Measurement of Utility in User Access of COVID-19 Literature via AI-powered Chatbot

Roland Oruche[1], Eric D. Milman[2], Xiyao Cheng[1], Megha Joish[1], Chaitra Kulkarni[2],
Agrim Sharma[1], Kerk Kee[2], Hariharan Regunath[1], Prasad Calyam[1]
University of Missouri[1],Texas Tech University[2]
Email: {rro2q2, xcwx3}@umsystem.edu; {sagrim23, mcjoish}@gmail.com; {regunathh}@health.missouri.edu;
{eric.d.milman, chaitra.kulkarni, kerk.kee}@ttu.edu; {calyamp}@missouri.edu

***Abstract*—The advent of COVID-19 has resulted in a data deluge in the medical literature archives and there is a need for publication analytics services to augment the clinical workflow of medical users (e.g., clinicians, researchers, medical students) for literature search. Publication analytics services such as KnowCOVID-19 science gateway with a chatbot interface viz., Vidura Advisor have been developed to mine relevant literature from archives such as e.g., CORD-19 open-source dataset. In this paper, we present a novel utility measurement framework that uses a statistical technique for z-score calculation to measure the utility in user access of COVID-19 literature with publication analytics services. Our framework approach builds on a usability study of the KnowCOVID-19 science gateway to identify challenges of user adoption of publication analytics at the individual level through the assessment of user performance and perception of factors such as user interface design, functionality, and derived information insights. In addition, we detail the software features (e.g., domain-specific topic filtering, data reports in terms of drugs/genes) within KnowCOVID-19 that support our measurement framework assessments. We evaluate our proposed framework through experiments on user performance and perception by comparing KnowCOVID-19 assisted by Vidura Advisor with a standard search engine i.e., Google Scholar over a set of clinical literature search tasks. The results from our usability study for assessing user performance using a single usability metric (z-score) show a 47% higher score in application utility with KnowCOVID-19 compared to Google Scholar.**

***Index Terms*—Publication Analytics, Statistical Utility Measurement, Diffusion of Innovations, Statistical Inferencing**

I. Introduction

The effect of COVID-19 has led to a massive data deluge in medical literature related to the pandemic. Medical users (e.g., clinicians, researchers, medical students), more than ever, are looking to extract information and insights (e.g., trending medical topics) from scientific literature in order to streamline the process of knowledge discovery within their respective medical disciplines.

In our previous work [1], we resolved the recurring problem of relying on expert knowledge to manually identify obscure topics among medical corpora by developing a science gateway application viz., KnowCOVID-19 for publication analytics. KnowCOVID-19 provides access to datasets, and tools specific to medical-domain users by utilizing a generative modeling and latent parameter estimation technique viz., Domain-Specific Topic Model (DSTM) [2], [3] to filter literature based on hierarchical evidence as suggested in the Level of Evidence Pyramid [4], [5]. In addition, we developed a context-aware conversational agent viz., Vidura Advisor [6], to mine user intents (i.e., queries) through a generative modeling algorithm in order to recommend relevant literature. Our science gateway application enables users to augment their manual workflow of literature search to filter high-quality publications at high-scale filtered on the premise of scientific rigor, as shown in Figure 1.

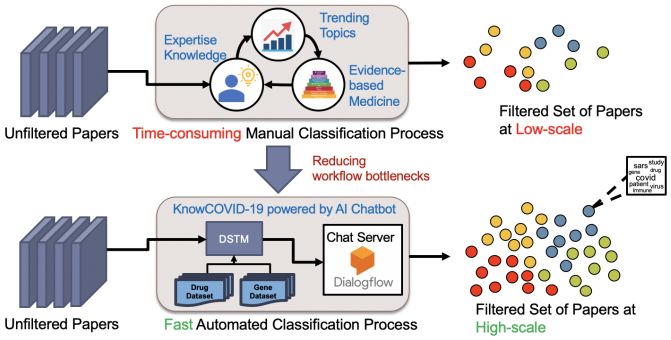


Fig. 1: Improvement of the manual literature search process using the KnowCOVID-19 workflow process for automatic filtering of COVID-19 publications at high-scale.

While we have previously demonstrated the performance of KnowCOVID-19 to filter high-quality medical literature using our topic modeling technique at high-scale [1], we found that there is a lack of pertinent measurement frameworks to validate the publication analytics processes for enabling clinical and bedside care practices. Based on the Diffusion of Innovations (DOI) theory [7], highly adopted applications that provide feasible user interfaces and tools for extracting high-quality information from high-dimensional datasets require widespread implementation and diffusion. These applications with advanced tools such as AI-powered chatbots for user guidance must meet the diverse needs of users for successful user adoption. Thus, a measurement of application usefulness must be performed to address the challenges faced by users in adopting a COVID-19 related publication analytics service.

In this paper, we develop a novel utility measurement framework that uses a statistical technique for z-score calculation to measure the utility of the KnowCOVID-19 science gateway

978-1-6654-2471-4/21/$31.00

workflow processes of publication analytics. We gather medical user input from a *perception assessment study* to design a framework that addresses the socio-technical challenges of DOI theory for user adoption. We define *utility* as a function of: (i) *user interface*, which are the features that make the website usable and intuitive, (ii) *functionality* of software components to perform domain-specific user tasks, and (iii) *insights* that measure the relevance of the information presented to users. The measurement framework entails metrics such as completion time, success rate, and difficulty score as a component of quantitative usability [8], [9], as well as newly defined metrics such as navigability, intuitiveness, document relevance, and document significance. The utility measurement framework leverages application engagement from medical users as input to calculate the quantitative scores that are in turn used for generating a single usability score (z-score) over application components (e.g., user interface, insights, functionality) as well the application as a whole. To support the measurement framework, we detail software components within the KnowCOVID-19 science gateway for publication analytics that include our domain-specific topic filtering and clinical data reports for evaluating user performance.

We evaluate our utility measurement framework in user access through a collection of medical user data over 6922 articles from the COVID-19 Open Research Dataset (CORD-19) [10]. We present a usability study that compares the user performance between Google Scholar and KnowCOVID-19 applications as *Condition 1*, and KnowCOVID-19 and KnowCOVID-19 assisted by Vidura as *Condition 2*. Eight medical users ranging from clinicians to medical students are evenly split over the two conditions to perform a set of clinical literature search tasks related to the evidence-based practice and COVID-19 area knowledge. The utility measurement framework uses the interactions of medical users with the applications as input to quantify their performance and assess the utility of the application through a z-score calculation. An example task may include determining whether the drug Lopinavir, which is used to treat SARS, can treat COVID-19 in patients. In addition, we capture the medical user perception during their application tasks, and also while interacting with the Vidura Advisor through a set of platform utility measures in the form of questionnaires [11]–[13]. Inferential statistics techniques such as MANOVA and Chi-square are used to demonstrate statistical significance in user performance and perception over the two conditions.

The rest of the paper is organized as follows: Section II discusses the related work. In Section III, we detail the results of our utility measurement framework requirements based on a user perception assessment study. In Section IV, we describe the statistical techniques used in our utility measurement framework to calculate z-scores, and also describe system components of KnowCOVID-19 assisted with Vidura Advisor. In Section V, we perform a usability experiment of literature search applications using our proposed utility measurement framework. Finally, Section VI concludes the paper.

II. Related Work

A. Utility Measurement for Technology Adoption

Application utility (or usefulness) has been long studied for the purposes of disbursement and adoption by users [7], [14]–[16]. The Technology Acceptance Model (TAM) proposed by Davis *et. al.* describes the factors behind the adoption of new technology based on valid measurements of utility. TAM details the concepts of: (i) *perceived usefulness*, the degree to which an application can enhance user work performance, and (ii) *perceived ease of use*, which studies the balance between the application learning curve and user performance. Similarly, Rogers [7] theorized Diffusion of Innovations (DOI) as the process by which new technologies, behaviors, or ideas are adopted in a given community, organization, or society. In its framework of innovation attributes, DOI presents the idea of *relative advantage* as the degree to which innovative technology is seen as better than the idea, program, or product it replaces.

The work in [17] extends TAM by developing a three-dimensional TAM on an innovative e-learning system for higher education. This three-dimensional framework demonstrates the extension of utility measurement through all aspects of system use, users, and system components, where perceived usefulness and ease of use are identified at each level. Similarly, authors in [18] utilize structural equation modeling [19] to statistically analyze TAM influenced metrics such as usefulness, ease of use, attitude, and intention as a measurement of acceptance for online banking. Min *et. al.* [20] studied the influence of DOI metrics such as relative advantage, compatibility, complexity, observability, and social influence on perceived usefulness and perceived ease of use for mobile ride-sharing applications. Authors in [21] adopted DOI to study strategies for implementation on mobile device integration in nursing curricula.

We build upon such previous works that demonstrate the practical use of the aforementioned theories of utility measurement, and show empirical evidence of the relative advantage that innovative technologies (i.e., science gateways) have over conventional applications. Uniquely, our work employs DOI theory to address individual-level adoption challenges (e.g., relative advantage, simplicity) by developing a utility measurement framework motivated by the feedback from users in a perception assessment study. We integrate a utility measurement framework over our KnowCOVID-19 science gateway assisted by Vidura Advisor and compare the application utility and relative advantage with existing literature search applications, such as Google Scholar.

B. Quantitative Approaches for Application Performance

Previous works have incorporated quantitative and statistical measurements for improving both application performance that affect human interaction [22]–[24]. The work in [25] addresses application usefulness challenges due to inadequate methods for utility measurement by conducting a statistical study that evaluates the usability and accessibility levels based on human

perception. This study was able to show the consistency of usefulness between a questionnaire-based evaluation as well as a performance-based evaluation using quantitative metrics such as task success rates, task completion time. Authors in [26] address the challenges of website navigation by proposing a mathematical programming model that is viewed as a specialized graph optimization problem. The study shows that using metrics such as the average number of paths per mini session greatly facilitates user navigation performance. In [27], authors develop a Social Media Panel (SMP) that evaluates effectiveness, efficiency, and user engagement against traditional website navigation tools. While quantitative measurements showed the application's efficiency with fewer clicks to complete a task, there were no significant results with statistical significance in demonstrating an acceleration of task completion on SMP. The work in [28] shows that by evaluating patients' tasks performance (e.g., success rate, time efficiency) and satisfaction (e.g., system usability scale) on mobile health systems, it is possible to alleviate the deficiencies of patients' perspective and interaction performance.

Quantitative frameworks such as [29] have developed an evaluation tool called, useR Interface evaluaTion frAmework (RITA), through the expansion of existing user interface evaluation tools on a software application with a modular architecture. The RITA framework utilizes three evaluation techniques (i.e., questionnaires, ergonomic quality inspection, electronic informer) for analyzing user performance and increasing user intent in interactive systems. A study in [30] leverages the BaLOReS evaluation framework [31] by formalizing a framework that quantifies aesthetic metrics of balance, linearity, sequentially, orthogonality, and regularity, where the aesthetic metrics are ranged between 0 and 1 on the absence of a measured principle (0), total fulfillment (1), or user acceptability (0.5). The work in [32] proposed an intelligent usability evaluation tool that automates the usability evaluation process through AI-based heuristic evaluations for enhancing the Internet users' experience and satisfaction on the Web.

Motivated by these previous studies, we develop a utility measurement framework that utilizes various quantitative metrics to calculate a z-score via a statistical technique for assessing user performance on the KnowCOVID-19 science gateway application. The calculation of our z-score combines each metric within their respective application component (e.g., user interface, functionality, insights) into a single usability score to identify the overall utility of the application.

III. Utility Framework Considerations From a Perception Assessment Study

Clinicians and researchers often set ambitious goals to advance medical knowledge via scientific discoveries in the medical world. Achieving this requires the acquisition of state-of-the-art techniques that are reported in the medical literature. While clinicians and researchers today often rely on online medical literature application such as PubMed, Scopus, or Google Scholar, the KnowCOVID-19 gateway looks to augment their clinical workflow and provide filtered papers according to the Levels of the Evidence pyramid [4]. In this section, we detail how we conducted a perception assessment study to identify the impressions of users about using a publication analytics service such as the KnowCOVID-19 and integrated tools such as Vidura Advisor for knowledge discovery. In addition, we outline the requirements to develop a framework that measures the utility of KnowCOVID-19 to demonstrate the relative advantage in individual user adoption compared to existing online literature search applications, such as Google Scholar.

A. Perception Assessment Study

Prior to developing KnowCOVID-19 and Vidura Advisor, we conducted a preliminary user perception research study where we interviewed 10 medical users to get their opinions on the advantages/disadvantages of their current manual literature search workflow. We also proposed the concept of automation with KnowCOVID-19 and the Vidura Advisor in a video presentation and asked questions regarding their impressions of such a method. Based on this, we obtained user opinions on suggested features that they desire in the workflow processes of KnowCOVID-19 with Vidura Advisor to improve their current literature search practices.

1) Perceptions on Manual COVID-19 Literature Search: Regarding users' traditional manual approach to searching for literature, several pitfalls became apparent in the context of COVID-19 research. Participants mentioned how the speed at which COVID-19 articles are being published online made traditional manual literature search approaches insufficient. While participants have expressed their comfort with using existing online medical literature search applications (e.g., PubMed, Google Scholar), they believed this deficiency can further lead to higher labor-intensive research and bedside care as well as slowing down their literature search workflow processes. Furthermore, high amounts of accessible literature are not peer-reviewed, which emphasized most of the participants' concerns with finding credible studies. Hence, without relying on the prestige of a journal, participants who seek to find peer-review literature go through a painstaking time commitment of perusing through each article to determine its quality. This time-consuming and risky assessment process is often subjective and requires both expertise and manual burden to be able to filter high-quality literature effectively.

2) Automated Search and Drill Down Options: After informing the participants on the concept of KnowCOVID-19 with Vidura Advisor via a video description, nearly all participants expressed their interests in the idea of filtering high-quality publications according to the Levels of Evidence. A few participants noted that this could address the significant challenge of manually parsing through new publications in a time-efficient manner. The inclusiveness of such a system was identified as a merit, as some participants noted that preprints and general news articles could influence the direction of their COVID-19 medical practice and bedside care. On the other hand, some participants suggested that KnowCOVID-19 should handle labeling literature based on the publication

venue, or lack thereof, in order to inform them of the publication type. A feature to facilitate the collaboration between researchers who share the same interests in topics and research objectives was also suggested. Several participants emphasized that the early application design would need to be intuitive, as a first impression of its ease of use would likely determine their intention to adopt the publication analytics service for further use. Overall, most participants agreed that KnowCOVID-19 could be a promising tool to improve their current manual literature search workflow.

3) Using Conversational Agent Assistance: The participants were also shown the concept behind an assisted conversational agent viz., Vidura Advisor that is integrated into KnowCOVID-19 for guidance on filtering COVID-19 literature. While most participants expressed their interest in using the Vidura Advisor, it was contingent that it could provide relevant and accurate assistance to their clinical tasks. Some participants expressed disdain for chatbots in general, as they prefer human assistance. It is important to note the participants who expressed this disdain either had a low-quality prior experience or did not find any effective use of chatbots due to the lack of intuitiveness of the application. All participants suggested the chatbot must rapidly respond to queries in an efficient manner for the AI-powered conversational agent to be adopted for long-term use. Furthermore, the function of Vidura Advisor should be used as a resource to complete the participants' searches rather than troubleshooting problems with application navigation. The participants' reactions to Vidura Advisor were mixed, however many noted the potential upside in obtaining assistance in COVID-19 publication analytics.

B. Utility Framework Requirements

Given the feedback from participants on the perception of KnowCOVID-19 with Vidura Advisor, we suppose that employing usability techniques in parallel to the science gateway development will promote user adoption. This is important to tangibly support the needs of clinicians and medical professionals who are using multiple search and drill-down tools for knowledge discovery related to the COVID-19 pandemic. In this context, we propose a framework to measure the usefulness of a publication analytics application that can allow for widespread user adoption to achieve medical research objectives and build collaborations for developing innovative solutions. To create such a framework to measure the utility of our KnowCOVID-19 science gateway, we outline the following requirements:

1) Assessment on Application Components: Assessing the utility of the COVID-19 literature search application is crucial for understanding the performance of medical users. While previous frameworks assess the utility of a platform from different aspects of the application (e.g., user interface, functionality, information insights), there presents a lack of an empirical evaluation that entails all aspects of the application development. Each level in the application development must be evaluated through a set of utility measurements. This segmented evaluation provides context on which components are more critical than others in terms of individual user adoption. Besides, this enables users to adopt online medical literature search applications based on multiple components. Hence, the utility measurement framework needs to fit the needs of consumers by assessing all aspects of the application and AI-assisted features (e.g., intelligent chatbot agents) that makes the application valuable for insights discovery.

2) Quantitative Metrics on Application Performance: A socio-technical perspective is required to measure the adoption of a COVID-19 publication analytics application. It can be assumed that user satisfaction is only as effective while performing literature searches as the tools/resources they are given. In other words, the abundance and relevance of literature search tools/resources can lead to improved knowledge discovery to foster innovations in pandemic-related solutions. Metrics that take into account application performance must therefore be quantitative and insightful. These metrics that encompass application performance should then be validated with users based on prior interactions on existing COVID-19 applications that feature literature search. The utility measurement framework should take advantage of the tools and content developed on KnowCOVID-19 and quantify the interaction between the software components and users who engage with it. Quantifying application performance can lead to better analysis of user satisfaction as opposed to solely relying on subjective methods.

3) Capturing User Perception based on Experience: According to DOI at the adoption level, users can accept, reject, or dis/continue the application. In addition to user performance, capturing medical users' perceptions of the application is important for understanding their impressions when using the KnowCOVID-19 with Vidura Advisor. These perceptions help to identify and assess the relative advantage KnowCOVID-19 has over existing online medical literature search applications. Diving into user perception serves as an effective tool for application feedback, as creators/developers of the application can iteratively improve their service and increase individual user adoption. From the perspective of a user, their perceptions can help influence more adoption to turn into implementation at the team level.

4) Platform Agnostic Utility Measurement: For effectively evaluating the relative advantage, the utility measurement framework must be applicable across various platforms related to KnowCOVID-19 with Vidura Advisor. Making the utility measurement framework platform agnostic allows for comparable analyses among similar online literature search applications. Quantitative measurements and statistical analysis techniques must be used to analyze user interaction and behavior for understanding the progression/digression and statistical significance between platforms, respectively. These quantitative metrics must assess the degree to which users can perform clinical tasks based on the tools/resources provided amongst different applications used for COVID-19 literature search. As a result, the utility measurement framework can remain consistent when evaluating between applications for literature search.

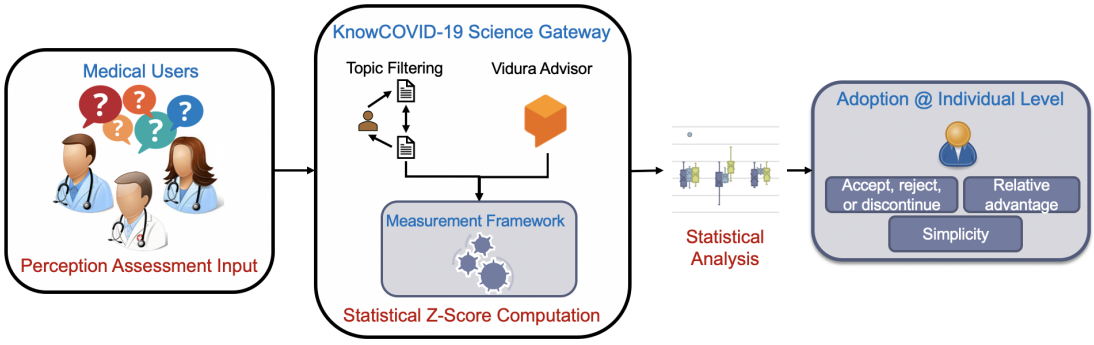


Fig. 2: Steps involved in the utility measurement framework when applied on the KnowCOVID-19 science gateway with Vidura Advisor to assess "Diffusion of Innovations" at the individual level.

IV. Utility Measurement Framework and KnowCOVID-19 Science Gateway Capabilities

In this section, we present our utility measurement framework based on the feedback from our perception assessment study that is utilized on top of the KnowCOVID-19 science gateway and the Vidura Advisor for individual-level adoption. As shown in Figure 2, the utility measurement framework implementation involves various steps that use medical user engagement over the KnowCOVID-19 with Vidura Advisor capabilities as input to calculate a z-score for a set of quantitative metrics using statistical techniques. Further analysis of these individual metrics are performed using statistical inferencing techniques such as MANOVA and Chi-square on user performance and perception to compare against applications such as Google Scholar for widespread adoption over clinical literature search tasks. In addition, we detail our KnowCOVID-19 science gateway with Vidura Advisor capabilities that are realized through software components to support our utility measurement framework implementation.

A. Utility Measurement Framework Description

1) Utility Performance Metrics: Our utility measurement framework is a combined function of three major aspects in the application development: user interface, functionality, and insights. Figure 3 illustrates how each utility component is composed of quantitative metrics that we define to assess the combined utility function. Table I describes the metrics used for our utility measurement framework to compute a score on application performance. Each of these aspects is part of the overall process that determines the utility of the science gateway application. While all aspects of the utility measurement framework can be analyzed separately, each component is critically important in evaluating the utility of the application. Herein, we define each of these aspects that together determine the usefulness of the application.

User Interface: Making an application feasible for user adoption relies on the ease of use of the application relative to its underlying purpose or intention. Hence, user performance depends on combination of prior or domain-specific knowledge with the application, and with the science gateway application capabilities. We define *user interface* as the ability

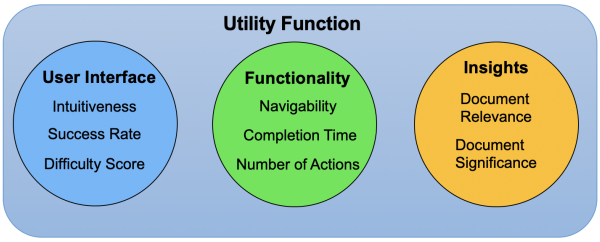


Fig. 3: Utility function obtained by combining all the quantitative metrics within each utility component of the application.

for an application to be usable and efficient from a science gateway design perspective with the following components:

- *Intuitiveness*: It measures the ease of use of the web application. This is calculated by dividing the overall time it takes over the number of actions performed on one task.
- *Success Rate*: This refers to the user's ability to complete a given task. We simply measure the success rate via binary indication (Y = 1, N = 0) of whether the user successfully completes the task at hand.
- *Difficulty Score*: This is a measurement of the user satisfaction based on how difficult a task was to perform. While it captures subjective user experience, we can quantify this metric by using a 5-point Likert scale [33].

Functionality: As the application has the necessary software components for providing user oriented features, it is important that the functions of the application work well to meet the users' task goals. In other words, the functions should enable the medical users to perform a specific action that aids in their literature search workflow process. Hence, the utility measurement framework details *functionality* as the ability for the science gateway components to properly function for a user on a given task with the following components:

- *Navigability*: It is the path taken from a start task to end task while operating a particular functionality. The score is normalized between 0 and 1 by calculating the ratio of the shortest path over the total path a medical user takes on a given task while using particular software functions.
- *Completion Time*: This is the duration (in seconds) of a participant completing a task while operating a particular functionality. We simply calculate this metric by subtract-

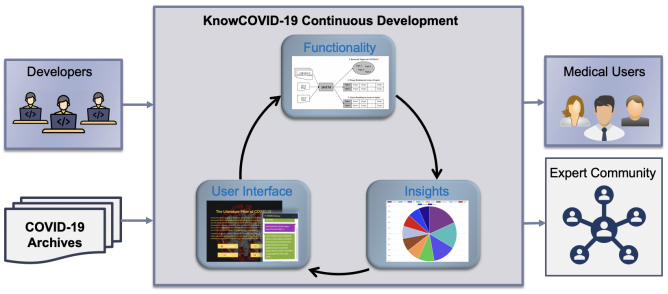


Fig. 4: Bridging the workflow for increasing utility of the KnowCOVID-19 with Vidura Advisor at the individual level.

ing the recorded end time from the start time of a given task while using particular software functions.

- *Number of Actions*: This is the number of events and interactions initiated by a user while operating a particular functionality. We define actions over a given task $A_T = \sum a$, where a is the set of valid actions on the website $(a_1, a_2, ..., a_n)$. These actions relate to use of particular software functions that include clickable links, completed search queries, switching between tabs, and enabling/disabling the chatbot interface.

Insights: The information the user receives from the application with or without the conversational agent assistance is crucial for furthering their research objectives. The results in terms of the literature retrieved must be accurate and reliable for knowledge discovery. This information should also be used for identifying the quality of evidence that relates to their clinical queries. We define *insights* as the relevance and importance of the information retrieved when performing a specified task on an application with the following components:

- *Document Relevance*: This denotes how well a retrieved document (or a set of documents) meets the information needs of the user. The relevance of a document is calculated by dividing the relevance score of each document by the total number of documents used to finish a task.
- *Document Significance*: Document significance utilizes the intuitiveness metric to calculate each document retrieved in a given task. The final score is the summation of all intuitiveness scores on each individual document retrieved and then averaged across all documents. This helps to assess the time it takes to find a relevant document in a result set.

2) Baselining Documents: In information retrieval, obtaining ground truth on the relevance of documents can be a difficult task. Given the collection of documents obtained from the CORD-19 dataset, we do not have the ground truth labels that indicate the relevance score between each query generated for performing COVID-19 related clinical tasks and each retrieval set. To alleviate this issue, we leverage the commonly known information retrieval method, Okapi BM25 [34], to compute the relevance scores as our ground truth for documents from the CORD-19 dataset. Okapi BM25 is a bag-of-word retrieval function that computes a relevance score and ranks a collection of documents based on the frequency of query terms that occurs in each document.

We define a given document D to be computed with a query Q with $q_1, q_2, ..., q_n$ query terms to obtain the relevance score. This score is computed by the probabilistic relevance scores $R(D, Q)$, which we can mathematically formalize as shown in Equation 1.

$$\sum_{i=1}^{n} IDF(q_i) \frac{f(D, q_i)(k+1)}{f(D, q_i) + k(1 - b + b(\frac{|D|}{avgdl}))}, \tag{1}$$

where $f(D, q_i)$ is the term frequency of a given query term q in a document D, $|D|$ is the length of the word document D, and $avgdl$ is the average length of the set of documents collected. Parameters k and b are predefined and can be optimized to increase the relevance score. In addition, we calculate the inverse document frequency score $IDF(q_i)$ of a given query term in Equation 2 as:

$$IDF(q_i) = \ln\Big(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\Big), \tag{2}$$

where N is the total number of documents in the corpus, and $n(q_i)$ is the number of documents containing q_i.

The result shows an aggregation of all relevance score values between the query and the collection of documents in the corpus. Given that our purpose is to compare the relevance scores of all query/document pairs on each application, we look to normalize the relevance score $R(D, Q)$ using a L2 norm that normalizes each non zero sample in the relevance query result set. The final output of the set of relevance ranges between 0-1, where scores that equal to one are the most relevant, while 0 shows no document relevance.

3) Utility Measurement Calculation: We define a utility measure score U as the combined score of all quantitative metrics used to assess application performance across various literature search applications. The utility components of user interface E, functionality C, and insights G make up the parameters of the utility function represented as $f(U) = f(E, C, G)$. We aggregate each aspect from the application development cycle to calculate a utility measurement score for assessing user performance. For this, we are motivated by [35] that combines standard utility measurements to assess user performance. Specifically, we accumulate all aspects of utility and average them into a z-score by the means of standardization. Given a single quantitative metric m of a particular utility component, we describe the z-score z of as the ratio between the difference of the *specification limit* $\bar{x}$, or benchmark score, from the mean x across all samples and the standard deviation σ, represented as:

$$z_m = \frac{x_m - \bar{x_m}}{\sigma_m}. \tag{3}$$

For a given utility component (e.g., user interface, functionality, insights), we calculate the summation the z-score of all metrics from that utility component in Equation 3 as:

TABLE I: Metric definitions used in the utility measurement framework that cover user interface, functionality, and insights in user tasks.

Metric	**Definition**	**Formula**	**Units**
Completion Time	Total number of seconds to complete a task	–	seconds (s)
Success Rate	Measures the successful completion rate of a task	–	1 = Y, 0 = N
Number of Actions	Number of events initiated by user input	$A_T = \sum a\epsilon(a_1, a_2, ..., a_n)$	–
Navigability [22], [26]	Path taken from start of task to end of task (normalized between 0 and 1)	$N_T = \frac{Min.pathcount}{Cur.pathcount}$	–
Intuitiveness [23], [24]	Ease of website use / interaction	$I_T = \frac{COMP-TIME(T_i)}{NUM-ACTIONS(T_i)}$	$\frac{s}{A}$
Document Relevance	Average relevance score of all retrieved documents	$R_T = \frac{\sum_{i=1}^{n} R(D_i,Q)}{\|D\|}$	–
Document Significance	Average intuitiveness score across all retrieved documents	$S_T = \frac{\sum_{i=1}^{n} I_T(D_i)}{\|D\|}$	–
Difficulty Score [33]	The task difficulty on a 5-point scale (1 = very easy, 5 = very difficult)	–	Likert Scale

$$z_\theta = \frac{\sum_{m=1}^{|\theta|} z_m}{|\theta|}, \tag{4}$$

where θ represents a particular utility component, $|\theta|$ is the length of quantitative metrics of a particular utility component, and z_m is the z-score of a given metric.

Additionally, task T is a set of n tasks represented by $\{t_1, ..., t_n\}$, where each task is the clinical assignment performed by a medical user on a given application. Using Equation 4, we calculate the z-score of given a task t and averaged it amongst the length of all quantitative metrics from each utility component, expressed as:

$$z_t = \frac{\sum_{i=1}^{|E|} z_i + \sum_{j=1}^{|C|} z_j + \sum_{i=k}^{|G|} z_k}{|E| + |C| + |G|}, \tag{5}$$

where z_i, z_j, and z_k are the calculated z-scores of the user interface, functionality, and insights components, respectively. $|E|$, $|C|$, and $|G|$ represent the total amount of quantitative metrics used for user interface, functionality, and insights, respectively. Finally, the total application standardization score is calculated by summing up the average z-scores of each utility component on a given task t and averaging them by the number of tasks performed on the application:

$$f(E, C, G)_\lambda = \frac{\sum_{t=1}^{|T|} z_t}{|T|}, \tag{6}$$

where λ is the literature search application, z_t is the z-score of a given task from Equation 5, and $|T|$ represents the total number of tasks performed on the application.

The final calculation of the application shown in Equation 6 represents the z-score of the application performance over a set of clinical tasks. In the following section, we detail the software capabilities of KnowCOVID-19 with the Vidura Advisor and show the ability of our utility measurement framework to be integrated into the science gateway application.

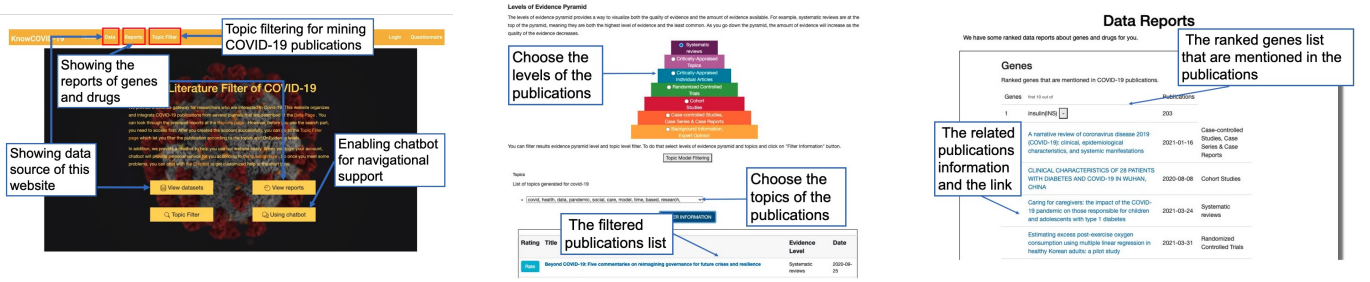
B. KnowCOVID-19 Science Gateway Capabilities

Figure 4 illustrates the software capabilities that aid in the utility of the KnowCOVID-19 science gateway. In these capabilities, we leverage the CORD-19 dataset to develop the KnowCOVID-19 science gateway user interface, functionality, and insights. Employing the utility measurement framework on top of KnowCOVID-19 helps bridge the workflow for increasing utility for users at the individual level.

1) System Components & Application Features: The KnowCOVID-19 science gateway is implemented using the Spring Boot [36] back-end development framework, which is a widely used framework in Java. Spring Boot is pre-configured and pre-sugared with a set of technologies that drastically minimize the manual efforts of configuration compared to conventional frameworks. We also use Apache Maven, which is a comprehensive build management tool to manage dependencies and versions, compile source code, run tests, package code into deployment-ready file formats, and deploy a final production code instance using Docker containers. In addition, we build the microservices for KnowCOVID-19 with Flask [37], which is a lightweight WSGI web application framework for seamless connection to the back-end Java application powered by Spring Boot.

As shown in Figure 5, the KnowCOVID-19 science gateway dashboard provides various features to enable users in effectively finding literature related to their clinical queries in a timely and automated manner. Herein, we detail each of the main features:

Data: The *Data* page provides context on the datasets we collected as well as links to view and download from its original source. For the KnowCOVID-19 science gateway, we collected the data from the CORD-19 dataset. This open-source dataset comprises of more than 280,000 scholarly articles from well-respected journals (e.g., New England Journal of Medicine, The Lancet Journal) as well as pre-prints (e.g., medRxiv, bioRxiv) about the novel coronavirus for the global research community. In addition, we also collected a set of drug and



(a) Dashboard capability.

(b) Topic Filtering capability.

(c) Reports capability.

Fig. 5: Three main capabilities of the KnowCOVID-19 science gateway application. The Dashboard capability introduces the science gateway and all related capabilities for users. The Topic Filtering capability provides the filtered literature list according to users' input. The Reports capability shows the genes and drugs reports for users along with related literature.

gene terms that are pertinent to the treatment, diagnosis, and prevention of COVID-19.

Reports: The set of drug and gene terms collected are important for users to investigate the key trends in the latest COVID-19 literature. This page gives users a text mining report on the amount or frequency of articles for each drug and gene term that appears - using a rule-based method. We organized the genes and drugs information to present a resultant set of documents to help users find which genes and drugs commonly appear in literature. Users have the ability to filter out the top 10 genes and drugs based on the amount of literature that has been published online. The resultant set of documents can be viewed by the evidence level or categorized based on the publication dates of the articles.

Topic Filter: Another main function of the KnowCOVID-19 gateway is to enable users to filter literature based on the Levels of Evidence. Based on our previous work in [1], we used an *evidence-based filtering* method that uses a pre-trained Domain-Specific Topic Model (DSTM) [2], [3] to find latent topics in the CORD-19 corpus, and the relevant set of drugs and gene terms. On this page, users can filter literature by selecting a specific level from the evidence pyramid. The DSTM displays a pre-defined set of topics for users to select and perform drill-down analysis. Correspondingly, the Topic Filter function searches the set of documents labeled under their respective evidence level and topic. The retrieved papers can then be viewed in their original source for literature filtering. This process shows how users can enhance their clinical workflow by minimizing the time to filter large sets of publications according to the evidence levels.

2) AI-assisted Conversational Agent for User Guidance: To help augment users' workflow processes, we employ an AI-based context-aware chatbot viz., Vidura Advisor that provides step-by-step navigation support and guides users in finding high-quality literature according to their clinical queries. The Vidura Advisor is integrated to fit our evidence-based filtering approach to enable users to filter documents. The Vidura Advisor is developed using Google DialogFlow, which is a natural language understanding platform (NLU). The Vidura Advisor learns the users' intentions based on user queries through a pre-trained NLU model that uses intents and entities to generate the tailored replies for users' inputs, which are meant to serve as questions topics. When users input the questions, the Vidura Advisor uses the input to match the intents through what is known as intent classification in order to extract useful information. For better performance, Vidura Advisor uses variation inputs about a question to train the NLU model. In addition, if users ask for replies that are not included in the intents, Vidura has fallback intent outputs to handle such cases.

Furthermore, we integrated the DSTM [2], [3] results into the Vidura Advisor training. For this, we first gathered a collection of unfiltered papers from the CORD-19 dataset. We then selected the publications which are most related to COVID-19 and that satisfy the user requirements. Also, we use a list of genes and drugs which are the most relevant to the COVID-19 literature. In addition, for generating topics with DSTM, we input related information with publications to train the model and identify the relationships among the collected publications and drug/gene terms using the Gibbs sampling parameter estimation algorithm from our DSTM. The results show the high frequency of drugs and genes information that are related to the literature to train the DSTM model. After finishing the training phrase, the DSTM can help analyze/visualize the most popular research topics in the literature. Moreover, the DSTM will rank the most commonly investigated drugs or genes based on each topic. We used this result to train the Vidura Advisor, so that the chatbot can also effectively help users to query COVID-19 relevant drugs and genes based on their research topics, or search relevant COVID-19 topics based on specific drugs and genes.

V. Performance Evaluation

In this section, we present an usability experiment of our utility measurement framework to assess user performance and perception of the KnowCOVID-19 science gateway with the Vidura Advisor. For this experiment, we collected over 6922 articles and pre-processed them from the CORD-19 dataset. The usability study was conducted with eight medical participants (three female and five male) from a Midwestern university and a Southwestern university in the United States. The occupation of the participants ranged from clinicians to

TABLE II: Usability questionnaire for measuring the perception of participants across metrics and user adoption.

Platform Usability Measures	
Measure	***Conceptual Definition***
Adapted USE Questionnaire (Lund, 2001)	
Usefulness	The utility of the platform
Ease of Use	How user-friendly the platform is
Ease of Learning	How quickly users can learn to use the platform
Satisfaction	Whether the user is happy with their experience with the platform and would recommend it to others
Adapted DOI Measures (Moore & Benbasat, 1991)	
Voluntariness	The degree to which users have a choice in using the platform at work
Relative Advantage	Whether the platform improves the efficiency of the user's workflow
Compatibility	Whether the platform would fit into a user's workstyle and routine
Image	The degree to which the use of the platform improves one's prestige at work
Simplicity/Ease of Use	The facility and comprehensibility of the platform
Result Demonstrability	How clear and explicable the outcomes of using the platform are
Visibility	To what degree the platform is known and used in one's organization
Trialability	Whether there was ample opportunity to try the platform before deciding to use it further
Chatbot Usability Measures	
Measure	***Conceptual Definition***
Adapted SASSI (Hone & Graham, 2000)	
System Response Accuracy	To what degree and how often the chatbot responds correctly
Likeability	How pleasant and easy the interaction with the chatbot is
Cognitive Demand	The level of stress and concentration experienced by the user while interacting with the chatbot
Annoyance	Whether interacting with the chatbot was repetitive, boring, or irritating in any way
Habitability	The confidence and level of understanding experienced by the user while interacting with the chatbot
Speed	The rate at which the chatbot responded

medical students. Areas of expertise included pulmonology, oncology, immunology, and more. Years of experience in the medical field ranged from less than one year to six years, and participants ranged from 23 to 35 years old. The usability experiment was conducted between the Google Scholar, KnowCOVID-19, and KnowCOVID-19 with the Vidura Advisor. The participants were required to perform a set of clinical tasks and retrieve papers related to COVID-19 related research. In addition, we also gauged the usability of Google Scholar, KnowCOVID-19, and KnowCOVID-19 with Vidura Advisor from each participant through a set of questionnaires [11]–[13], that use a 5-point Likert scale to assess usability metrics such as usefulness, ease of use, and relative advantage, as shown in Table II.

A. Application Performance Measurement

1) Experimental Setup: Participants' screens were observed and recorded in confidence via Zoom. This was followed by a debriefing of the study and an opportunity to ask questions based on their experience with the platform. The recorded video of the participants was privately stored for research to further analyze their performance and record down the metrics suggested by our utility measurement framework. Each participant was assigned two tasks to perform in their assigned experimental conditions. These clinical tasks were reviewed and approved by a clinician revolving around the literature search process related to COVID-19. The tasks are described as follows:

- *Task 1 (easy task):* You want to find the most rigorous studies on COVID-19 based on Levels of Evidence published since 2020. First, if needed, research Levels of Evidence to determine the most rigorous category. Then complete your search by finding 3 articles.
- *Task 2 (difficult task):* Find a publication (or set of literature) that details the treatment of Chloroquine (CQ) and Hydroxychloroquine (HCQ) for COVID-19. First, find publication(s) that details treatment for CQ/HCQ. Next, please explain whether CQ/HCQ is useful for treatment for COVID-19.

The participants were assigned tasks over a condition that followed the within-subject design. *Condition 1 (C1)* involved participants completing Task 1 and Task 2 utilizing Google Scholar, followed by completing the same tasks using KnowCOVID-19. Similarly, *Condition 2 (C2)* required participants to complete Task 1 and Task 2 using KnowCOVID-19 without the guidance of Vidura Advisor, followed by completing the same task with Vidura Advisor. Half of the participants tested on *C1* (Google Scholar vs. KnowCOVID-19) while the other half tested on *C2* (KnowCOVID-19 vs. KnowCOVID-19 with Vidura). Using the utility measurement framework, we observe the application performance for the participant with regards to each task and manually record this information to conduct an analysis. For computing the scores related to document relevance, we leverage the standard Okapi BM25 information retrieval method detailed in Section IV-A.

TABLE III: Performance results with statistical significance for continuous metrics over the candidate applications: Google Scholar, KnowCOVID-19, and KnowCOVID-19 with Vidura.

Condition	**Avg. Significance**		**Task 1**		**Task 2**	
	M	***SD***	***M***	***SD***	***M***	***SD***
Google Scholar (n = 4)	19.68	5.80				
Completion Time (s)			209.50	64.16	105.25	21.01
Navigability			0.50	0.15	0.65	0.26
Intuitiveness			16.36	3.67	32.46	10.24
Relevance Score			0.12	0.17	0.46	0.25
KnowCOVID-19 (n = 8)	19.84	8.77				
Completion Time (s)			236.87	156.25	321.50	128.37
Navigability			0.54	0.27	0.46	0.20
Intuitiveness			16.40	6.19	24.24	14.91
Relevance Score			0.16	0.15	0.21	0.14
KnowCOVID-19 with Vidura (n = 4)	12.85	6.43				
Completion Time (s)			278.75	178.04	320.75	147.06
Navigability			0.43	0.20	0.26	0.11
Intuitiveness			15.37	6.11	13.45	8.46
Relevance Score			0.12	0.12	0.17	0.18

TABLE IV: z-scores calculation results over each task to assess the utility across multiple applications.

Platform	**Task 1**	**Task 2**	**Application Score**
Google Scholar	0.48	-2.57	-1.05
KnowCOVID-19	0.05	0.05	0.05
KnowCOVID-19 with Vidura	-0.20	-0.03	-0.12

2) Statistical Technique for Z-score Calculation: The calculation of the z-scores given our quantitative metrics are crucial for understanding the utility of a publication analytics application for literature search. Given our statistical technique to compute the z-scores of an application from Equation 6, we had to standardize our quantitative metrics to demonstrate its z-score equivalent. We solve this by looking at the distribution of medical user performance to compute the statistics (e.g., mean M and standard deviation SD) of the continuous and discrete metrics over each assigned clinical task. For finding the specification limit $\bar{x}$ (benchmark score) referenced from Equation 3, we average the values of participants who succeeded on a clinical task for completion time, navigability, number of actions, intuitiveness, and document significance.

For the remaining metrics such as success rate, document relevance, difficulty score, we look at existing works (e.g., SUM [35], Okapi BM25 [34], Likert [33]) to guide us in benchmarking each metric for finding the specific limit for our z-score calculation. First, we leverage the technique used in SUM [35] to find the specification limit by computing the ratio of defects (or failed tasks) over the total tasks attempted. Second, using the Okapi BM25 in Equation 1, we set a predefined number $K = 20$ to determine the most relevant papers in our corpus based on a clinical query. Finally, we use the recommended 5-point Likert scale value for determining task difficulty as 4.0/5.0. Each metric is then calculated using our z-score formula and converted to a percentage using a standard normal distribution table.

3) User Performance Results: The results of the user performance experiments are shown in Table III. Participants in *Condition 1* (i.e., Google Scholar vs. KnowCOVID-19) showed consistently higher mean scores M across metrics such as navigability, intuitiveness, and average document significance when using KnowCOVID-19 over Task 1 (easy clinical task), but reported lower scores over the much difficult Task 2. Furthermore, the participants in *Condition 2* (i.e., KnowCOVID-19 vs. KnowCOVID-19 with Vidura) did not show any improvement when using the Vidura Advisor on KnowCOVID-19 across all quantitative metrics. This could be because of the lack of perceived usefulness from participants about the Vidura Advisor that could add in knowledge discovery for literature search over COVID-19 publications. In addition, we performed a statistical inference analysis for independent metrics using Chi-square for each of the conditions over the two clinical tasks. This did not show a significant difference in success rates between the two conditions, nor did a one-way between-groups multivariate analysis of variance (MANOVA) yield significant results for the other performative measures (F = 1.69, p = 0.18; Wilks' Lambda = 0.11; partial eta squared = 0.66.)

Furthermore, we report the results of the z-scores of each application over each clinical task, which are shown in Table IV. Google Scholar had the highest averaged score (z=0.48) compared to KnowCOVID-19 (z=0.05), and KnowCOVID-19 with Vidura (z=-0.20) over Task 1. The participants assessment of finding rigorous papers according to evidence-based practice techniques highly depended on the citation count as well as the content quality of the article. In Task 2, KnowCOVID-19 reported the highest averages score (z=0.05) among the three applications for COVID-19 literature search. The overall assessment of each application showed that KnowCOVID-19 had the highest utility application score

TABLE V: Results of the medical user perception measurements over the candidate applications: Google Scholar, KnowCOVID-19, and KnowCOVID-19 with Vidura.

	Google Scholar		**KnowCOVID-19**		**KnowCOVID-19 with Vidura**	
Measure	***M***	***SD***	***M***	***SD***	***M***	***SD***
Adapted USE Questionnaire (Lund, 2001)						
Usefulness	3.30	0.42	3.76	0.91	3.44	0.40
Ease of Use	3.75	1.41	3.80	0.65	3.56	1.14
Ease of Learning	4.50	0.70	4.53	0.56	4.17	1.04
Satisfaction	3.50	0.71	3.84	1.04	3.05	0.82
Adapted DOI Measures (Moore & Benbasat, 1991)						
Voluntariness	2.50	2.12	3.40	1.52	3.38	1.25
Relative Advantage	3.90	0.14	3.96	0.67	3.50	0.35
Compatibility	3.33	0.94	3.60	1.16	3.58	0.17
Image	4.33	0.47	3.07	0.60	2.33	0.94
Simplicity/Ease of Use	3.88	0.88	4.15	0.49	3.56	1.07
Result Demonstrability	4.25	0.00	3.95	0.76	3.81	0.55
Visibility	2.75	0.35	3.10	0.42	2.75	0.50
Trialability	4.25	0.35	4.40	0.42	4.00	0.41
SASSI (Hone & Graham, 2000)						
System Response Accuracy					3.19	0.87
Likeability					3.44	.83
Cognitive Demand					3.56	0.77
Annoyance					3.40	0.28
Habitability					3.06	0.59
Speed					4.25	0.65

(z=0.05), which is 47% higher than Google Scholar (z=-1.05). KnowCOVID-19 with Vidura performed slightly worse than KnowCOVID-19 with a 7% lower score (z=-0.12), which stemmed from most of the users' perceptions of the chatbot for feasible use of their clinical literature search tasks.

B. Participant Interviews & Usability Questionnaire Analysis

Following the completion of their tasks, participants were briefly interviewed about their experience with the resources they were assigned, after which they completed a questionnaire covering the usability of the platforms they used. During the interviews, participants expressed how they appreciated the simplicity of KnowCOVID-19 and how it facilitated finding articles filtered by Levels of Evidence. This was particularly expressed in comparison to finding articles on Google Scholar. A multivariate inferential statistics analysis using MANOVA was conducted to measure statistical significance between the Google Scholar, KnowCOVID-19, and the Vidura Advisor.

1) User Adoption Questionnaire: Table V captures the perception of users through a 5-point Likert scale over a set of questionnaires we adopted for assessing user adoption. We gauge the usability of KnowCOVID-19 by leveraging the USE questionnaire [11] to measure and quantify notable metrics such as usefulness, ease of use, ease of learning, and satisfaction. Furthermore, DOI measures have also been adapted from [12] to measure voluntariness, relative advantage, compatibility, image, simplicity/ease of use, result demonstrability, visibility, and trialability. Both measures utilize the 5-point Likert scale type questions. To measure the usability of the Vidura chatbot, SASSI questionnaire [13], [38] was adapted to measure the chatbot's system response accuracy, likeability, cognitive demand, annoyance, habitability, and speed as measured by the 5-point Likert scale type questions.

2) User Perception Results: According to the USE questionnaire, KnowCOVID-19 reported the highest mean scores according to usefulness, ease of use, ease of learning, and satisfaction of the application compared to Google Scholar. The DOI measures also showed consistently higher scores when using KnowCOVID-19 compared to the subsequent platforms for literature search usability. Due to the absence of an AI-assisted chatbot agent in Google Scholar and KnowCOVID-19, we report the mean scores adopted from the SASSI questionnaire to give context on the Vidura Advisor based on its system response accuracy, likeability, cognitive demand, annoyance, habitability, and speed. The idea score of acceptance of a technology according to Hone & Graham [38] was 4.0/5.0, or 80% (i.e., 20% for annoyance). The metrics did not show acceptance of the technology feature in almost all metrics except for annoyance and speed. Overall, a one-way between-groups multivariate analysis of variance (MANOVA) did not find any significant difference between platforms (Google Scholar, KnowCOVID-19, and KnowCOVID-19 with Vidura), (F (9, 1) = 2.07, p = 0.50; Wilks' Lambda = 0.05; partial eta squared = 0.95.)

Participants also had some suggestions for improvement. Regarding topic model filtering, one user suggested that it would be helpful to be able to check which topics were desired individually rather than having to choose from a list. Multiple users expressed a desire to be able to sort articles by date. In the Reports section, participants would like to see extension

of the filtering capabilities to include Level of Evidence as well. While most participants were able to complete their tasks successfully, mistaken queries slowed their progress at times.

C. Discussion

Despite the upside in conceptualizing the quantitative analysis in application performance with the utility measurement framework, much improvements need to be made to enhance the framework. The recording of such entries made the process for users to be time-consuming and laborious. The current framework lacks automated features to calculate various metrics without the need to validate with manual efforts. Not to mention, this can cause a problem from human error as the start and end time may not have been accurately recorded across all tasks. Similarly, our navigability metric was manually calculated as we assessed the number of actions. Previous studies such as [26] study human interaction between pages through a graph optimization problem. It is necessary to alleviate this deficiency by formalizing the links between web pages and features through a graph representation. Allowing automated features and structured flow between web pages and features could reduce the burden of manually tracking down application performance influencing human behavior on the KnowCOVID-19 science gateway.

Furthermore, the perceptions of the users after performing clinical trials on the KnowCOVID-19 science gateway were solely about the application itself. Participants had concerns over the results set, as they believed sorting the retrieved articles in alphabetical order was not sufficient for them to find relevant information according to their query. Some participants suggested adding more advanced search features that can help them filter the results down even further. Other participants noted that the personalization of queries was important when performing search methods, and they sometimes did not find any use in utilizing the pre-defined topics provided by the DSTM [3]. Other participants after the study suggested the application needed to be more intuitive and that there is a high learning curve of performing clinically-related tasks. Thus, improvements to the overall KnowCOVID-19 science gateway with the Vidura Advisor were noted by the users for higher utility and adoption amongst individual users.

VI. Conclusion

In this paper, we presented a utility measurement framework that addresses the individual-level adoption challenges among medical users (e.g., clinicians, researchers, medical students) in COVID-19 applications for literature search. The design of the utility measurement framework considered the feedback from medical users in a perception assessment study regarding their perceptions of the KnowCOVID-19 science gateway and Vidura Advisor for publication analytics. We defined utility to be a function of user interface, functionality, and insights that leverages a statistical technique to calculate a z-score over a set of quantitative metrics. We detailed how our utility measurement framework is supported by the KnowCOVID-19 science gateway through the software capabilities that assist users in publication analytics tasks such as text mining and literature search.

We validated the utility of our measurement framework by developing a usability study that compares the user performance via z-score calculation between Google Scholar, KnowCOVID-19, and KnowCOVID-19 with Vidura based on our quantitative metrics. In the first part of our study, we assigned participants to test between two platforms either under *Condition 1* (Google Scholar vs. KnowCOVID-19) or *Condition 2* (KnowCOVID-19 vs. KnowCOVID-19 with Vidura). In the second part of the study, we followed up with a usability study questionnaire to capture users' experiences and perceptions of Google Scholar, KnowCOVID-19, and the Vidura Advisor. While our results demonstrate a 47% higher utility z-score calculation using KnowCOVID-19 over Google Scholar, inferential statistic analyses using Chi-square and MANOVA for user performance and perception did not show any significant difference between the applications. This leads us to further develop the KnowCOVID-19 with the Vidura Advisor for improving the utility measurement scores and statistical significance for publication analytics.

In future works, we plan to iteratively develop our KnowCOVID-19 science gateway assisted with the Vidura Advisor to improve our utility measurement scores for improved adoption at the individual level as well as implementation at the team level in the medical community. In addition, we will investigate how the generation of these scores can guide us to develop a user proficiency estimation to better present information to the medical users. This will allow our AI-based intelligent agent i.e., Vidura Advisor to refine the type of guidance presented to the user based on their performance while using the science gateway capabilities.

Acknowledgment

This work is supported by the National Science Foundation under awards: OAC-2006816 and OAC-2007100. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] R. Oruche, V. Gundlapalli, A. P. Biswal, P. Calyam, M. L. Alarcon, Y. Zhang, N. R. Bhamidipati, A. Malladi, and H. Regunath, "Evidence-based recommender system for a covid-19 publication analytics service," *IEEE Access*, 2021.

[2] Y. Zhang, P. Calyam, T. Joshi, S. Nair, and D. Xu, "Domain-specific topic model for knowledge discovery through conversational agents in data intensive scientific communities," in *2018 IEEE International Conference on Big Data (Big Data)*, 2018, pp. 4886–4895.

[3] ——, "Domain-specific topic model for knowledge discovery in computational and data-intensive scientific communities," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[4] D. Timm, "Evidence Matters," *Journal of the Medical Library Association: JMLA*, vol. 94, no. 4, p. 480, 2006.

[5] M. H. Murad, N. Asi, M. Alsawas, and F. Alahdab, "New Evidence Pyramid," *BMJ Evidence-Based Medicine*, vol. 21, no. 4, pp. 125–127, 2016.

[6] A. A. Chandrashekara, R. K. M. Talluri, S. S. Sivarathri, R. Mitra, P. Calyam, K. Kee, and S. Nair, "Fuzzy-based conversational recommender for data-intensive science gateway applications," in *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 2018, pp. 4870–4875.

[7] E. M. Rogers, *Diffusion of innovations*. Simon and Schuster, 2010.

[8] J. Sauro and J. R. Lewis, "Correlations among prototypical usability metrics: Evidence for the construct of usability," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2009, pp. 1609–1618.

[9] C.-M. Karat, R. Campbell, and T. Fiegel, "Comparison of empirical testing and walkthrough methods in user interface evaluation," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1992, pp. 397–404.

[10] M. Ekin Eren, N. Solovyev, E. Raff, C. Nicholas, and B. Johnson, "COVID-19 Kaggle Literature Organization," *arXiv e-prints*, pp. arXiv–2008, 2020.

[11] A. M. Lund, "Measuring usability with the use questionnaire12," *Usability interface*, vol. 8, no. 2, pp. 3–6, 2001.

[12] G. C. Moore and I. Benbasat, "Development of an instrument to measure the perceptions of adopting an information technology innovation," *Information systems research*, vol. 2, no. 3, pp. 192–222, 1991.

[13] K. S. Hone and R. Graham, "Towards a tool for the subjective assessment of speech system interfaces (sassi)," *Natural Language Engineering*, vol. 6, no. 3-4, pp. 287–303, 2000.

[14] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS quarterly*, pp. 319–340, 1989.

[15] F. F. Fuller, "Concerns of teachers: A developmental conceptualization," *American educational research journal*, vol. 6, no. 2, pp. 207–226, 1969.

[16] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS quarterly*, pp. 425–478, 2003.

[17] D. Persico, S. Manca, and F. Pozzi, "Adapting the technology acceptance model to evaluate the innovative potential of e-learning systems," *Computers in Human Behavior*, vol. 30, pp. 614–622, 2014.

[18] S. Samar, M. Ghani, and F. Alnaser, "Predicting customer's intentions to use internet banking: the role of technology acceptance model (tam) in e-banking," *Management Science Letters*, vol. 7, no. 11, pp. 513–524, 2017.

[19] R. H. Hoyle, *Structural equation modeling: Concepts, issues, and applications*. Sage, 1995.

[20] S. Min, K. K. F. So, and M. Jeong, "Consumer adoption of the uber mobile application: Insights from diffusion of innovation theory and technology acceptance model," *Journal of Travel & Tourism Marketing*, vol. 36, no. 7, pp. 770–783, 2019.

[21] G. J. Doyle, B. Garrett, and L. M. Currie, "Integrating mobile devices into nursing curricula: Opportunities for implementation using rogers' diffusion of innovation model," *Nurse education today*, vol. 34, no. 5, pp. 775–782, 2014.

[22] Y. Zhang, H. Zhu, and S. Greenwood, "Web site complexity metrics for measuring navigability," in *Fourth International Conference onQuality Software, 2004. QSIC 2004. Proceedings.* IEEE, 2004, pp. 172–179.

[23] M. N. Islam, "Exploring the intuitiveness of iconic, textual and icon with texts signs for designing user-intuitive web interfaces," in *2015 18th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2015, pp. 450–455.

[24] S. R. Serge, J. A. Stevens, and L. Eifert, "Make it usable: highlighting the importance of improving the intuitiveness and usability of a computer-based training simulation," in *2015 Winter Simulation Conference (WSC)*. IEEE, 2015, pp. 1056–1067.

[25] S. Roy, P. K. Pattnaik, and R. Mall, "A quantitative approach to evaluate usability of academic websites based on human perception," *Egyptian Informatics Journal*, vol. 15, no. 3, pp. 159–167, 2014.

[26] M. Chen and Y. U. Ryu, "Facilitating effective user navigation through website structure improvement," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 3, pp. 571–588, 2011.

[27] N. Nizam, C. Watters, and A. Gruzd, "Improving website navigation with the wisdom of crowds," in *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, 2016, pp. 337–339.

[28] M. Georgsson and N. Staggers, "Quantifying usability: an evaluation of a diabetes mhealth system on effectiveness, efficiency, and satisfaction metrics with associated user characteristics," *Journal of the American Medical Informatics Association*, vol. 23, no. 1, pp. 5–11, 2016.

[29] S. Charfi, H. Ezzedine, and C. Kolski, "Rita: a user interface evaluation framework." *J. Univers. Comput. Sci.*, vol. 21, no. 4, pp. 526–560, 2015.

[30] S. G. López, F. M. Simarro, and P. G. López, "Balores: A framework for quantitative user interface evaluation," in *New Trends in Interaction, Virtual Reality and Modeling*. Springer, 2013, pp. 127–143.

[31] S. González, F. Montero, and P. González, "Balores: a suite of principles and metrics for graphical user interface evaluation," in *Proceedings of the 13th International Conference on Interacción Persona-Ordenador*, 2012, pp. 1–2.

[32] A. Dingli and S. Cassar, "An intelligent framework for website usability," *Advances in Human-Computer Interaction*, vol. 2014, 2014.

[33] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, "Likert scale: Explored and explained," *British Journal of Applied Science & Technology*, vol. 7, no. 4, p. 396, 2015.

[34] S. Robertson and H. Zaragoza, *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc, 2009.

[35] J. Sauro and E. Kindlund, "A method to standardize usability metrics into a single score," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2005, pp. 401–409.

[36] "Spring boot introduction," https://spring.io/projects/spring-boot, accessed: 2021-11-05.

[37] "Flask introduction," https://www.palletsprojects.com/p/flask/, accessed: 2021-11-05.

[38] K. Hone, "Usability measurement for speech systems: Sassi revisited," in *SIGCHI Conference Paper, Toronto*, 2014.