Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

SPECIAL ISSUE JOURNAL

Science Gateway Adoption using Plug-in Middleware for Evidence-based Healthcare Data Management

Roland Oruche[1] | Eric Milman[2] | Mauro Lemus[1] | Xiyao Cheng[1] | Ashish Pandey[1] | Songjie Wang[1] | Prasad Calyam[1] | Kerk Kee[2]

[1]Department of Electrical Engineering and Computer Science, University of Missouri, Missouri, USA
[2]College of Media and Communication, Texas Tech University, Texas, USA

Correspondence
Prasad Calyam
Email: calyamp@missouri.edu

Summary

There is a growing need for next-generation science gateways to increase the accessibility of emerging large-scale datasets for data consumers (e.g., clinicians, researchers) who aim to combat COVID-19-related challenges. Such science gateways that enable access to distributed computing resources for large-scale data management need to be made more programmable, extensible and scalable. In this paper, we propose a novel socio-technical approach for developing a next-generation healthcare science gateway viz., *OnTimeEvidence* that addresses data consumer challenges surrounding the COVID-19 pandemic related data analytics. OnTimeEvidence implements an intelligent agent viz., *Vidura Advisor* that integrates an evidence-based filtering method to transform manual practices and improve scalability of data analytics. It also features a plug-in management middleware that improves the programmability and extensibility of the science gateway capabilities using microservices. Lastly, we present a usability study that shows the important factors from data consumers' perspective to adopt OnTimeEvidence with chatbot-assisted middleware support to increase their productivity and collaborations to access vast publication archives for rapid knowledge discovery tasks.

KEYWORDS:
Science Gateways, Microservices, Intelligent Middleware, Intelligent Agents, Diffusion of Innovations

1 | INTRODUCTION

Managing the emerging collections of large-scale medical datasets such as scientific publications and electronic health records (EHRs) can be a challenging task for medical data consumers (e.g., clinicians, medical researchers) who need to make timely decisions for combating COVID-19-related issues. Data consumers are constantly faced with complex tasks that are labor-intensive and require domain-specific knowledge discovery over medical information for critical decision making. When synthesizing scientific literature for knowledge discovery, data consumers often rely clinical methodologies such as Levels of Evidence[1,2] to improve information reliability and reduce the quantity of literature by prioritizing scientific rigor (e.g., expert opinions to systematic reviews and meta-analyses). The challenge of manually filtering high-volume of literature based on evidence-based methods presents the need from data consumers to adopt next-generation science gateways to gain access to emerging large-scale datasets and resources for developing timely pandemic-related solutions.

Science gateways previously have been able to alleviate the aforementioned challenge for data consumers across scientific domains, including bioinformatics, neuroscience, physics, chemistry, and material science[3,4]. They are capable of hiding the complexities for domain users via easy-to-use domain application interfaces to access distributed computing resources and emerging datasets that tailor their needs for a variety of scientific tasks related to research and education. They also allow data providers (e.g., administrators, developers) to manage and provide access to data consumers distributed computing resources and large-scale datasets for large-scale data management. However, most science gateways today are built for specific purposes with pre-defined workflows, user interfaces, and fixed computing resources. Such a state-of-practice makes it difficult for: (a) data consumers, whose science gateway needs constantly evolve in terms of e.g., data-intensive workflow automation, and (b) data providers, whose require dynamic choice of cloud computing platform resources to match growing data handling demands. The next-generation of science gateways we envision for knowledge discovery (e.g., through data navigation, searching, sharing) that can handle constantly growing datasets and corresponding analytical tools need to be programmable, extensible, and scalable.

For this, the architecture designs of next-generation science gateways need the ability to automate guidance in user interfaces while dynamically providing data resources (e.g., advanced analytical tools for COVID-19 publication archives) based on the ever-growing community needs. Also, science gateway administrators/developers are often challenged to integrate advanced technologies (e.g., knowledge bases, recommenders, machine learning tools) due to architecture design limitations in today's science gateways that do not support programmability. In addition, the lack of programmability makes it difficult to manually orchestrate handling of a rightly scaled number of cloud resources for hosting microservices and storing data from disparate sources in order to achieve high application performance. As a result, data providers need application programming interfaces and guided interface methods to avoid manual integration of advanced technologies to acquire/manage/process data to generate data processing pipelines that meet evolving data consumer demands in large-scale data management. To overcome above issues, some key solution approaches must be considered: (i) motivated by Diffusion of Innovations theory[5], science gateways must address challenges of widespread adoption at the individual level amongst data consumers through the implementation of guided intelligent interfaces, and dynamic provisioning of large-scale medical resources and datasets, and (ii) science gateways need to be modernized with 'plug-in' management middleware support for data providers to programmatically increase their extensibility and scalability to meet the users' growing large-scale data management needs.

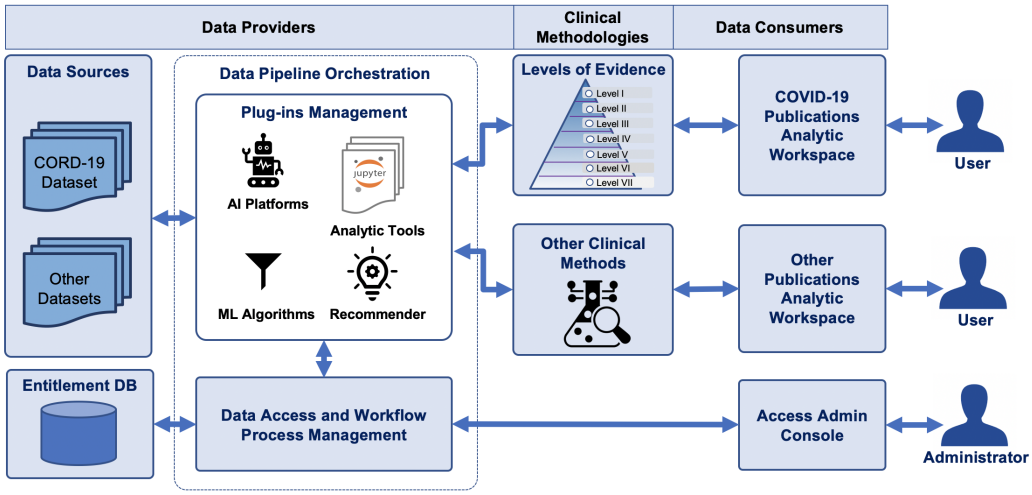

FIGURE 1 The Vidura plug-in management middleware components required for the data pipeline orchestration to help data providers and data consumers in handling publication analytics for OnTimeEvidence workspace.

In this paper, we propose a novel socio-technical approach that addresses the aforementioned challenges surrounding the COVID-19 pandemic by developing a next-generation science gateway viz., *OnTimeEvidence*. OnTimeEvidence helps data consumers with easy access to publication archives and related analytics tools for developing rapid, data-intensive solutions in the context of COVID-19 publications (i.e., publication analytics) and its correlation to EHRs. It implements an intelligent agent viz., *Vidura Advisor* that integrates a plug-in management middleware approach for programmable science gateways. The benefits from the Vidura plug-in management approach to develop OnTimeEvidence include its ability to generalize across

programmable science via microservice architectures for application decoupling and allowing data consumers and data providers to customize and incorporate domain-specific components. The Vidura plug-in management middleware development is inspired by a "pluganized" management framework developed in[6], which enables plug-in of network protocols as extensions to support fast/secure data transmission.

Figure 1 shows the creation of OnTimeEvidence and other literature workspaces through our generalized Vidura plug-in management middleware. This allows data providers to customize and orchestrate data processing pipelines for augmenting the clinical methodology workflows of data consumers. OnTimeEvidence uses a social approach that guides the personalized design of the intelligent agent to foster *adoption* of next-generation science gateways by data consumers in data analytics and knowledge-driven collaborations. This approach of OnTimeEvidence with the Vidura plug-in middleware for *adoption* at the individual level can be further extended to other socio-technical aspects of *implementation* and *diffusion* for data consumers at the organization and community levels, respectively. The technical approach features a microservice-based workflow automation over distributed computing resources in order to support data provider tasks of data lake creation from disparate data sources, data transformation, and access/management control. The source code of OnTimeEvidence is publicly available on GitHub[7]. OnTimeEvidence features application programming interfaces (APIs) in the Vidura plug-in management middleware for creation of domain-specific end-to-end pipelines with diverse infrastructure, customized processes, detailed monitoring, and flexible programmability for existing science gateways such as OnTimeEvidence. The APIs address the challenges of developing next-generation science gateways with the following benefits:

- *Modularity:* Vidura plug-in management middleware leverages microservice architectures that decouples the science gateway application code, and enables addition of new services that function independently but interconnect to each other. Consequently, the microservices code can be reused by processes that have similar execution behavior across multiple science gateways.

- *Extensibility:* Vidura plug-in management middleware allows science gateway data providers to easily extend the middleware with additional application components or plug-ins such as, e.g., multi-cloud based workflows with templates, execution pipelines involving machine learning (ML) algorithms, and more.

- *Scalability:* Vidura plug-in management middleware allows users to reserve pre-configured and ready-to-use cloud infrastructure resources in order to help them scale their workloads as and when needed. Replacing a microservice for a different scale of resource needs allows for dynamic resource management for changing user workflow needs.

- *Programmability:* Vidura plug-in management middleware helps science gateway providers to program, register and upload various components into their existing setup using a customizable application interface.

Lastly, we present a usability study that shows the important needs from data consumers to adopt our OnTimeEvidence science gateway that integrates the Vidura plug-in management middleware to increase their productivity and collaborations to access vast emerging publication archives for rapid knowledge discovery tasks. While our solution approach addresses the needs for both data consumers and data providers, the scope of the presented usability study is directly aimed to increase the adoption of OnTimeEvidence for data consumers. In the usability study, we conduct a qualitative assessment with 10 participants ranging from clinicians, researchers and healthcare practitioners. We asked a set of questions regarding their current literature review workflow process, as well as their suggestions on adopting the OnTimeEvidence science gateway with Vidura Advisor. Based on their expertise and domain knowledge, all the participants highly approve of the publication analytics features of the OnTimeEvidence science gateway supported by our Vidura plug-in management middleware. Further, most of the participants expressed interest to leverage our Vidura Advisor and the proposed plug-ins for clinical guidance as well as to keep up with the deluge of emerging scientific literature, especially in the area of COVID-19 related research.

The remainder of the paper is organized as follows: Section 2 presents related work. Section 3 provides background and motivation for OnTimeEvidence with the Vidura plug-in management middleware for both data providers and data consumers. Section 4 presents an overview of OnTimeEvidence and details the Vidura plug-in management middleware design and development. In Section 5, we present the *adoption* usability study to evaluate our OnTimeEvidence next-generation science gateway with the Vidura plug-in management middleware from a data consumer perspective. Finally, Section 6 concludes the paper.

2 | RELATED WORK

2.1 | COVID-19-based Platforms for Data Consumers

Previous works have sought to develop platforms to resolve the issues of accessibility, management, and analytics of COVID-19-related datasets for data consumers. Ahmed *et. al.*[8] develops an end-to-end data management the workflow process viz., COVID-19 Data Summarization and Visualization (DSV), which handles large-scale surveillance data on the outbreak of the virus for pandemic response in the World Health Organization (WHO) African Region. Similarly, Peddireddy *et. al.*[9] created a COVID-19 Surveillance Dashboard to provide a visualization tool for healthcare consumers and authorities to compare, organize, and track large-scale surveillance data in pandemic progression. This platform enables healthcare consumers/authorities to customize visually-guided interfaces in the form of dashboards to make timely decisions in responding to pandemic-related issues. iResponse[10] addressed the challenges in monitoring and multi-term planning of COVID-19 related challenges by developing a data-driven framework for coordination and pandemic management. Authors in[11] developed an IoT-based framework that performs real-time application tracking and monitoring of COVID-19 data. The framework aggregates emerging data from disparate sources into a cloud-based infrastructure where healthcare consumers such as physicians can manage patients through data processing pipelines.

The dissemination of COVID-19 literature has also been critical in the healthcare community in fostering knowledge discovery and critical decision making. Previous works have demonstrated the need to develop literature analytics-based platforms to address data management and analytic challenges related to the pandemic. IBM Deep Search[12] leverages ML models to extract content from ingested documents using the Corprus Conversion Service with high accuracy. They leverage the Corpus Processing Service[13] to apply natural language processing techniques (e.g., Named-Entity Recognition, Fact Extraction) to build knowledge graphs for data consumers in performing contextual information retrieval and knowledge discovery. COVID-Scholar[14] handles the needs of data consumers by developing a literature analytic platform that collects, processes, and manages over 150,000 documents for actionable insights. It continuously integrates new documents via data pipelines and processes this information using natural language processing techniques such as SciBERT[15] and dimensionality reduction. This enables both the classification of articles based on design type and visual representations of word embeddings. Other services such as the work in Moran *et. al.*[16] implement a visual literature analytics platform viz., PLATIPUS, provisions data consumers guided user interfaces to explore their clinical queries of interest related to the pandemic. Similarly, CoronaCentral[17] offers a data-driven visualization platform to enable consumers to feasibly search through emerging COVID-19 literature via a detailed categorization scheme.

In this work, our novelty is in the development of OnTimeEvidence, a next-generation science gateway that leverages an evidence-based filtering method[18] and a plug-in management middleware viz., Vidura Advisor, for provisioning and managing of large collections of COVID-19 and related EHR datasets with suitable computational resources. Through the implementation of the Vidura plug-in management middleware, OnTimeEvidence integrates visual- and data-driven models to increase the accessibility for data consumers in performing knowledge discovery and critical decision making surrounding the COVID-19 pandemic. The Evidence-based filtering is based on the Domain-specific Topic Model (DSTM)[19] to discover the relationships over words and tools/resources (e.g., drugs and genes) related to the COVID-19 pandemic.

2.2 | Middleware and Guided Interfaces for Science Gateways

Designing and developing a successful science gateway takes a significant amount of time, funding and personnel effort. However, science gateways have to continuously evolve to adapt to the changing needs in scientific research/education tasks. Leveraging advanced technologies can help science gateways to address this issue. One such technology involves use of microservices via RESTful APIs. For example, GenApp[20] leverages decoupling of application code from the science gateway to allow researchers to specify only the input and output parameters to run their command line applications via a graphical interface. Agave[21], a science-as-a-service API platform, was built largely with Docker container based microservices to seamlessly integrate API management, capacity scaling, and community contributions to provide platform services, science APIs and support services.

A number of science gateways are looking to integrate easily reusable and transferable building blocks. Apache Airavata[22] provides a software suite to compose, manage, execute, and monitor large-scale applications and workflows for science gateways. PaaSage[23] supports the design and deployment of multi-cloud applications by optimizing and customizing workflows in science gateways. Agave[21] offers platform-as-a-service for hybrid cloud computing and data management purposes and is being adopted

as the API layer by several science gateways, such as CyVerse[24]. Globus Galaxies[25] is a domain-independent, cloud-based science gateway providing a web-based interface for creating, executing, sharing, and reusing workflows composed of arbitrary applications, tools, and scripts. MiCADO[26], a microservice-based application orchestrator middleware, offers scalable Docker container-based microservice deployment by integrating services from federated private and public cloud resource providers.

Many of the existing science gateways provide comprehensive user services and convenient workflow automation; however, they largely lack the capability for users to customize programmable plug-ins (e.g., customized recommenders, ML tools for workflow processes, or dynamically configurable environments). The integration of plug-ins requires modular programming to extensively maintain and deploy heterogeneous components. Our middleware development is inspired by leading science gateways and frameworks that allow highly portable and reusable building blocks in next-generation science gateway development, as well as decoupling of system components to allow users to customize their workflows and computing tasks. We use the plug-in management middleware approach for integrating multiple plug-ins to hide the complexities of managing and deploying the plug-ins to the users.

Another novelty is in the use of intelligent agents to design and implement guided web interfaces for data providers/consumers to integrate, manage, execute, chain and scale plug-ins in science gateways. Intelligent agents in the form of chatbots have been popular in the development for human support in various applications[27]. The use of chatbot services has been demonstrated in fields such as customer service, banking, and healthcare[28,29,30,31]. The primary use behind intelligent agents is to conduct an online chat conversation via text or text-to-speech, in lieu of providing direct contact of a human agent. With such chatbot guided interfaces, users can be familiar with the system quickly and get instant instructions. On the other hand, the developers can get feedback from users with the content from users' input. This provides a great deal of convenience for both users and developers. Although the promise of using chatbots has been demonstrated in examples such as customer service, banking, healthcare, there has been limited work in using chatbot guided user interfaces in the context of science gateways.

Our work in this paper builds on early prior work on intelligent agents that showed their importance in the context of science gateways while enabling data consumers to keep up with the emerging resources/technologies in various domains, such as neuroscience and bioinformatics[32]. We also leverage work in[33], where the authors assessed the critical need to improve the adoption and diffusion among science gateway users by providing intelligent agents that are capable of using fuzzy logic for improving computational workflows for researchers. Lastly, the work in[34] leverages chatbot technology as an intricate part of a recommender system to guide researchers and educators, specifically in the neuroscience domain, to improve their pre-defined workflow through RESTful web services. This chatbot-assisted recommender system aimed to help researchers/educators with various levels of domain proficiency. Similarly, authors in[35] address the users' growing adoption of science gateways, which has led to the increase of more inexperienced users who are new to science gateways, by developing an ontology-based e-learning environment. This e-learning supports guided interfaces that reduce the knowledge gap of inexperienced scientists and helps them to effectively develop data-driven insights pertinent to their research workflows.

2.3 | Socio-Technical Factors for Innovation Diffusion

Socio-technical factors are important in the design of science gateways because they need user interfaces that provide community-level access to distributed computing resources as well as provide tools to help execute workflows and automate data integration in a sustainable fashion. Therefore, attention to why users adopt and how they interact with science gateways are important social science research questions to consider. Without good answers, science gateways would not have widespread implementation, and also may not be successful in ultimately increasing user satisfaction e.g., meeting data consumers' growing needs for a variety of scientific tasks, or data providers' requirement to customize scientific workflows. The theory on diffusion of innovations (DOI) theory offers relevant insights to find good answers. Originally proposed by Everett Rogers[5], DOI refers to the complex process by which new technologies or ideas are adopted on a societal level, and involves descriptions of the phases of adoption, which type of people fall into these phases, and what actions can be taken to best promote adoption.

Social scientists[36] have identified 10 attributes that promote user adoption of science gateways (and cyberinfrastructure broadly). First, user adoption is driven by needs, as users generally do not have the time to explore a science gateway that does not help them address a particular scientific need. Second, user adoption is driven by being able to find the science gateways listed at some popular repositories with organized access. Third, user adoption is driven by allowing potential adopters to try out the science gateway features without much initial investment of time. Fourth, user adoption (except for the pioneering users) is driven by the opportunity to observe what their peers find beneficial in a science gateway and how the peers obtained benefits in their productivity and collaborations. Fifth, user adoption is driven by the evidence that the science gateways are much

better than stand-alone command-line tools and techniques in accomplishing respective scientific tasks. Sixth, user adoption is driven by the simplicity of the science gateway, as users often do not have time to learn the inner workings of complex tools to handle big data management. Seventh, user adoption is driven by how the science gateways fit into their existing routines and the routines of their collaborators. Eighth, user adoption (especially the late majority who adopt after about half of the population has adopted) is driven by having a community behind the science gateways. Ninth, user adoption is driven by having very readable and updated online documentations of the science gateway capabilities. Lastly, user adoption is potentially also driven by a science gateway's ability to adapt from one scientific domain to another.

Furthermore, social scientists[37] have identified seven strategies to promote science gateways to potential users, based on different short-term and long-term goals. For short-term activities aiming at active promotion, the team behind a science gateway needs to raise awareness of the gateway with success stories, provide personalized demonstrations, train users (online and offline), and network with the scientific community. Long-term goals of stimulating organic adoption requires the team behind a science gateway to build relationships with trust, stimulate persuasion via word-of-mouth, provide state-of-the-art tools and distributed computing resources, as well as maintain sustainable documentation online. While adoption is the first step towards the spreading of an innovation, it does little good if it is not also *accepted and continuously being used* long-term. Even with science gateways facilitating the use of cyberinfrastructure, the technology remains complicated if it is primarily developed for advanced users. Unfortunately, this can limit the adoption and ultimate acceptance of the technology. To address this issue, authors in[37] explored various strategies informed by DOI and the technology acceptance model (TAM)[38], with the latter adding perceived ease of use—how simple an innovation is to implement—and perceived usefulness—how much an innovation can improve one's workflow—to the aforementioned DOI acceptance attributes. Other related works such as[39] documented that successful and sustainable projects tend to employ open-source software licenses in their science gateways, and they often put well-known best practices in the software engineering community into practice.

Given the above literature review, in order to promote meaningful adoption and implementation of next-generation science gateways, this work in the context of the exemplar OnTimeEvidence science gateway with the Vidura plug-in management middleware is guided by the aforementioned 10 attributes. Especially, our focus is on determining the user needs for such a next-generation science gateway through qualitative interviews similar to those conducted in market research, and how a chatbot (in this case, Vidura Advisor) can be designed to improve simplicity/reduce complexity, direct users to online documentations, and facilitate a robust community behind the gateways, to highlight a few strategies.

3 | BACKGROUND & MOTIVATION

In this section we introduce the OnTimeEvidence science gateway component requirements to provide the data pipeline process between data providers and data consumers. OnTimeEvidence implements a plug-in management middleware guided by an intelligent agent viz., Vidura Advisor, which allows data providers to customize and integrate multiple plug-ins and reduces the complexities of managing and deploying various features for data consumers, as illustrated in Figure 2. Thereby, OnTimeEvidence provides cloud-based data pipeline orchestration for COVID-19 evidence-based text (e.g., publications) and data (e.g., EHRs) analytics. OnTimeEvidence also helps data consumers to automatically filter high-quality publications for identifying tools, topics and other important criteria for feasible, pandemic-related solutions. Its primary function is to retrieve large quantities of scientific literature on a variety of healthcare related topics, with the important added capability of sorting this literature based on a hierarchy of increasingly rigorous study designs known as the Levels of Evidence pyramid[11,12]. With this capability, data consumers will be able to find relevant studies based on the best science, facilitating and speeding up the research process in an increasingly diluted field of study caused due to inundation of publications with varying levels of evidence.

Data providers can access diverse datasets and integrate them using the OnTimeEvidence science gateway via RESTful interfaces. A data pipeline orchestration layer manages the plug-ins to extract, process and fulfill the datasets from data providers to data consumers. This layer includes the plug-ins manager to process the data as well as enable data access and workflow manager customization to transform and share large-scale datasets. Once data from providers is available within the science gateway environment, it is extracted and processed using ML model-based plug-ins. The result sets generated by the extraction process are made available for data consumers via recommenders and analytic tool plug-ins. From the data consumers side, researchers' access to the data pipeline functions is managed by the data access and workflow process module, and is dependent on specific roles they will have access to datasets via an entitlement feature. Specifically, the data access and workflow process

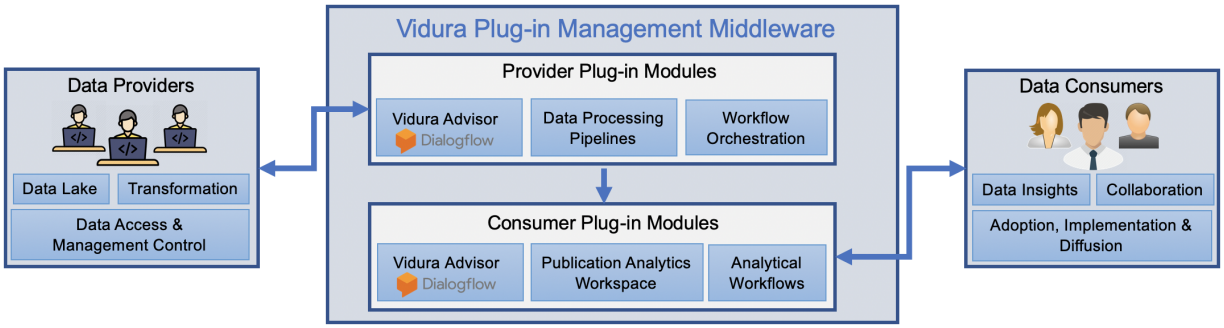

FIGURE 2 Overview of the Vidura plug-in management middleware components to help both data providers and data consumers.

management module uses the entitlement database to manage the users' profile information to conduct the role-based data and analytic resources entitlement access, and allows administrators to manage the systems resources and users.

In the following, we detail the provider and consumer side requirements relate them to the socio-technical requirements for promoting adoption of the OnTimeEvidence science gateway.

3.1 | Provider Side Requirements

The OnTimeEvidence science gateway can integrate multiple data sources via RESTful APIs or open SQL interfaces. It allows flexible and scalable integration of data sources related to healthcare literature. To achieve this step our science gateway needs to adopt best practices from healthcare data management standards, utilize state-of-the-art healthcare data processing pipelines solutions, and guarantee an efficient data workflow between providers and consumers via a guided orchestration process. In the following, we describe current solutions for each stage of the process, and indicate how our approach leverages those solutions.

Healthcare Data Management: Healthcare data providers have specific requirements to store and share healthcare data, keep those records secure, provide analytic services related to big health data, and preserve data privacy. In the context of data accessibility, secure options using Blockchain-based[40] secure storage layer have been proposed to bring the level of security required to store EHRs in shared or cloud environments. Role-based access control (RBAC)[41] approaches are used to share data among multiple data consumers (i.e., physicians, researchers, government institutions, private organizations). Such access mechanisms are needed in place to facilitate the data sharing process while ensuring data confidentiality and integrity[42]. RBAC restricts platform users' permissions to their roles and only permits users access to have privileges that they absolutely need to perform their job functions. For example, healthcare students of an organization should not have access to digital financial records of patients. The OnTimeEvidence science gateway needs to ensure the security and privacy standards required by the data providers by having in place the role-based and entitlement feature mechanisms to process and fulfill the data related actions.

Healthcare Data Processing Pipelines: Multiple approaches are in place to provide framework-based solutions for the deployment of healthcare related data processing pipelines. Distributed cloud-based framework are used to enable researchers to process and train their ML models[43], perform effective analytic tasks on complex phenotype datasets[44], or resolve the critical concern of optimizing supply-demand needs[45]. Recent solutions are increasingly leveraging ML and data processing pipelines to effectively manage the complexity and size of healthcare datasets. In a recent work related to the COVID-19 pandemic, authors in[46] resolve the need to handle increasing rate of COVID-19 related data through a data-driven model deployed on a cloud-based platform that predicts the growth of the pandemic. Authors in[47] develop an intelligent framework of emerging ML-based technologies for helping with the COVID-19 pandemic response. Their work suggests that these disruptive technologies can be integrated in smart devices using cloud platforms. Authors in[48] seek to resolve the image segmentation problem in COVID-19 chest X-rays by developing a novel ML framework that utilizes slime mold and whale optimization algorithms. This problem involves a threshold mechanism that builds a binomial classification to determine whether a patient has the COVID-19 virus. Similarly, the authors in[49] address the issue of providing accurate classification of COVID-19 in CT scans by developing a deep learning architecture that leverages a semi-supervised few-shot segmentation algorithm for image segmentation.

The OnTimeEvidence science gateway needs to adopt best practices in the above approaches by e.g., including a ML-based recommender as a plug-in module to extract and process source datasets and put place the role-based access control mechanisms to retrieve and deliver the required data for knowledge discovery of data consumers.

3.2 | Consumer Side Requirements

Given the time-consuming nature and complexity of the processes used by data consumers (e.g., clinicians, researchers) to search, process and analyze healthcare literature, augmentation through consumer-side plug-in tools that help automate and scale the workflows are valuable. Such tools can greatly lessen the burden of data consumers to handle large collections of publications to be analyzed for a research task, and provide new insights for knowledge discovery based on evidence-based information filters to take decisive actions. For related publication analytics, the following requirements are outlined to meet the data consumers need to improve the efficiency and effectiveness of the workflow tasks in the OnTimeEvidence science gateway.

Data Analytics Toolkits and Workspaces: A plug-in management approach at the application level is an effective way for data consumers to have accessible analytics tools and workspaces for performing effective analysis over publication archives. First, such data analytics toolkits must provide a wide capability for data consumers to perform a variety of computational tasks. Data consumers must be provided general information on such toolkits and information on the most feasible analytics tools to use depending on their search query. Second, the analytics workspace must provide an intuitive interface for data consumers to immediately perform their analysis over publication archives. The OnTimeEvidence science gateway should be capable of supporting a guided user interface to navigate users in discovering insights through a step-by-step approach.

Data Source and Timeline Management: The system should provide multiple options and allow data consumers to select the sources of articles (i.e., CDC, NIH, medRxiv) that are relevant for their research. Besides the source of the publications, data consumers should be able to narrow the search by indicating the publication date window where the search will take place. Consumers should be able to get guidance about which articles pertain to their areas of research. In this way, science gateway data consumers will have the ability to manage a large number of short-term new publications with content relevant to the novel COVID-19 or EHRs. They should also be able to correlate findings with longer-term publications with content related to previous coronavirus studies (e.g., MERS, SARS) that are related epidemics in terms of infectious diseases.

Multiple Search and Drill-Down Options: The system should allow data consumers to leverage recommender modules according to evidence-based practice standards. OnTimeEvidence should allow support for evidence based-filtering and social filtering. Evidence-based filtering provides ML-based analytics to guide users in narrowing down their articles list by selecting topics pertinent to their clinical queries (by applying a topic model), focusing on a specific clinical category (by applying a category model), or by running predefined queries on processed data. Social filtering helps the users to leverage the models from the evidence-based filtering approach to search and filter results for important clinical questions. The recommender modules should be integrated as plug-ins to support the dynamic needs of data consumers. In an effort to keep science gateway consumers informed on the advanced capabilities of the recommender modules, the consumers must be provided assistance and guidance on how these algorithms process literature and provide critical insights across the publication dataset. In addition, explanations should be provided on how these insights can foster effective collaboration with other researchers dealing with similar research objectives.

User-intent based Analysis/Visualization: Having a good list of properly filtered articles from multiple reliable sources does not fully alleviate the challenges data consumers face in attaining their goals. In addition, useful analytic tools need to be provided to help them make topic associations, find similarities, and visualize and share results. By meet this requirement, OnTimeEvidence should allow data consumers to: (a) define and manage the size of the result-set required for analysis, (b) present the result in a tabular structure including multiple parameters (e.g., title of the article, source, publication date, Level of Evidence, clinical category), and (c) let them control the content's presentation by using actions such as show/hide parameters, and sort or filter records based on the information of parameters being selected. Consumers should be provided information on how to conveniently access information about articles such as the abstract section content, the entire document retrieval from the source, or the ability to tag documents as favorites for easy future reference. Analytics tool-sets should provide the functionality to organize and group the articles based on types of design study (e.g., systematic reviews, meta-analysis), clinical

category, and subsequently allow them to run correlations on those subsets and mine knowledge patterns related to previous such pandemics (i.e., 2012-13 MERS, 2002-03 SARS) to find similarities. Lastly, a dashboard functionality should be provided to display the results using multiple data visualization artifacts (e.g., tables, charts) for easy reading, interpretation, findings sharing, and collaboration.

Collaboration Tools to Crowdsource Expertise: Effective collaboration tools are needed to help data consumers to share their findings, correlate their conclusions with results from similar works, discuss topics with colleagues and collaborate with broader projects. To help data consumers discern useful information from the vast and fast-growing articles in journal archives, evidence-based filtering and social filtering using ML-based analytics and expert collaborations through a social network plug-in are important to relieve data consumers from the burden of manual search processes. Thus, data consumers can better concentrate on analyzing the latest and relevant content, sharing their insightful findings, providing critical feedback, and producing new knowledge to combat the pandemic response issues.

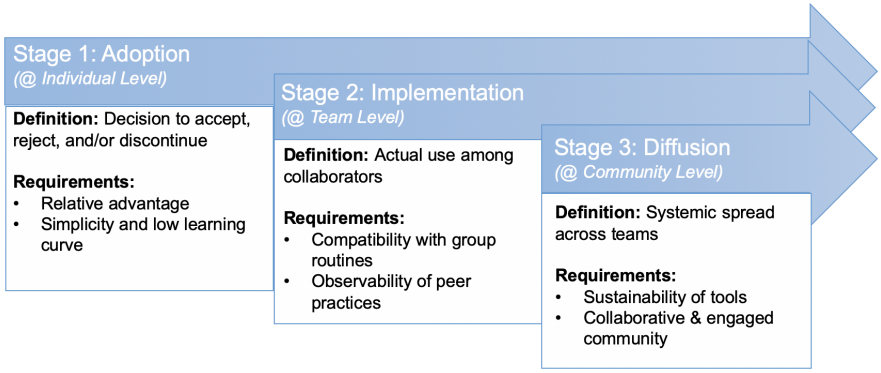

FIGURE 3 Three stages of science gateway development that consider the socio-technical aspects of adoption at the individual level, implementation at the team level and diffusion at the community level.

3.3 | Understanding Requirements from a Socio-Technical Perspective

As discussed earlier, science gateways meaningful impact can be limited when there is little to no widespread implementation and diffusion across the user community. Thus, the needs of data consumers must be addressed to fully understand how science gateways can be adopted. To better understand the growing need of data consumers using the OnTimeEvidence science gateway, we must outline the data consumer requirements in the socio-technical stages of adoption at the individual level, implementation at the team level, as well as diffusion at the community level. Figure 3 describes the stages required for data consumers to acquire and maintain the use of OnTimeEvidence and its functional needs for both data consumers and data providers:

Adoption: At the individual level, this refers to the decision to accept, reject and/or discontinue using the science gateway. The application being offered needs to have a relative advantage over other existing science gateway applications. This can include leveraging the latest technologies, accessible datasets and a variety of analytics tools for discovering data-driven solutions. Data consumers must also find the user interface intuitive, such as point and click operations, or guided visual cues. Thus, the OnTimeEvidence science gateway must handle such requirements to maintain the trust and satisfaction of the application users.

Implementation: Implementation at the team level is defined as the actual use of the science gateway among a team of collaborators. When individual consumers adopt such technologies, this allows consumers to connect with others who are in a similar field for trustworthy collaborations. In this context, data consumers require the science gateway to be compatible with group routines. In addition, these collaborators must require peer observation among their colleagues in order to develop their practice and or methods when using the science gateway application. Thus, the OnTimeEvidence science gateway must manifest its capability to allow a team of data consumers to collaborate through the application in order to accomplish their research objectives.

Diffusion: Diffusion at the community level is the systematic spread of science gateway across teams. As teams are being formed, a community of data consumers must continue to promote quality research and further innovative thinking through the spread and use of the science gateway. Data consumers require the sustainability of tools needed to make bold scientific advancements within the consumer community. Furthermore, in order to maintain the systematic spread among teams, the community must be active and highly engaged. A science gateway such as OnTimeEvidence must ensure the growth of the community to sustain its widespread adoption.

In the context of this study, we focus our attention on the first stage i.e., adoption from an individual standpoint. Our proposed OnTimeEvidence with Vidura plug-in management middleware assesses the needs for both data providers and data consumers from a socio-technical perspective, and focuses on the needs of data consumers to effectively adopt the OnTimeEvidence in their healthcare data analysis workflows.

4 | ONTIMEEVIDENCE SCIENCE GATEWAY

In this section, we present development details of the OnTimeEvidence science gateway that leverages the Vidura plug-in management middleware. We first detail the analytics components (e.g., ML modules, knowledge bases, intelligent interfaces) for large-scale data management in OnTimeEvidence. Following this, we describe how the Vidura plug-in management middleware is integrated within OnTimeEvidence at the infrastructure layer for data providers, and at the application layer for data consumers. In this context, we also detail how the Vidura plug-in management middleware can be generally customized within other science gateways to serve scientific domains broadly.

4.1 | Analytics Components for Large-scale Data Management

OnTimeEvidence as shown in Figure 4 augments the practices for science gateway users for both the data providers (e.g., administrators, developers) and data consumers (e.g., clinicians, researchers), and provides guided interfaces to accomplish their respective tasks. By enabling an end-to-end process, OnTimeEvidence allows data consumers to customize data analytics workflows, in which data consumers access COVID-19 literature in their scientific research tasks. In clinical fields, researchers commonly follow a systematic literature review procedure known as the evidence-based practice[50]. Data consumers commonly adopt this method for synthesizing and reviewing articles based on the inherent evidence levels that are pertinent to their research.

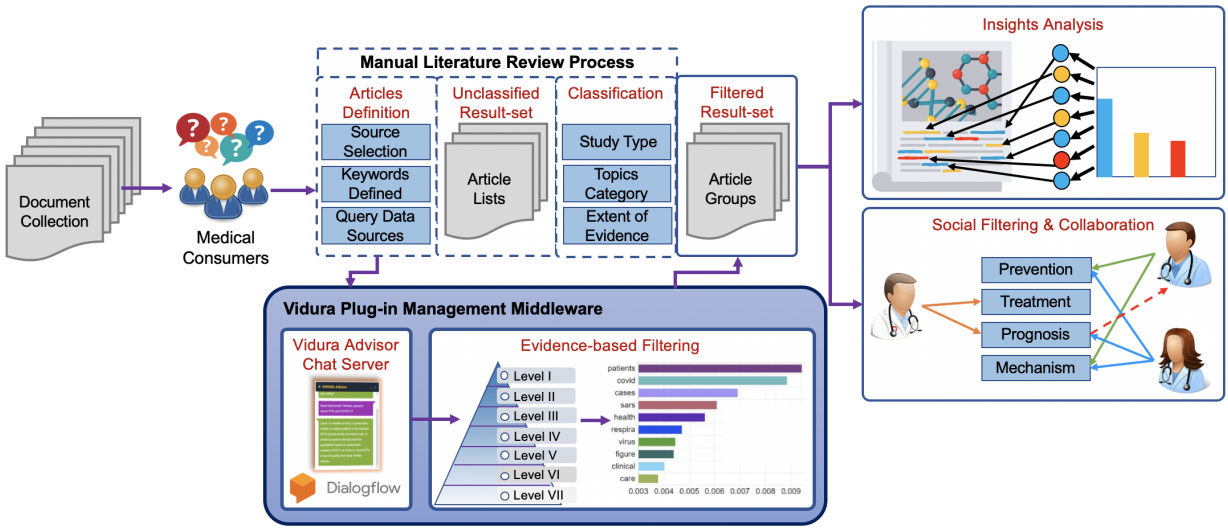

FIGURE 4 The end-to-end process of extracting evidence-based information from users' inputs and providing analytics/visualization for obtaining insights and sharing them on a social network platform.

Specifically, OnTimeEvidence implements a hierarchical evidence-based framework, Levels of Evidence pyramid[1,2] that can filter out publications with high quality information (e.g., background information to systematic reviews) versus publications that have low evidence levels such as opinions or focused case studies.

To simplify literature data selection and analysis for COVID-19 researchers, the components in the OnTimeEvidence cater to data consumers' needs by generating indicative responses to clinical queries that allows them to reduce the manual steps in their scientific workflows. The implemented tasks performed include: (i) a literature selection form that allows a user to query search terms related to the Levels of Evidence and other clinical information related to COVID-19 (e.g., *What is the status of Nucleic Acid Amplification tests with PCR used for COVID-19 and SARS-CoV-2? What is the status of serology tests with Immunoglobulin M (IgM) and Immunoglobulin G (IgG) used for COVID-19?*), (ii) the functionality to process the literature selection, and in response, generation of a new workspace (e.g., Jupyter notebook) with plug-in capable tools needed to discover novel insights related to COVID-19 clinical and bedside care studies, and (iii) use the workspace plug-in to allow the user to conduct a publication and/or collaborative analysis and store the results for sharing them via a social network platform. The above OnTimeEvidence component for literature selection and analysis uses a relational data structure and process to upload and store the metadata related to the COVID-19 literature. For the purposes of this work, we have collected over 10,000 publication records from the Kaggle COVID-19 Open Research Dataset (CORD-19)[51].

We detail the AI-based plug-in modules that have been integrated within OnTimeEvidence that include: Evidence-based filtering[18], Social filtering, and Vidura Advisor[34].

Evidence-based Filtering: This AI-based module integrates a probabilistic topic modeling algorithm viz., Domain-specific Topic Model (DSTM)[19,52] that uses domain-specific resources (e.g., drugs/genes) to discover hidden topic distributions within COVID-19 scientific literature. This is an extension of the Latent Dirichlet Allocation (LDA) model[53] that represents a topic as a multinomial distribution over a set of pre-defined vocabularies, and a document as a multinomial distribution over topics. DSTM improves the LDA algorithm by parameterizing drug and gene terms to find relevant COVID-19 research topics and trending words based on the topic distributions. DSTM in OnTimeEvidence automatically learns the latent patterns within the CORD-19 dataset. First, users input of a collection of COVID-19 articles from the CORD-19 dataset[51], a knowledge base of drug terms from the *COVID-19 Vaccine Tracker*[54], and a knowledge base of gene terms from the *Virtual Incident Procurement (ViPR)*[55] to train the DSTM. Then after training, the DSTM can help analyze/visualize the most popular research topics in the collection of publications. The DSTM will then rank the most commonly investigated drug or gene terms based on each topic. The DSTM enables data consumers to query relevant COVID-19 drugs and genes based on their research topics.

Social Filtering: Social filtering is an AI-based module that implements an "user-category" recommendation approach, motivated by[56]. In this approach, data consumers subscribe to a social network to receive notifications within their feeds based on the relevant findings or research interest tags similar of data consumers. The user-category recommendation takes an input rating of papers based on publication tagging to suggest data consumer collaborations with other expert data consumers with similar interests. The Social Filtering module is implemented via *HumHub*, which is a purpose-driven, open-source social network that allows data consumers to publish content and connect to other subscribers. The social plane provides an organized method of: (i) helping data consumers to obtain useful resources, and further enables them to perform expertise sharing, and (ii) allows the data consumer community to crowdsource the effort to analyze/visualize information to help answer clinical research questions.

Vidura Advisor: The Vidura Advisor plug-in agent which guides data consumers to navigate the OnTimeEvidence features is implemented through Google Dialogflow, which is a natural language understanding (NLU) platform. Such an implementation makes it easy to design and integrate the Vidura Advisor intelligent agent user interface into mobile apps, web applications, smart devices and interactive voice response systems[57]. The Vidura Advisor learns the data consumer intentions (e.g., user queries) through a pre-trained NLU model that uses *intents* and *entities* to generate the replies for users' inputs. The intents are meant to serve as questions topics. When users input the questions, 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 asked for some replies which are not included in the intents, Vidura Advisor has fallback intent outputs[58] to handle such cases.

4.2 | OnTimeEvidence System Components

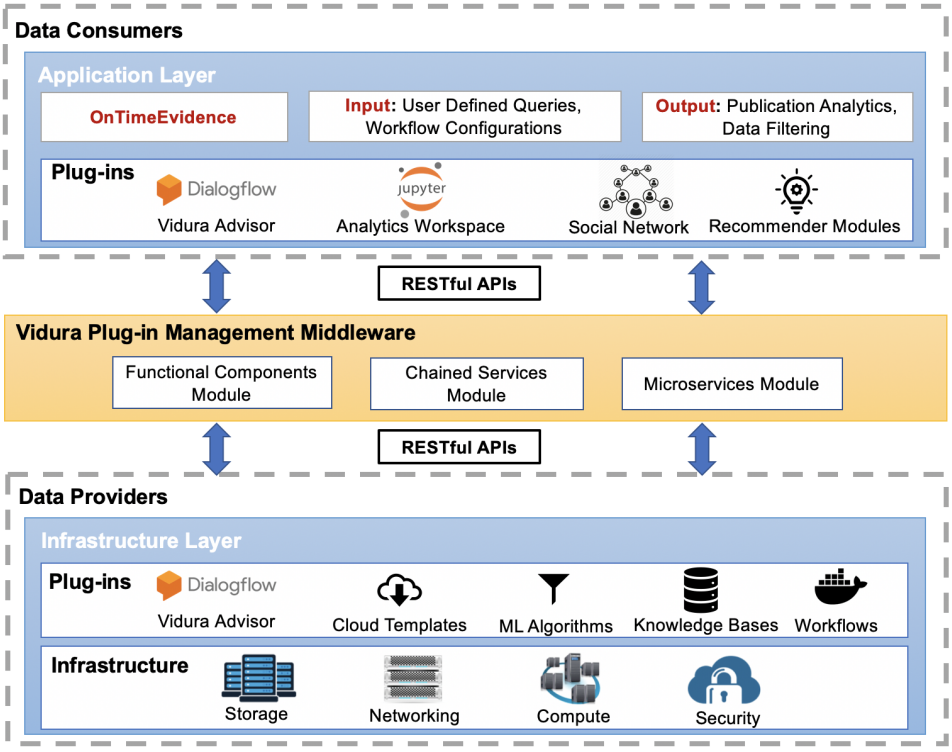

FIGURE 5 Multi-layered Plug-ins Management Design.

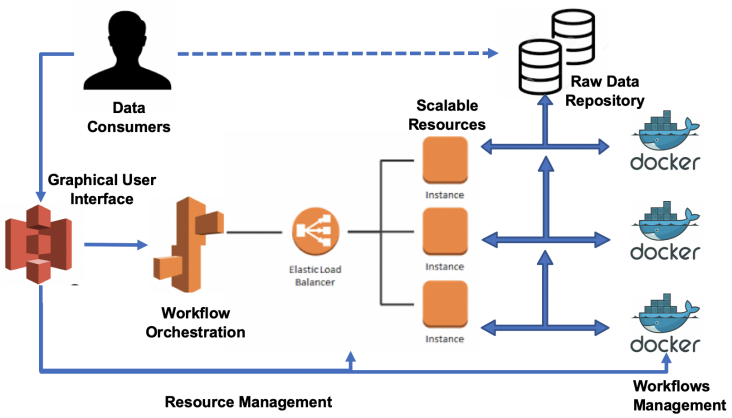

FIGURE 6 Workflow orchestration illustration and associated components within the OnTimeEvidence science gateway.

We present our plug-in management middleware development details that leverages a intelligent agent i.e., Vidura Advisor, which augments the practices for science gateway users for both the data providers and consumers, and provides guided interfaces to accomplish their respective tasks.

As shown in Figure 5, the middleware components provide navigational support for data management and executing computational workflows for the OnTimeEvidence science gateways users. The Vidura Advisor provides guided interfaces to the dynamic plug-in technologies at both the infrastructure level for data providers and at the application level for data consumers. These middleware components help in handling the management, orchestration, and rendering/execution/monitoring of plug-ins as microservices that are integrated in the OnTimeEvidence science gateway. In the following, we present and detail its three layers i.e., Infrastructure Layer, Vidura Plug-in Management Middleware Layer, and Application Layer, which are interfaced

through RESTful APIs.

4.2.1 | Infrastructure Layer for Data Providers

The Infrastructure Layer primarily supports data provider roles, i.e., administrators and developers. Data providers can rely on the middleware capabilities to supply computing, storage, and networking resources to support the needs of science gateway application users. The needs include collecting datasets from disparate sources, performing transformation of structured and unstructured data through data lakes or data warehouses, as well as providing cloud and access control mechanisms for authorizing privileges to data consumers at the application level. Data providers will have the capability to interact directly with the management middleware through various plug-ins (e.g., cloud templates, ML models, knowledge bases). Data providers can also transform data processing pipelines and associated data into workflows, which can then be offered as plug-in pipelines in the workflow orchestration tasks.

Workflow Customization: Provisioning of the data for analytics and custom pipelines are enabled by our middleware using containers and remote repositories for management of computation resources. Our plug-ins interface with distributed resources and control them in a centralized manner. New analytics pipelines and containerized workflows can be plugged into OnTimeEvidence platform with ease and orchestrated using our implemented orchestration plug-in mechanism. As shown in Figure 6, the containerized workflows and remote data repositories can be pulled into any of the worker nodes and executed on demand by the users interacting with the OnTimeEvidence graphical user interface (GUI). The GUI is built to be customizable and can act as host interface for many pluggable workflows. The GUI is interactive and captures all users requirement and workflow orchestration demands for further enabling compute resource allocations via a pluggable recommender service. A RESTful web service enabled back-end server further improves the customizability of plug-ins and workflow orchestration within the platform.

Analytics Workspace: Once healthcare related literature datasets are available within the OnTimeEvidence science gateway, researchers, who have been entitled to access these datasets and use analytic workspace resources (i.e., Jupyter notebooks), can request these resources and make use of the required data. In this way, users can have access to customized analytic workflows related to specialized literature such as published COVID-19 articles. Using the plug-in services such as the Vidura Advisor and the Level of Evidence literature search framework, researchers can search and filter documents and successfully extract relevant literature pertinent to their research. For instance, by using the Vidura Advisor plug-in, researchers can leverage the OnTimeEvidence features to reduce the burden of manual searches by: (i) filling a literature selection form that allows a user to query search terms related to the Levels of Evidence, (ii) making the selection of articles from the dataset generated, and (iii) using an analytic workspace (e.g., JupyterLab environment) to conduct a publication and/or collaborative analysis and store the results for sharing via a social network platform.

4.2.2 | Chatbot-assisted Plug-in Management Middleware

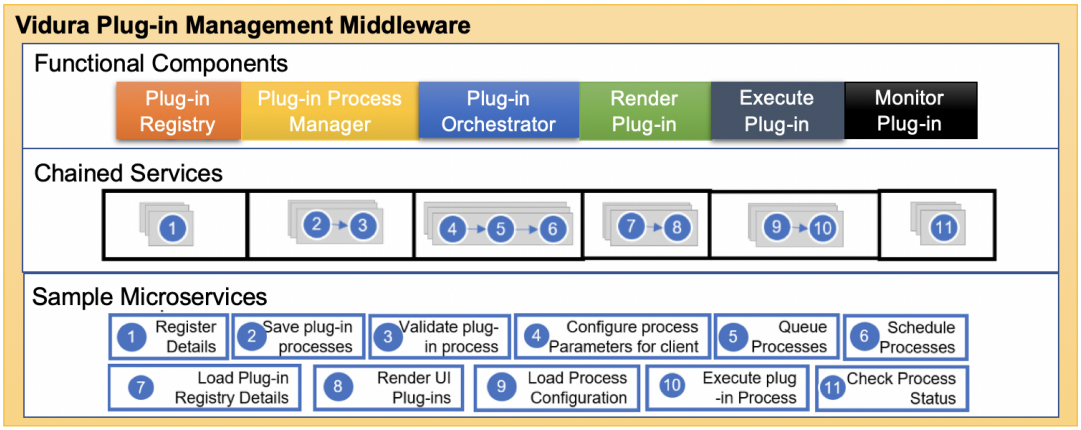

FIGURE 7 Multi-layered Vidura Plug-ins Management Design.

We implement the Vidura plug-in management middleware as an end-to-end framework that provides data consumers plug-in management capabilities as shown in Figure 7. The Vidura plug-in management middleware purpose is to integrate, manage and execute different plug-ins using functional components implemented as microservices. Once integrated, augmented interfaces are provided to end-users (e.g., data consumers) when using the domain-specific gateways to access and use the capabilities provided by the plug-ins. We can see in Figure 7 how several application-specific or infrastructure-specific microservices can be chained to customize capabilities and allow information sharing between the different microservices to execute specific functional components. In the following, we detail the three components of the Vidura plug-in management middleware: Functional Components, Chained Services and Sample Microservices.

Functional Components: Herein, we detail the functionality with the following plug-in components:

- *Plug-in Registry* is the repository for all clients and plug-ins related data. It also includes the metadata of plug-ins–which is the configuration to execute the processes related to the client's plug-in selection–as well as the science gateway application details (i.e., OnTimeEvidence). This metadata is formatted using JSON and stored in the database when add or update actions are taken on plug-ins/processes. Our current implementation of this registry uses the Hibernate framework for object-relational mapping over a relational database i.e., a MySQL backend.

- *Plug-in Process Manager* is used to create plug-in processes for a next-generation programmable science gateway such as our OnTimeEvidence. Plug-in management and execution can be broken down into multiple processes. We implemented microservices that help data consumers to configure and execute processes for specific plug-ins. This component is also responsible as a client to the Plug-in Registry for persistently storing all process details, metadata information related to each process of plug-ins.

- *Plug-in Orchestrator* abstracts the configuration of the middleware queuing layer for individual clients. Specifically, it configures execution parameters for different processes of plug-ins, and consequently queues all processes of plug-ins. This component is also responsible as a client to the Plug-in Registry for persistently storing all parameter configurations and queuing requests for processes in a queue.

- *Render, Execute and Monitor Plug-in* provides the web-based user interface to the application layer using a client software development kit that we implemented. The Execute plug-ins component executes different processes for specific plug-ins based on data consumer requests, and the Monitor plug-ins component checks the execution status of the plug-ins.

Chained Services: In science gateways such as our OnTimeEvidence, data consumers have different application needs in terms of certain features for their scientific workflows. To handle such a diverse need while maintaining high performance of the Vidura plug-in management middleware, we have implemented chained services that use the output from plug-ins executed as microservices from the Functional Components layer. This provides data consumers customizable features on the OnTimeEvidence science gateway through the support of the Vidura Advisor plug-in at the application level. The chained services handle the dynamic end-user need of programmable features and allow data consumers to specifically choose which application features are needed for their scientific workflows through the assistance of the Vidura Advisor plug-in.

As the data processing tools and workspaces are managed through the Functional Components layer, data consumers can configure their workflows by selecting which recommender modules will be pertinent for performing knowledge discovery over a COVID-19 dataset. For example, if a data consumer were to use OnTimeEvidence with the intent to find data-driven insights over a chosen publication archive, the data consumer can customize their application to include evidence-based filtering (as mentioned previously in Section 3) by interacting with the Vidura Advisor plug-in. The data consumer can interact with the Vidura Advisor plug-in at the application level by asking questions and receiving answers through its guided interface. The Vidura Advisor plug-in then recommends what specific services (in this case evidence-based filtering) the consumer should use. By leveraging the chatbot-assisted plug-in, data consumers can thus have more guidance on which application plug-ins are feasible for executing their workflows to accomplish their tasks.

Sample Microservices We have implemented microservices that follow the standard practices of RESTful API design using the Java programming language. All microservices have been developed in the back-end using Spring Boot, 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. In addition, we used 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.

The microservice architecture involves enabling flexible interactions between multiple services. Each instance of a service exposes a remote RESTful API at a particular location (host and port), and the number of service instances as well as their locations change dynamically. In this case, a combination of service registry and client-side service discovery allow services to find and communicate with each other without hard-coding the host names and ports. The service registry handles details of services such as their instances and locations. Service instances are registered with the service registry on startup and are de-registered on shutdown. We have implemented microservices for both service registry and discovery client using Spring cloud[59] and Eureka. We have also developed a gateway edge service using Spring cloud and Zuul to enable dynamic routing in our middleware.

In addition, we have implemented the OAuth 2.0 security protocol with Spring Boot and Spring Security to provide authentication support to science gateway clients. It enables third-party applications (e.g., GitHub API, Google API) to obtain limited access to web applications. This allows for science gateway data providers to enable access control to plug-in services they want to provision. With the integration of OAuth 2.0, we validate users by allowing them to sign-on to the web application with necessary authorized permissions.

The implementation of our Vidura plug-in management middleware allows data providers to integrate and monitor the use of plug-ins that can be generalized across science gateway applications. Some of the benefits of generalizing our middleware are as follows: (i) it enables data providers to modularize the plug-in services used to develop microservices in science gateway applications such as our OnTimeEvidence. These design benefits can support plug-ins to independently operate by decoupling processes into microservices, (ii) it supports customizable design and deployment to augment scientific workflows; the architecture uses Docker containers to construct deployment patterns across the distributed resources to optimize the pre-defined workflows commonly used in a domain science community, and (iii) it provides the flexibility for disparate code bases to be integrated through microservices; Data providers typically create microservices using their preferred coding language and our architecture allows data providers to use their preferred coding language for creating customizable science gateways.

The process of using our plug-in management middleware is broadly applicable to any intelligent agent implementations that are driven by e.g., recommender modules, knowledge bases, and Jupyter Notebooks within various scientific domains such as neuroscience and bioinformatics[60,61]. In the context of this study, we apply our Vidura plug-in management middleware to OnTimeEvidence for the COVID-19 domain. The knowledge bases we have included involve the collection of documents from the COVID-19 Open Research Dataset (CORD-19) as well as drug/gene terms that have been collected from *COVID-19 Vaccine Tracker*[54] and *Virtual Incident Procurement (ViPR)*[55]. The recommender modules include the Evidence-based Filtering using the Domain-Specific Topic Model (DSTM)[19] and Social Filtering[56] adopted from our prior work[18]. Figure 8 illustrates the process to configure middleware components to add and execute the different microservices. The steps for a science gateway data provider to customize plug-in management modules for a given science gateway are as follows:

- **Step-1**: register different microservices to be used on the science gateway using the Vidura-guided plug-in interface. This interface uses Plug-in Registry services to register information related to the microservices for the science gateway.
- **Step-2:** register scientific plug-in processes such as data collection/processing parameters, and knowledge base, such that all the necessary information is provided to execute relevant microservices in the science gateway using the 'Plug-in Workflow Manager'.
- **Step-3**: add the science gateway (i.e., OnTimeEvidence in this work) client and link microservice modules with the client to allow users to use microservices on the science gateway interface.
- **Step-4**: configure parameters of plug-in processes and queue processes of the microservices for the science gateway client.
- **Step-5**: execute all plug-in processes using provided configurations and publish the microservice outputs to end users on the science gateway interface.

4.2.3 | Application Layer for Data Consumers

The Application Layer assumes data consumer roles of healthcare professionals (e.g., clinicians and researchers). Its implementation in OnTimeEvidence features a user interface where data analytics pipelines (e.g., analytics workspace, recommender

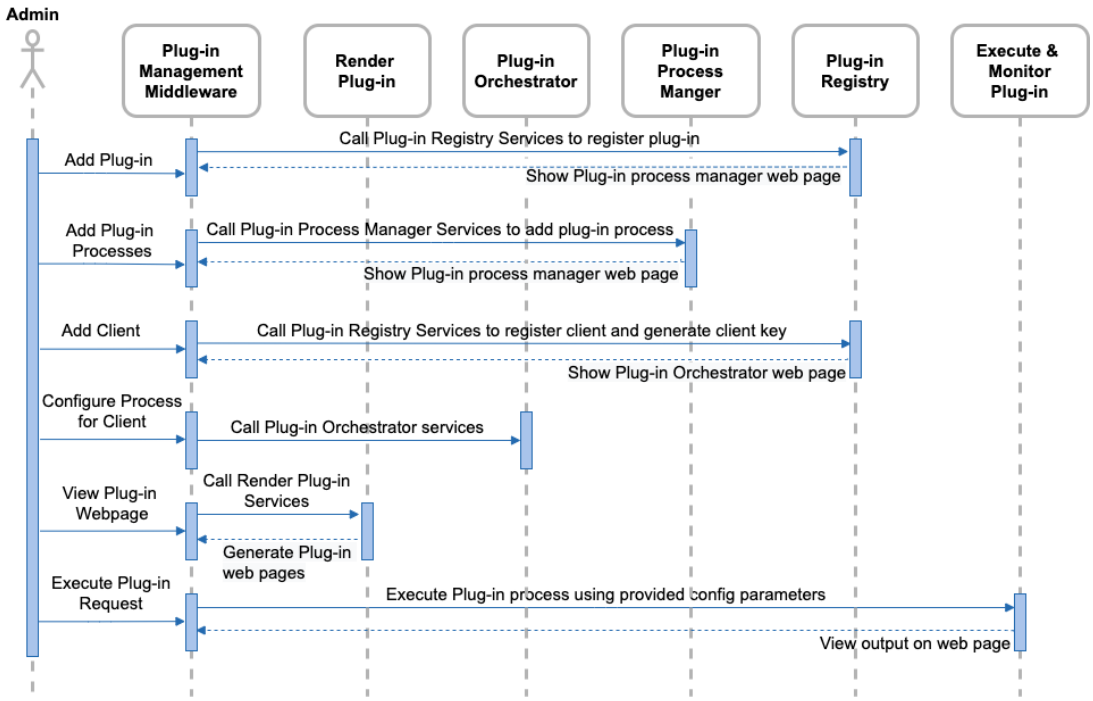

FIGURE 8 Sequence diagram showing steps to Add, Manage, Execute Plug-ins customized for a science gateway using the plug-in management middleware component interactions.

modules, social network) are implemented as plugs-in. These plug-ins enable the data consumer to dynamically integrate analytics tools and guided workspaces directly onto OnTimeEvidence. To interact with OnTimeEvidence application for finding insights in COVID-19 related text data (e.g., publications), data consumers can define queries to filter the set of articles they want to retrieve based on our pre-trained recommender module plug-ins (e.g., Evidence-based Filtering and Social Filtering). The data consumer can also set up their workflow configurations to have access to specific or multiple data sets such as COVID-19 publication datasets and EHRs. The output of the OnTimeEvidence (Figure 4) shows a set of results from the analysis conducted by the data consumer which includes publication analytics of COVID-19 literature and the filtering of specific data based on the defined query. Data consumers can then leverage this information to collaborate with other consumers using the social network plug-in as a separate application platform in which they can diffuse their insights and/or clinical queries. In the social science perspective, this meets the needs of data consumers via plug-in capabilities to engender widespread adoption from easy to use and configurable analytics tools/resources as well as collaboration at a team level (implementation) through a social network.

5 | USABILITY STUDY

To demonstrate the utility of the OnTimeEvidence science gateway that integrates the Vidura plug-in management middleware, we conduct a usability study on the COVID-19 literature workflow process and use the Levels of Evidence pyramid methodology[1,2]. Specifically, we conduct a qualitative interviews with 10 participants to assess their possible adoption of the OnTimeEvidence science gateway equipped with the Vidura plug-in management middleware. We report below both the social science methodologies and the findings from the user interviews.

5.1 | Interviewing Methodology and Thematic Analysis Techniques

Qualitative interviewing is an appropriate social science methodology for exploring emerging ideas, to seek participants' input with open-ended questions[62,63]. In this study, we employed snowball methodology[64] for participant recruitment. Based on the recommendations of a clinician associated with a midwestern university, four participants were recruited during the first wave. Then, based on the recommendations of the first wave, six more participants were recruited for the second wave. Collectively, we conducted a series of 10 qualitative interviews during Spring of 2021 with clinicians, medical researchers, biochemists,

veterinarians, and medical librarians. Among the participants, three are females and seven are males. The longest interview was 58:34 minutes and the shortest was 19:55 minutes, averaging around 32:42 minutes.

All interviews were conducted on Zoom. The interview questions asked participants to describe their current literature search practices, tools used, ideal features of databases, etc. Each participant was shown a short video regarding our OnTimeEvidence science gateway. The video went into a brief description of the application and addressed the challenge of manual publications analysis over emerging COVID-19 publications. Considering our target participants who were data consumers, the video did not go into detail on the specifics of Vidura plug-in middleware and its features that are more relevant to data providers. However, the participants received the context of Vidura Advisor and its capabilities with respect to the OnTimeEvidence science gateway. After showing them a short video of the developed capabilities as illustrated in Figure 4, we asked for their impressions of OnTimeEvidence and the Vidura Advisor. The interviews were transcribed verbatim using an AI-powered transcription service, Otter.ai. Then transcripts were systematically analyzed using the grounded theory methodology[63] for thematic analysis, a technique commonly performed in the social sciences. Although the findings are considered preliminary, what these interview participants shared reviewed how our prototype of OnTimeEvidence and Vidura Advisor may provide solutions to some of the challenges clinicians and medical researchers persistently face today.

5.1.1 | Preliminary Findings from the Interviews

We report the findings in terms of three sections of participants' impressions of OnTimeEvidence, usefulness of the Levels of Evidence pyramid, and the potential impact of the Vidura Advisor based guided interfaces.

Impressions of OnTimeEvidence: The 10 participants and medical researchers interviewed all had favorable first impressions of OnTimeEvidence. Many expressed that a problem with current research on COVID-19 is that the rate at which research is being published is much faster than any one person could possibly keep up with. Further, relying on measures of quality such as impact factors of journals and citation counts of articles are particularly limiting with the newest research (due to there being inadequate time for the impact factors and citation counts to accumulate). Additionally, many relevant studies have been pre-published in terms of pre-prints due to the urgency of disseminating findings about COVID-19. These pre-prints did not have time to undergo rigorous peer reviews, thus making all available articles a mix of quality.

Sometimes, researchers have difficulty finding studies with the strongest evidence in the more reputable journals. One interview participant stated that "the journal is becoming less and less important at this stage in time. In fact, the community realizes that [some high impact journals] put things out there that are maybe the hot topic and trending thing, but often don't have all the...actual data that you need" [P08]. This participant and others noted that expert opinions are not always the best evidence. Another participant noted that the main reason for his enthusiasm for OnTimeEvidence is its method of using the Levels of Evidence pyramid (discussed in the next section). This participant [P05] indicated that if OnTimeEvidence can filter and screen for randomized, double-blind, placebo controlled, multi-center trials, findings from these articles will be more compelling than under-powered studies, observational studies, and expert opinions.

Usefulness of the Levels of Evidence Pyramid: OnTimeEvidence addresses this concern by utilizing the Levels of Evidence pyramid in order to further sort articles not only by topic and date, but also by how rigorous the research evidence is in a given article. This is especially important with articles on COVID-19 because, as mentioned above, many articles on the subject are pre-prints, which lack rigorous peer reviews. Having ML models detect the quality of research in each article using the levels of evidence pyramid regardless of the source of publication or the number of citations they have garnered circumvents the problem raised by many of the participants interviewed. Seeing this, those who were interviewed expressed a great deal of excitement about the capabilities of OnTimeEvidence.

Many spoke of the amount of time this could save them, as well as addressing their concerns that they were missing important information in their traditional approach to literature searches. One participant expressed how useful this would be in reducing time spent combing through articles:

> *I think what you're describing is that this system is...reading the paper for me. Exactly that which I said is going to take forever. And I have to do it by hand. That's what this is doing for me… [And] it's actually measuring the evidence that way, what I would be doing by hand, that's what it's doing in each of these articles. Because that's great.* [P09]

What this participant expressed is a concern that most participants interviewed echoed, which was how time-consuming literature searches can be. By doing much of the busywork needed to find relevant and quality articles with the highest level of

evidence, OnTimeEvidence can save valuable time that clinicians and researchers absolutely need in order to most effectively do their job, including treating patients and save lives with the reliable research evidence in the available literature.

Helpfulness of the Vidura Advisor: In order to further facilitate the use of OnTimeEvidence, the capabilities of the chatbot i.e., Vidura Advisor was also discussed. Participants were largely open to the idea of interacting with a chatbot expressing that having the ability to enable or disable a chatbot as needed could provide some needed support during the learning phase of utilizing a resource such as OnTimeEvidence. One participant stated that having a chatbot available could be the difference between effectively utilizing a resource and giving up on it: "*We've all been on a website where we cannot figure out how to do the next step. And sometimes you just give up and go home. And so it would be good to have a way to get over those obstacles as you are working through the...database*" [P01]. Even if the chatbot could assist a small group of new users, it can be helpful to these users who may otherwise give up using all the OnTimeEvidence capabilities due to frustration in navigation of the user interface or specific analytics features.

Participants suggested that some helpful features for the chatbot would include ease of use, help with search inquiries, learn functionalities, and offer suggestions based on common search errors. They emphasized that a good first impression is key, with the chatbot needing to respond correctly to queries and offering relevant options. One way a chatbot could do this is by offering "multiple choice options," offering specific choices or "all of the above" [P09] regarding the parameters of studies being searched for. By implementing such features, Vidura can provide an essential service to users of OnTimeEvidence, allowing for easier navigation of the science gateway and its features and assuring that users come back due to successful, efficient inquiries. When asked if they would be interested and willing to participate in a usability study of the gateway and chatbot in the next phase of our study, all 10 participants agreed to be contacted for a follow-up study. This as a sign of their enthusiasm for the promise of our OnTimeEvidence coupled with the Vidura Advisor that interfaces with the underlying plug-in management middleware.

6 | CONCLUSION

In this paper, we have detailed a novel socio-technical approach to handle the needs of data consumers accessing large-scale datasets and computational resources surrounding the COVID-19 pandemic by developing a next-generation science gateway viz., OnTimeEvidence. OnTimeEvidence provides publication analytics tools for data consumers (e.g., clinicians, researchers) for performing knowledge discovery over large-scale publication archives related to critical areas such as COVID-19 pandemic response. It implements an intelligent agent viz., Vidura Advisor, which leverages a set of plug-in middleware components to allow data providers (e.g., administrators, developers) to orchestrate data processing pipelines and provision data consumers with access to large-scale publication datasets. The Vidura plug-in middleware management is capable of being generally extended to other science gateways based on a set of customization steps. We have presented details on how the Vidura plug-in management middleware can be integrated at the infrastructure level for data providers and at the application level for data consumers. We have also studied the important social context of widespread adoption requirements of such technologies from data consumers perspective, which then guides the implementation and diffusion of OnTimeEvidence with user teams at community-scale.

We demonstrate the usefulness of the OnTimeEvidence science gateway with Vidura plug-in middleware management approach by conducting a usability study. Herein, we conducted a qualitative assessment with 10 participants and asked questions regarding the adoption of the OnTimeEvidence science gateway as well as the usefulness of the Vidura Advisor that manages the underlying plug-in management middleware. Our results suggest that the participants are highly inclined to use a science gateway application such as OnTimeEvidence, while also believing the Vidura Advisor can be impactful for guiding them into using the science gateway functionalities. The Vidura Advisor through the plug-in management middleware aims to provide effective guidance for science gateway users (e.g., data providers and data consumers) and handle the dynamic, growing need of consumers in an effort to ultimately foster widespread adoption at a community-level.

While this study focuses on the socio-technical stage of adoption, our future work is to allow for users to routinely interact with our implementation of the OnTimeEvidence capabilities through more detailed and long-term usability evaluations. To study the effectiveness of the Vidura Advisor, we also plan to study various metrics behind the dialogue between data consumers and the chatbot plug-in to measure the effectiveness of the intelligent agent in helping improve the productivity and collaboration tasks of data providers and data consumers. In addition, we plan to address the adoption challenges for data providers by evaluating the usefulness of our Vidura plug-in management middleware within OnTimeEvidence through an extended usability study.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation under awards: OAC-1730655, 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. Timm D. Evidence Matters. *Journal of the Medical Library Association: JMLA* 2006; 94(4): 480.
2. Murad MH, Asi N, Alsawas M, Alahdab F. New Evidence Pyramid. *BMJ Evidence-Based Medicine* 2016; 21(4): 125–127.
3. Demeler B. UltraScan: a comprehensive data analysis software package for analytical ultracentrifugation experiments. *Modern analytical ultracentrifugation: techniques and methods* 2005: 210–229.
4. Sivarathri SS, Calyam P, Zhang Y, et al. Chatbot Guided Domain-science Knowledge Discovery in a Science Gateway Application. *Proceedings of Gateways* 2019: 1–4.
5. Rogers EM. *Diffusion of innovations*. Simon and Schuster . 2010.
6. De Coninck Q, Michel F, Piraux M, et al. Pluginizing quic. *Proceedings of the ACM Special Interest Group on Data Communication* 2019: 59–74.
7. OnTimeEvidence GitHub Repository. Accessed January 2022. [Online]; . Available at https://github.com/Xiyao-Cheng/OnTimeEvidence.
8. Ahmed K, Bukhari MA, Mlanda T, et al. Novel approach to support rapid data collection, management, and visualization during the COVID-19 outbreak response in the world health organization African region: Development of a data summarization and visualization tool. *JMIR Public Health and Surveillance* 2020; 6(4): e20355.
9. Peddireddy AS, Xie D, Patil P, et al. From 5vs to 6cs: Operationalizing epidemic data management with covid-19 surveillance. In: IEEE. ; 2020: 1380–1387.
10. Alam F, Almaghthawi A, Katib I, Albeshri A, Mehmood R. IResponse: An AI and IoT-enabled framework for autonomous COVID-19 pandemic management. *Sustainability* 2021; 13(7): 3797.
11. Otoom M, Otoum N, Alzubaidi MA, Etoom Y, Banihani R. An IoT-based framework for early identification and monitoring of COVID-19 cases. *Biomedical Signal Processing and Control* 2020; 62: 102149.
12. Use Deep Search to Explore the COVID-19 Corpus. 2020; .
13. Staar PW, Dolfi M, Auer C. Corpus processing service: A Knowledge Graph platform to perform deep data exploration on corpora. *Applied AI Letters* 2020; 1(2): e20.
14. Trewartha A, Dagdelen J, Huo H, et al. COVIDScholar: An automated COVID-19 research aggregation and analysis platform. *arXiv preprint arXiv:2012.03891* 2020.
15. Beltagy I, Lo K, Cohan A. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* 2019.
16. Moran A, Hampton S, Dowson S, et al. Online Interactive Platform for COVID-19 Literature Visual Analytics: Platform Development Study. *Journal of Medical Internet Research* 2021; 23(7): e26995.
17. Lever J, Altman RB. Analyzing the vast coronavirus literature with CoronaCentral. *Proceedings of the National Academy of Sciences* 2021; 118(23).
18. Oruche R, Gundlapalli V, Biswal AP, et al. Evidence-based Recommender System for a COVID-19 Publication Analytics Service. *IEEE Access* 2021.

19. Zhang Y, Calyam P, Joshi T, Nair S, Xu D. Domain-specific Topic Model for Knowledge Discovery through Conversational Agents in Data Intensive Scientific Communities. In: IEEE. ; 2018: 4886–4895.

20. Savelyev A, Brookes E. GenApp: Extensible tool for rapid generation of web and native GUI applications. *Future Generation Computer Systems* 2019; 94: 929–936.

21. Dooley R, Brandt SR, Fonner J. The Agave Platform: An open, science-as-a-service platform for digital science. *Proceedings of the Practice and Experience on Advanced Research Computing* 2018: 1–8.

22. Pierce M, Marru S, Demeler B, Singh R, Gorbet G. The apache airavata application programming interface: overview and evaluation with the UltraScan science gateway. *Gateway Workshop* 2014.

23. Phase D. Cloud application modelling and execution language (CAMEL) and the PaaSage workflow. *Advances in Service-Oriented and Cloud Computing: Workshops of ESOCC 2015, Taormina, Italy, September 15-17, 2015, Revised Selected Papers* 2016; 567: 437.

24. Merchant N, Lyons E, Goff S, et al. The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS biology* 2016; 14(1): e1002342.

25. Madduri R, Chard K, Chard R, et al. The Globus Galaxies platform: delivering science gateways as a service. *Concurrency and Computation: Practice and Experience* 2015; 27(16): 4344–4360.

26. Kiss T, Kacsuk P, Kovács J, et al. Micado—microservice-based cloud application-level dynamic orchestrator. *Future Generation Computer Systems* 2019; 94: 937–946.

27. Albayrak N, Özdemir A, Zeydan E. An overview of artificial intelligence based chatbots and an example chatbot application. *2018 26th Signal processing and communications applications conference (SIU)* 2018: 1–4.

28. Ni L, Lu C, Liu N, Liu J. Mandy: Towards a smart primary care chatbot application. *International symposium on knowledge and systems sciences* 2017: 38–52.

29. Baby CJ, Khan FA, Swathi J. Home automation using IoT and a chatbot using natural language processing. *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)* 2017: 1–6.

30. Muslih M, Somantri , Supardi D, et al. Developing smart workspace based IOT with artificial intelligence using telegram chatbot. *2018 International Conference on Computing, Engineering, and Design (ICCED)* 2018: 230–234.

31. Kar R, Haldar R. Applying chatbots to the internet of things: Opportunities and architectural elements. *arXiv preprint arXiv:1611.03799* 2016.

32. Vekaria K, Calyam P, Sivarathri SS, et al. Recommender-as-a-service with chatbot guided domain-science knowledge discovery in a science gateway. *Concurrency and Computation: Practice and Experience* 2020: e6080.

33. Chandrashekara AA, Talluri RKM, Sivarathri SS, et al. Fuzzy-based conversational recommender for data-intensive science gateway applications. *2018 IEEE International Conference on Big Data (Big Data)* 2018: 4870–4875.

34. Sivarathri SS, Calyam P, Zhang Y, et al. Chatbot Guided Domain-science Knowledge Discovery in a Science Gateway Application. *Proceedings of Gateways* 2019: 1–4.

35. Kiss T, Bolotov A, Pierantoni G, et al. Science gateways with embedded ontology-based E-learning support. In: CEUR Workshop Proceedings. ; 2020.

36. Kee K, Sleiman M, Williams M, Stewart D. Sciserver compute: The 10 attributes that drive adoption and diffusion of computational tools in e-science. *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale* 2016: 1–8.

37. Kee K, Le B, Jitkajornwanich K. If you build it, promote it, and they trust you, then they will come: Diffusion strategies for science gateways and cyberinfrastructure adoption to harness big data in the science, technology, engineering, and mathematics (STEM) community. *Concurrency and Computation: Practice and Experience* 2021: e6192.

38. Davis F. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly* 1989: 319–340.

39. Stewart C, Simms S, Plale B, Link M, Hancock D, Fox G. What is cyberinfrastructure. *Proceedings of the 38th annual ACM SIGUCCS fall conference: navigation and discovery* 2010: 37–44.

40. Yue X, Wang H, Jin D, Li M, Jiang W. Healthcare data gateways: found healthcare intelligence on blockchain with novel privacy risk control. *Journal of medical systems* 2016; 40(10): 1–8.

41. Dinakarrao SMP, Sayadi H, Makrani HM, Nowzari C, Rafatirad S, Homayoun H. Lightweight node-level malware detection and network-level malware confinement in iot networks. *2019 Design, Automation & Test in Europe Conference & Exhibition (DATE)* 2019: 776–781.

42. Matos DR, Pardal ML, Adao P, Silva AR, Correia M. Securing electronic health records in the cloud. *Proceedings of the 1st Workshop on Privacy by Design in Distributed Systems* 2018: 1–6.

43. García ÁL, De Lucas JM, Antonacci M, et al. A cloud-based framework for machine learning workloads and applications. *IEEE access* 2020; 8: 18681–18692.

44. Bayat A, Szul P, O'Brien AR, et al. VariantSpark: Cloud-based machine learning for association study of complex phenotype and large-scale genomic data. *GigaScience* 2020; 9(8): giaa077.

45. Simmhan Y, Aman S, Kumbhare A, et al. Cloud-based software platform for big data analytics in smart grids. *Computing in Science & Engineering* 2013; 15(4): 38–47.

46. Tuli S, Tuli S, Tuli R, Gill SS. Predicting the growth and trend of COVID-19 pandemic using machine learning and cloud computing. *Internet of Things* 2020; 11: 100222.

47. Abdel-Basset M, Chang V, Nabeeh NA. An intelligent framework using disruptive technologies for COVID-19 analysis. *Technological Forecasting and Social Change* 2020: 120431.

48. Abdel-Basset M, Chang V, Mohamed R. HSMA_WOA: A hybrid novel Slime mould algorithm with whale optimization algorithm for tackling the image segmentation problem of chest X-ray images. *Applied Soft Computing* 2020; 95: 106642.

49. Abdel-Basset M, Chang V, Hawash H, Chakrabortty RK, Ryan M. FSS-2019-nCov: A deep learning architecture for semi-supervised few-shot segmentation of COVID-19 infection. *Knowledge-Based Systems* 2021; 212: 106647.

50. Sackett DL. Evidence-based medicine. *Seminars in perinatology* 1997; 21(1): 3–5.

51. Eren ME, Solovyev N, Raff E, Nicholas C, Johnson B. COVID-19 Kaggle Literature Organization. *Proceedings of the ACM Symposium on Document Engineering 2020* 2020: 1–4.

52. Zhang Y, Calyam P, Joshi T, Nair S, Xu D. Domain-specific Topic Model for Knowledge Discovery in Computational and Data-Intensive Scientific Communities. *IEEE Transactions on Knowledge and Data Engineering* 2021.

53. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. *Journal of machine Learning research* 2003; 3(Jan): 993–1022.

54. Milken Institute . COVID-19 TREATMENT AND VACCINE TRACKER. 2020; .

55. Pickett BE, Sadat EL, Zhang Y, et al. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic acids research* 2012; 40(D1): D593–D598.

56. Tan Z, He L. An efficient similarity measure for user-based collaborative filtering recommender systems inspired by the physical resonance principle. *IEEE Access* 2017; 5: 27211–27228.

57. Google Dialogflow (2021). [Online]; . Available at https://cloud.google.com/dialogflow/docs.

58. Ranavare SS, Kamath R. Artificial intelligence based chatbot for placement activity at college using dialogflow. *Our Heritage* 2020; 68(30): 4806–4814.

59. Spring Cloud Netflix (2020). [Online]; . Available at https://spring.io/projects/spring-cloud-netflix.

60. Vekaria KB, Calyam P, Oruche R, Zhang Y, Wang S. "Bring-your-own" Plug-in Management Middleware for Programmable Science Gateways. *Science Gateways Community Institute* 2020.

61. Vekaria K, Calyam P, Sivarathri SS, et al. Recommender-as-a-service with chatbot guided domain-science knowledge discovery in a science gateway. *Concurrency and Computation: Practice and Experience* 2021; 33(19): e6080.

62. Kee KF, Schrock AR. Telephone interviewing as a qualitative methodology for researching cyberinfrastructure and virtual organizations. *Second international handbook of internet research* 2020: 351–365.

63. Kee KF, Thompson-Hayes M. Conducting effective interviews about virtual work: Gathering and analyzing data using a grounded theory approach. *Virtual work and human interaction research* 2012: 192–212.

64. Baxter LA, Babbie ER. *The basics of communication research*. Cengage Learning . 2003.