accordingly.

#### Category 3: Suboptimal Information Display

This category of usability issues refers to a less efficient and effective way of information representation and display. Some usability issues were covered by heuristics such as recognition rather than recall, aesthetic and minimalist design, and visibility of system status. Others can be resolved by refining the information architecture and hierarchy. These usability issues especially helped the research team refine the presentation of abstracts and their key sentences. While the original abstracts may be too long, the key sentences automatically generated by the system were lack of logic. This inspired the team to come up with a solution to highlight the key sentences within the abstract, giving the key sentences a context and reducing the cognitive load of reading the abstract.

#### Category 4: Implementation Errors

Usability issues in this category were due to implementation errors, which were not intended and had nothing to do with the design and can be resolved technically. Capturing this type of usability issues before the system rollout can significantly increase user acceptance toward the system. One example was that the PubMed identifier (PMID) being searched in the searching box was misplaced by a document identifier after clicking the search button. A thorough testing plan should be executed to catch these minor issues.

### User Roles in Literature Search

As shown in Table 2, there were two groups of participants (Groups A and B) who gave very different usability scores to the system, which may be due to their different roles in literature search. It is not uncommon to see asymmetric roles (e.g., managers versus knowledge workers) in collaborative information seeking [16]. Reviewing the inductive color coding of our qualitative data confirmed that there were three different roles in our participants: Administrator, Expert, and Principal Investigator. Figure 3 shows the persona of each role.

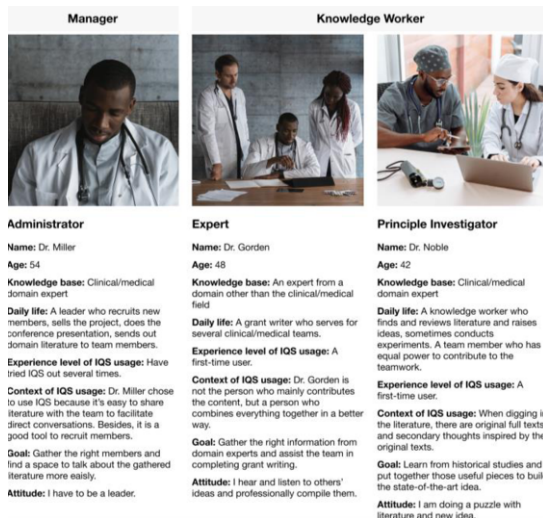


Figure 3 – Personas based on the semi-structured interviews.

These three roles overlapped with each other, but each had distinct goals and processes in literature search. The principal investigators and experts were kinds of knowledge workers, while the administrators were focused on research project management. The qualitative analysis also found that the research team structured can affect the boundary between managers and knowledge workers. When there was a hierarchy in a research team, there was a more explicit boundary between these two roles. On the other hand, situations where the boundary was nebulous and everyone on the team worked as a knowledge worker occurred in more equal collaborations.

In addition, the qualitative analysis shows that Group A and B likely map to different roles. The participants in Group B (N=9) were more like knowledge workers who worked mostly in equal collaborations. The participants in Group A (N=3) were more like managers working in hierarchical collaborations. These participants needed to find influential articles in a short time to get a sense of the topic but leave the detailed search to others in the team. They also used literature search to identify potential collaborators, more than knowledge discovery and hypothesis generation. This finding suggests that our system needs to be redesigned to support collaborative information seeking and pay attention to the needs of both managers and knowledge workers in specific context.

## Discussion

### Key Findings

We developed a COVID-19 literature search engine with well-performed ranking algorithm and conducted a user-centered evaluation with 12 biomedical researchers in our institution. The usability of the system is above average, although the distribution was bipolar. The qualitative analysis revealed four categories of usability issues, helping the research team to rethink and redesign the system. The analysis also showed distinct user roles and needs that can affect the perceived usefulness of the system. Based on the key findings, we shared two lessons learned for researchers to develop their own COVID-19 literature search engines.

### Lesson 1: Conduct Phased User-Centered Evaluation

Our study indicated that literature search engines supports the suggestions of user-centered evaluation to iteratively improve its usability both in a lab and in a field and further increase long-term user adoption. [17] Even with well-trained designers to refine the prototype system using heuristics in the first place, the gaps of mental model between designers and users were still significant, leading to multiple usability issues. While most of COVID-19 search engines focus on improving the relevance of retrieved documents, more attentions should be given to system usability to better support the process of knowledge discovery and collaboration.

In addition, the gaps of mental model between designers and users could lead to different user interface display representations. Users would consider more on how to get heuristic and intuitive information from a search engine interface, while designers would focus on aesthetics or simplicity of the interface to increase the visibility. Since these two preferences may overlap with but also deviate from each other, and even evolve over time, a phased user-center evaluation can help identify the gaps and bridge them.

### Lesson 2: Consider User Roles and Collaborations

Our usability rating showed mixed results in SUS scores, which likely resulted from the distinct user roles in the team collaborations. Although we did not have a large group of participants, our analysis was able to uncover three potential user roles and explain how they may change the perceived usefulness of the system. Since team-based science becomes essential, biomedical researchers may have different roles in the literature search and collaboration. The system design should refer to theories and models in the field of collaborative information seeking and accommodate different user groups as well as support their information needs.

Additionally, it is worth investigating the literature search behaviors of researchers in various cultural groups, especially those which have been conquering COVID-19, to make our system more impactful. Taiwan has been widely applauded for its successful management of the pandemic using smart contact tracing, automated alert messaging, and health insurance data [18]. Several COVID-19 dashboards have been developed to deliver public health messages, and the general public follows the rules to keep COVID-19 from spreading. Such societies with successful stories may lead to unique COVID-19 research questions and literature search strategies. It would be worth investigating societal and cultural differences and redesigning our system to support global collaborations and partnerships in COVID-19 research.

### Limitations

This study has several limitations. First, it has a relatively small and unrepresentative sample of users. However, this number is sufficient to identify most of the usability issues of our system since normally five participants could identify 85% of issues [17]. Second, the participant demographics were unbalanced, with most being male and in their mid-career. After the system rollout, we will collect user profiles and click behaviors with a larger user group to improve our understanding of the system usability. Finally, we only tested three of the four system search modes due to limited time with the participants. While we combined multiple methods to collect user feedback, the time constraint prevented us from collecting more nuance user behaviors and feedback.

## Conclusions

We developed COVID-IQS and summarized its usability issues in four categories and shared two lessons learned. We have refined the system based on the findings of the present study and rolled out the system for the researchers in our institution to use. We will also conduct user-centered evaluation in phases and redesign the system to incorporate different user groups and needs. Meanwhile, we will enhance the repository features of COVID-IQS to support collaborative information seeking. COVID-IQS will be integrated into our clinical research informatics infrastructure to support the COVID-19 research planning in our institution. COVID-IQS can be shared with other institutions globally and be applied to non-COVID-19 literatures to make a broader impact.

## Acknowledgements

This research was funded by the COVID-19 Research Pilot Award co-sponsored by the College of Medicine and the Office of Research at the University of Cincinnati (PI: Wu, April 2020 to April 2021). We thank Ms. Chen Xin and Ms. Tianran (Vicky) Chen for their effort on interface design and heuristic evaluation. We also thank Ms. Anunita Nattam and Mr. Ruthvik Abbu for manuscript proofreading.

## References

- [1] L. Lu Wang *et al.*, "CORD-19: The Covid-19 Open Research Dataset," *ArXiv*, Apr. 2020.
- [2] L. J. Kricka *et al.*, "Artificial Intelligence-Powered Search Tools and Resources in the Fight Against COVID-19," *EJIFCC*, vol. 31, no. 2, pp. 106–116, Jun. 2020.
- [3] A. Abd-Alrazaq *et al.*, "A Comprehensive Overview of the COVID-19 Literature: Machine Learning-Based Bibliometric Analysis," *J. Med. Internet Res.*, vol. 23, no. 3, p. e23703, Mar. 2021, doi: 10.2196/23703.
- [4] G. Cernile *et al.*, "Network graph representation of COVID-19 scientific publications to aid knowledge discovery," *BMJ Health Care Inform.*, vol. 28, no. 1, Jan. 2021, doi: 10.1136/bmjhci-2020-100254.
- [5] M. V. Powers-Fletcher *et al.*, "Convergence in Viral Outbreak Research: Using Natural Language Processing to Define Network Bridges in the Bench-Bed-side-Population Paradigm," *Harv. Data Sci. Rev.*, Jan. 2021, doi: 10.1162/99608f92.cc479d52.
- [6] Q. Chen, A. Allot, and Z. Lu, "Keep up with the latest coronavirus research," *Nature*, vol. 579, no. 7798, pp. 193–193, Mar. 2020, doi: 10.1038/d41586-020-00694-1.
- [7] C.-H. Wei, H.-Y. Kao, and Z. Lu, "PubTator: a web-based text mining tool for assisting biocuration," *Nucleic Acids Res.*, vol. 41, no. Web Server issue, pp. W518–522, Jul. 2013, doi: 10.1093/nar/gkt441.
- [8] A. Trewartha *et al.*, "COVIDScholar: An automated COVID-19 research aggregation and analysis platform," *ArXiv201203891 Cs*, Dec. 2020, Accessed: Apr. 23, 2021. [Online]. Available: <http://arxiv.org/abs/2012.03891>.
- [9] K. Roberts *et al.*, "TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19," *J. Am. Med. Inform. Assoc.*, vol. 27, no. 9, pp. 1431–1436, Sep. 2020, doi: 10.1093/jamia/ocaa091.
- [10] E. Voorhees *et al.*, "TREC-COVID: constructing a pandemic information retrieval test collection," *ACM SIGIR Forum*, vol. 54, no. 1, pp. 1–12, Jun. 2020, doi: 10.1145/3451964.3451965.
- [11] C. Gormley and Z. Tong, *Elasticsearch: the definitive guide*, First edition. Beijing ; Sebastopol, CA: O'Reilly, 2015.
- [12] D. Demner-Fushman, W. J. Rogers, and A. R. Aronson, "MetaMap Lite: an evaluation of a new Java implementation of MetaMap," *J. Am. Med. Inform. Assoc. JAMIA*, vol. 24, no. 4, pp. 841–844, Jul. 2017, doi: 10.1093/jamia/ocw177.
- [13] J. Nielsen and R. Molich, "Heuristic evaluation of user interfaces," in *Proceedings of the SIGCHI conference on Human factors in computing systems Empowering people - CHI '90*, Seattle, Washington, United States, 1990, pp. 249–256, doi: 10.1145/97243.97281.
- [14] J. Brooke, "System usability scale (SUS): a quick-and-dirty method of system evaluation user information," *Read. UK Digit. Equip. Co Ltd*, vol. 43, 1986.
- [15] M. Maguire and B. Delahunt, "Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars," *AISHE-J. J. Teach. Learn. High. Educ.*, vol. 9, no. 3, 2017, [Online]. Available: <http://ojs.aishe.org/index.php/aishe-j/article/view/335>.
- [16] D. A. Norman, *The design of everyday things*, Revised and Expanded edition. New York, New York: Basic Books, 2013.
- [17] C. Shah, "Collaborative information seeking: Collaborative Information Seeking," *J. Assoc. Inf. Sci. Technol.*, vol. 65, no. 2, pp. 215–236, Feb. 2014, doi: 10.1002/asi.22977.
- [18] C.-M. Chen *et al.*, "Containing COVID-19 Among 627,386 Persons in Contact With the Diamond Princess Cruise Ship Passengers Who Disembarked in Taiwan: Big Data Analytics," *J. Med. Internet Res.*, vol. 22, no. 5, p. e19540, May 2020, doi: 10.2196/19540.

## Address for correspondence

Danny T.Y. Wu, PhD, MSI, FAMIA  
 Assistant Professor, Biomedical Informatics & Pediatrics  
 University of Cincinnati College of Medicine  
 231 Albert Sabin Way, ML0840, Cincinnati, OH 45229 USA. Email: [wutz@ucmail.uc.edu](mailto:wutz@ucmail.uc.edu)