CHAPTER 6

Cloud-based data pipeline orchestration platform for COVID-19 evidence-based analytics

Mauro Lemus Alarcon, Roland Oruche, Ashish Pandey, and Prasad Calyam
Electrical Engineering and Computer Science Department, University of Missouri-Columbia, Columbia, MO, United States

1 Challenges in COVID-19 data handling

Accessing massive collections of prior medical literature and handling the ongoing data deluge create challenges for health-care data consumers (e.g., clinicians and researchers) who need to make timely data-driven decisions related to the COVID-19 pandemic response. Current practice still heavily relies on time-consuming and onerous manual methods to search, compile, and select the articles that are relevant for gaining insights to shape outcomes (Ioannidis, Salholz-Hillel, Boyack, & Baas, 2020). The COVID-19 pandemic demands swift actions from researchers and clinicians, and there is a dire need for robust tools to help them manage the datasets in research tasks, and also to enable them collaborate with other experts based on critical evidence (Kricka et al., 2020). The tools also need to be integrated within unified data-sharing platforms that increase accessibility to specialized literature and support data analytics automation to expedite, e.g., search and analysis processes. Even more importantly, the tools need to be accessible in a flexible and scalable manner by utilizing cloud-based deployments with necessary interfaces to integrate open-source tools and health-care social networks.

1.1 Cloud- and AI-based data pipeline platform

Data pipelines are increasingly being used to combine data from multiple sources, allow access to multiple users, and include multiple data analytics tools to orchestrate data collection and processing. To handle such data

Novel AI and Data Science Advancements for Sustainability in the Era of COVID-19
https://doi.org/10.1016/B978-0-323-90054-6.00003-9

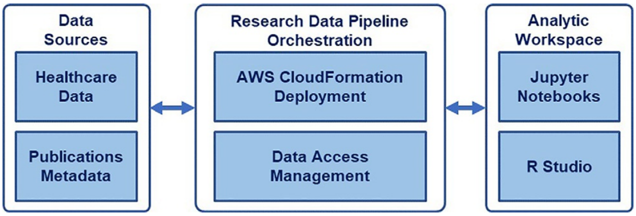

Fig. 1 Cloud-based data pipeline orchestration platform components for COVID-19 evidence-based analytics.

pipelines, exemplar open technologies such as Observational Health Data Sciences and Informatics (OHDSI) (Hripcsak et al., 2015) have been developed, which are yet to be customized and explored for the COVID-19 response purposes. The open architecture of such technologies makes it feasible to integrate open-source data analytics tools and create interactive interfaces to perform data processing. In this chapter, we describe "OnTimeEvidence," a cloud-based data pipeline orchestration platform built on OHDSI for COVID-19 evidence-based text (e.g., publications) and data (e.g., electronic health records) analytics as shown in Fig. 1. More specifically, we describe how OnTimeEvidence leverages the concept of a modern data pipeline platform that uses a technology agnostic architecture on the Amazon Web Services (AWS) cloud platform (Wiggins, 2018) and integrates AI-based analytic tools.

Both structured and unstructured data from multiple sources [i.e., Synthetic Public Use Files (SynPUF) (Borton et al., 2010) health-care data, Kaggle COVID-19 Open Research Dataset (CORD-19) (Ekin Eren, Solovyev, Raff, Nicholas, & Johnson, 2020)] can be stored in a repository that uses redshift data-warehouse services in AWS, and follows the Common Data Model (CDM) standard (Makadia & Ryan, 2014). The repository can be utilized for multiple data processing tasks involving, e.g., natural language processing, machine learning depending on assigned roles of users who have relevant entitlements for managing data access and processing. Users manage the data processing tasks in their customized analytics workspaces that provide the necessary tools to retrieve and analyze the data with user-friendly Jupyter Notebooks and R Studio interfaces. A particular open-source tool that we integrate in OnTimeEvidence is the Domain-specific Topic Model (DSTM)-based publication analytics tool (Zhang, Calyam, Joshi, Nair, & Xu, 2018) that helps users in inferring latent patterns across COVID-19 or other scientific domain literature documents. Using a detailed case study, we show how OnTimeEvidence helps health-care data consumers to submit data requests and retrieve the data in a secure, consistent, and standard manner to analyze COVID-19-related literature. Users are also provided with access to analytic tools which help them to (a) conduct knowledge discovery tasks while reducing the manual burden in compiling and analyzing large datasets and (b) run data analytic processes on disparate systems with minimal automation.

1.2 Chapter organization

In this book chapter, we first present the background of open technologies and issues around cloud-hosted data processing pipelines with AI-based tools. Next, we introduce our OnTimeEvidence platform and detail its architecture components. Following this, we present a case study to show the benefits of deploying OnTimeEvidence to help with COVID-19-related data analytics requiring data access control and use of versatile analytic workspaces. Lastly, we list a set of open issues for how cloud- and AI-based platforms could be further developed to not only handle COVID-19 crisis, but also support needs of health-care data consumers to unlock the promise of "precision medicine" that can help better cure cancer and other diseases (Friedman, Letai, Fisher, & Flaherty, 2015).

2 Background and related works

In this section, we first provide a background on existing exemplar cloud-based data pipeline orchestration solutions (i.e., the OHDSI on AWS platform) that motivate our OnTimeEvidence platform design. Next, we describe best practices in cloud-based health-care data analytics and sharing. Lastly, we summarize latest advances in cloud-based data processing pipeline schemes that can benefit health-care data consumers.

2.1 OHDSI on AWS infrastructure

The OHDSI program is committed to promote the importance of health data analytics through the development and release of open-source data analytics tools (i.e., ATLAS, ACHILLES, ATHENA) (Hripcsak et al., 2015). These tools have common features which allow them to interact with a CDM (Makadia & Ryan, 2014) that can be implemented using multiple database management systems (e.g., Postgresql, Redshift). Through proper extraction, transformation, and loading (ETL) processes, disparate structured and unstructured data sources can be integrated into the CDM repository under a well-defined data structure that will allow the analytic tools to utilize templates to run standardized data analytic processes and generate insightful results.

Our OnTimeEvidence platform builds on the open-source OHDSI on AWS solution, and extends the out-of-the-box automated CloudFormation deployment that includes a Redshift data-warehouse infrastructure instance. This instance hosts the CDM repository physical model, and the data analytic tools that allow OnTimeEvidence users to interact with the CDM. As part of the OnTimeEvidence deployment, we have loaded the SynPUF dataset into the CDM. The deployment of this platform takes a few hours but it removes the manual burden in the design, development, and deployment efforts required by a regular IT infrastructure process with manual steps. We complemented the data repository by adding tables to store the CORD-19 metadata about COVID-19 literature, and loaded the related dataset into those tables. On top of this infrastructure, we developed a centralized role-based data access model to provide entitlements to authorize users to access both datasets and analytic resources. To facilitate the data retrieval and analysis tasks, we developed embedded data request forms within a JupyterLab environment that enables researchers to submit data requests and retrieve the required data within the Jupyter workspace. Once the user-requested data is available, users have access to various analytic tools, and can run correlation rules on multiple datasets. Thereby, they can produce relevant diagnostics to discover insights for COVID-19-related research tasks.

2.2 Cloud-based health-care data management

Multiple solutions have been developed to store and share health-care data in cloud environments, keep those records secure in such environments, provide analytic services related to health big data, and preserve data privacy. In the context of data accessibility, the work in Health-care Data Gateway (Yue, Wang, Jin, Li, & Jiang, 2016) aims to securely store Electronic Health Record (EHR) data in a cloud-based platform and uses a Blockchain-based secure storage layer. Data sharing is supported among multiple users (i.e., physicians, researchers, government institutions, private organizations) based on role assignments. Similarly, in their work, Matos, Pardal, Adao, Silva, and Correia (2018) proposed a system to store EHR in a public cloud, and their focus was on ensuring data confidentiality and integrity by using an access control mechanism based on the lattice model. Using such a model, users can define a hierarchy of data access levels (i.e., private, clinic, research, public). Identify and access management solutions focusing more on cloud-based data sharing while preserving privacy as have been proposed by Hörandner, Krenn, Migliavacca, Thiemer, and Zwattendorfer (2016), Sharma, Chen, and Sheth (2018), and Barik, Dubey, and Mankodiya (2017). In these works, once a user authentication is complete, data access from multiple client devices can be allowed or patient data can be collected from multiple sources. However, none of these prior works provide user-customized analytics workspaces, which in turn leads to users having to manage any retrieved data manually outside of their platforms.

Access management best practices related to centralization need to be carefully designed (Cohen & Nissim, 2018). Among the access management best practices, role-based access control (RBAC) has been the most popular one (Dinakarrao et al., 2019). RBAC restricts platform users' permissions to their roles and only permits users access to privileges that they absolutely need to perform their job functions. For example, health-care students of an organization should not have access to digital financial records of patients. In addition, RBAC also helps facilitate identity security, operational processes, and cybersecurity visibility. As part of our access management best practices, it is important to assign clear, delineated roles to all users. Ideally, this includes privileged users such as faculty members with Institutional Review Board (IRB) approved projects having more entitlements compared to regular users such as students. Moreover, RBAC implementations need to ensure that no user should receive permissions outside his/her role. However, if projects demand the assignment of temporary privileges, those privileges should expire within a set time limit to ensure long-term security of the data access.

2.3 Cloud-based data processing pipelines

Prior works have exemplified the need to provide open-source, cloud-based frameworks for the deployment of data processing pipelines. In the work by García et al. (2020) a distributed cloud-based framework has been developed, viz., DEEP to enable researchers to process and train their machine learning data science models. The DEEP framework integrates serverless architecture to ease the transition from deployment to production. In a similar fashion, VariantSpark (Bayat et al., 2020) is a distributed machine learning framework that performs association analysis to effectively identify variants with complex phenotypes. It features a multilayer parallelization that allows the framework to scale the whole genome population dataset for developing an in-depth analysis using its machine learning pipelines. Simmhan et al. (2013) resolved the critical concern of optimizing supply-demand needs of customers in a Smart Grid Project by developing a robust cloud framework that leverages machine learning and data processing pipelines.

In recent work related to the COVID-19 pandemic, Tuli, Tuli, Tuli, and Gill (2020) resolved the need to handle increasing rate of COVID-19 through a data-driven model deployed on a cloud-based framework that predicts the growth of the pandemic. Abdel-Basset, Chang, and Nabeeh (2020) developed an intelligent framework of emerging AI-based technologies for helping with the COVID-19 pandemic response. Their work suggests that these disruptive technologies can be integrated in IoT and IoMT devices using cloud platforms. Abdel-Basset, Chang, & Mohamed, 2020 sought to resolve the image segmentation problem in COVID-19 chest X-rays by developing a novel machine learning 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, Abdel-Basset, Chang, Hawash, Chakrabortty, and Ryan (2021) addressed the issue of providing accurate classification of COVID-19 in CT scans by developing a deep learning architecture that leverages a semisupervised few-shot segmentation algorithm for image segmentation. The work by Otoom, Otoum, Alzubaidi, Etoom, and Banihani (2020) presents an Internet of things (IoT)-based framework that performs real-time monitoring and tracking of COVID-19 data. The framework entails the aggregation of data from multiple resources in a cloud infrastructure where stakeholders (e.g., health physicians) can monitor patients through data processing algorithms. A study by Ashraf et al. (2020) developed a smart surveillance system for effective remote monitoring of human health conditions and close interactions. Similarly, Hossain, Muhammad, and Guizani (2020) developed a mass surveillance system through a hierarchical edge computing service using 5G wireless connectivity and deep learning algorithms.

OnTimeEvidence builds upon the above works and uses data-driven models deployed on OHDSI to allow researchers to effectively perform knowledge discovery pertinent to the COVID-19-related datasets. We develop a data-driven modeling scheme through our existing Domain-specific Topic Model (DSTM) (Zhang et al., 2018), which is an extension of the Latent Dirichlet allocation (Blei, Ng, & Jordan, 2003) to discover the relationships between words and tools/resources (e.g., drugs and genes) related to the COVID-19 pandemic. In addition, we also use Gibbs sampling algorithm (Griffiths & Steyvers, 2004) to infer latent patterns within the COVID-19 domain in an unsupervised manner.

3 OnTimeEvidence architecture and component implementation

In this section, we introduce our OnTimeEvidence platform and its components as illustrated in Fig. 2. The core component of the data pipeline orchestration is built on top of the open-source OHDSI on AWS. The AWS CloudFormation is used to deploy OnTimeEvidence along with the data access and process management module, entitlement database, and access admin console. We leverage the JupyterLab included with OHDSI on AWS to facilitate users' data access and interaction, and create extensions such as the user data request forms and the data processing models in order to provide the analytic workspace for health-care data and COVID-19 publications analysis as well as result sharing. We have uploaded the SynPUF Medicare and the CORD-19 datasets to a relational repository on the OHDSI Red-shift data warehouse service, and the related health-care data and COVID-19-related literature information are available for process testing and validation of user utility.

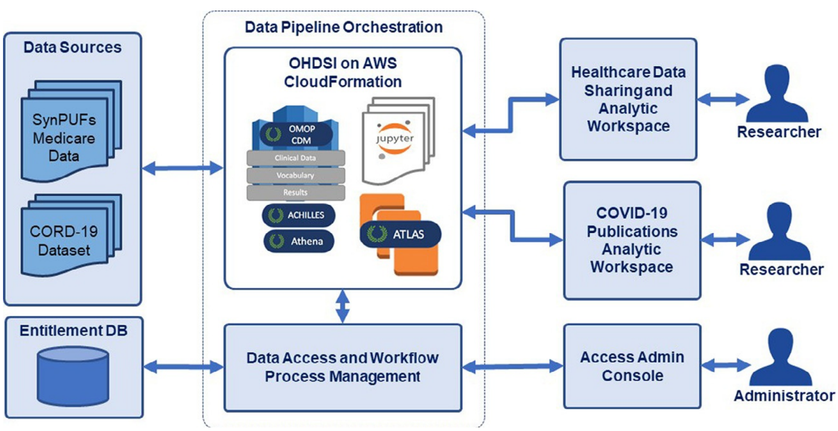

Fig. 2 Components of OnTimeEvidence data pipeline orchestration built on top of the OHDSI on AWS infrastructure.

3.1 OHDSI components of OnTimeEvidence

OnTimeEvidence uses various OHDSI components featuring in-built data repositories and software capabilities. The deployment of OHDSI components provides an enterprise class, multiuser, and scalable health-care data sharing and analytics functionality. As shown in Fig. 2, OHDSI components include the OMOP-CDM deployed on a Redshift data warehouse. The CDM schema allows the integration of disparate data sources into a common format (model) and common representation (terminology, vocabulary, coding) allowing the definition and execution of standard analytic processes. Once data is available in the CDM, evidence knowledge can be generated using the analytic tools included with the OHDSI on AWS platform (i.e., Athena, ATLAS), and the analytic models and tools available in the analytic workspace available via Jupyter Notebooks or R-Studio. The OHDSI components include out-of-the-box open source analytic tools such as (i) ATLAS, a web-based application for researchers to conduct analyses on data loaded to the OMOP-CDM through the creation of cohorts based on drug exposure or diagnosis of a particular condition. The cohort results are visualized in the tool's user interface, or stored in a relational repository to be used by other analytic tools; (ii) ACHILLES, an application used to analyze the database hosting the CDM and evaluate data quality; and (iii) ATHENA, a tool that is used to generate and load standardized data vocabularies into the CDM repository.

OnTimeEvidence extends OHDSI through the integration of the following new components we have developed: (i) a role-based user access and workflow management component to keep control on the authentication and authorization of the data, and ensure data privacy and security compliance; this functionality allows users to submit data requests, which are fulfilled by OnTimeEvidence based on the role-based user access privileges and (ii) the functionality that allows users to perform publication analytics (considering knowledge pattern mining of hundreds or thousands of articles) and big data analytics (considering millions of patient-related records) relevant to COVID-19. The original OHDSI features are complemented by the above two components and the result is a robust platform for researchers to analyze data with open-source tools, and find correlations as well as gain insights all within the same platform. Consequently, OnTimeEvidence reduces the burden of researchers to handle large-scale data compilation and analysis in cloud infrastructures, and enables them to focus on their scientific research goals instead.

3.2 Access and authorization management

Herein, we provide the details regarding the data access and workflow management components in OnTimeEvidence. For user management, we created a centralized role-based access control mechanism to manage users' credentials and privileges through an administrator interface. The user account creation workflow process steps are illustrated in Fig. 3. Users are first asked via the OnTimeEvidence web interface to provide essential data to validate their privileges and necessary data requirements for their analytics tasks. Users are accordingly granted role privileges by the administrator using roles such as student, faculty, or independent researcher depending on the data access requirements and user status in an institution. Once the administrator verifies user credentials, a user account corresponding to a group is created on the OHDSI database server through an administrator web interface that runs customized shell scripts in the backend to automate the user account creation and group mapping. After user account creation, administrators can also dynamically regulate the level of access provided to users thus protecting sensitive and proprietary data. A large number of roles consisting of a combination of access to different attributes can be stored in the entitlement database to allow administrators on creating user accounts with flexibility in access control rules governing the platform.

The control actions performed by the administrator utilize the HTTPS protocol to connect and communicate with the application and database server in a secure manner. Administrator password and user access request form are secured on a separate database system (disjoint with the original

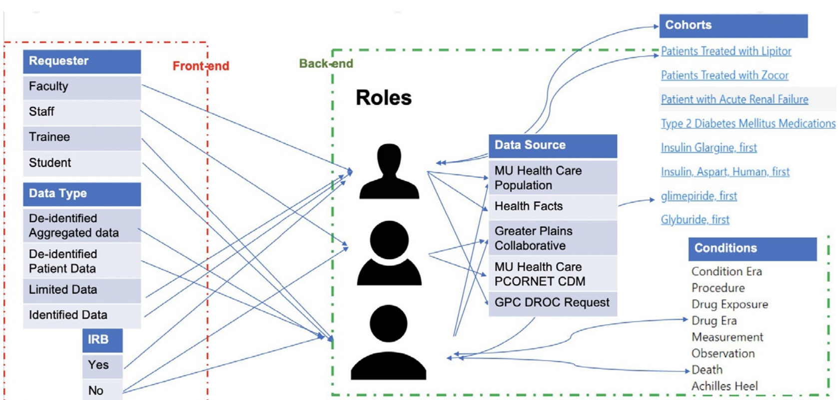

Fig. 3 Role-based access control enabled by a web interface and entitlement database using automated shell scripts.

OHDSI components, e.g., CDM) identified as the entitlement database server in the architecture schematics shown in Fig. 2. The role-based access control and separation of concern for user access request form and administrator authentication on a different entitlement (database) server reduces the attack surface on the overall platform. To increase security, we enforce strong passwords (at least eight characters with uppercase and lowercase letters, numbers, and special characters) for authentication of users and the administrator. We have added further security measures by using the *bcryptjs* library for hashing the password and other user data at rest. Field level encryption is used for protecting sensitive user data such as details about the project or their privileges. This ensures that the actual explicit password and data are not accessible to hackers even if they have access to the entitlement database. To save the users from malicious security attacks, strict protocols such as form and data validations are put in place at each layer of the control flow. Consequently, a malicious user cannot execute attacks such as SQL injection attacks on the application or database server. Data validation at the application layer is also performed to prevent users from storing undesired data on the web server.

3.3 COVID-19 literature selection and analysis

The OnTimeEvidence platform can be used with customized analytics workflows in which health-care data consumers (e.g., clinicians, health professionals) 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 (Sackett, 1997). Health-care data consumers commonly adopt this method for synthesizing and reviewing articles based on the inherent evidence levels that are pertinent to their research. Specifically, a hierarchical evidence-based framework, viz., Levels of Evidence Pyramid (Murad, Asi, Alsawas, & Alahdab, 2016), is used. The Levels of Evidence Pyramid illustrate the reduced quantity of publications with respect to the increase in high-quality information (e.g., background information to systematic reviews). However, it remains a challenge for clinical researchers to sort and filter information based on high-quality evidence in a timely manner.

To simplify literature data selection and analysis for COVID-19 researchers, we implement new components in the OnTimeEvidence platform to cater to the user's needs by reducing their manual steps in scientific workflows as illustrated in Fig. 4. The implemented tasks performed include:

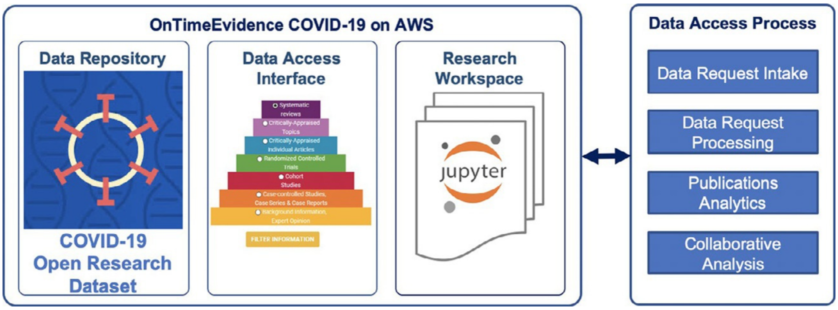

Fig. 4 COVID-19 literature selection and analysis workflow process in OnTimeEvidence for knowledge discovery.

(i) a literature selection form that allows a user to query search terms related to the levels of evidence, (ii) the functionality to process the literature selection, and, in response, generation of a new Jupyter Notebook with an embedded SQL query to execute the information requested by the user, and (iii) use the JupyterLab workspace to allow the user to conduct a publication and/or collaborative analysis and store the results for sharing them via the platform.

The above OnTimeEvidence component for literature selection and analysis described 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) (Ekin Eren et al., 2020). This dataset is stored on an Amazon Redshift database server that hosts CDM data models as shown in Fig. 2. With the publication archives stored in the Red-Shift cluster, the user is able to search for articles by using the request form through a data access request form on a JupyterLab interface. The form allows the user to submit a COVID-19 literature selection based on the selected Level of Evidence. Following this approach, the data access request form processes the health-care data consumers' requests and generates the SQL query to retrieve the related literature selected. In this process, we have developed an intuitive client-interface form that is rendered on the JupyterLab workspace within the default view for data consumers, i.e., the request form is displayed whenever users have been given accesses by the administrator in the workspace, and they can use it to submit a new publications data access request.

Fig. 5 illustrates the process of the CDM to execute the SQL query and generate a new Jupyter Notebook on the user's workspace once the data

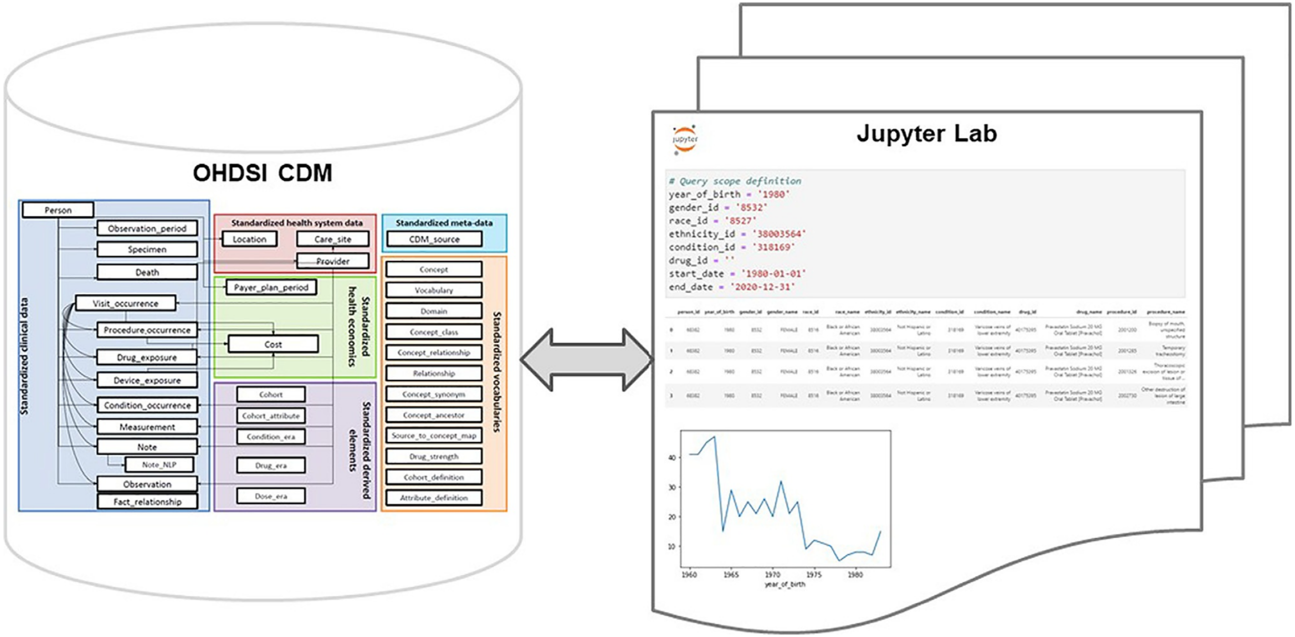

Fig. 5 Integration between a Jupyter Notebook and the CDM via a SQL query to retrieve data and perform data analytics for a given COVID-19 research task.

request form is submitted and processed from the user's query. The user will use this notebook to execute the query against the CDM and retrieve the required data. The user does not have to know the structure or content of the CDM repository to retrieve data as the query provides the required definition to fulfill the data request. However, if the user has some knowledge of the CDM structure and data content, it will be possible for the user to modify the query and retrieve a new dataset. The scope of the data being accessed by the user will be limited by the user's role. Therefore, even if the user attempts to access data not allowed for the related role, the corresponding query will not work due to the access security mechanism. This feature allows the user to explore only the data that the user has access to when using the initial SQL query statement or modify the queries in the user data request submission. New SynPUF data following the CDM model can also be added and processed in the OnTimeEvidence platform for varying analytics challenges by the researchers and administrators. Once the SQL query provides a required list of selected articles, the user will be able to conduct analytic tasks using the Python libraries available in the JupyterLab environment, generate the required results, and store or share those results within the same environment.

3.4 Data processing using domain-specific topic model

With the implementation of new components on top of OHDSI, OnTimeEvidence enables the collection of external resources and deployment of machine learning models via open-source tools. In this section, we illustrate the utility of open-source tools in OnTimeEvidence to filter high-quality information and reduce the time-expensive workflow steps involved in performing knowledge discovery over COVID-19 publication archives using data processing pipelines.

Particularly, we detail our Domain-specific Topic Model (DSTM) (Zhang et al., 2018)-based tool that can be used to deploy statistical and deep generative models for guiding researchers to rapidly discover high-quality information (in terms of Evidence Levels) from aggregated medical resources (e.g., publication databases, information on drugs and genes). DSTM extends the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) model to discover domain-specific topics and latent knowledge patterns among the topics and users' interests. The LDA model is a powerful tool that is capable of representing a document through a Dirichlet distribution of random topics and, in parallel, represents each topic as a distribution of key terms. Our DSTM also utilizes inference algorithms such as variational inference and Gibbs sampling (Griffiths & Steyvers, 2004) to infer those latent parameters. DSTM learns two distinct multinomial distributions (Dai, Ding, Wahba, et al., 2013) to generate random topics: (i) topic distributions that are distributions over terms and (ii) document distributions that also are distributions over topics. Given that LDA generates random topics in the scientific literature, we have configured our DSTM tool to generate domain-specific topics through the aggregation of scientific terms related to the COVID-19 research areas.

In utilizing the LDA model for identifying the distribution of topics and their respective terms, we leveraged our DSTM (Zhang et al., 2018) to discover the latent patterns of specific scientific drug and gene terms in salient medical information from the COVID-19 Vaccine Tracker (Milken Institute, 2020) and Virtual Incident Procurement (ViPR) (Pickett et al., 2012). Clinical researchers commonly refer to important criteria including drugs and gene tools to further study the issues relating to infectious disease, immunology, and epidemic/pandemic control. Hence, we simplify the computational complexity of our generative model by generating each word based on a drug or a gene.

As shown in Fig. 6, our DSTM works automatically learns the latent patterns underlying the datasets. Each document represents the collected document from the CORD-19 Kaggle dataset. We decompose a scientific paper into a list of information about the research topic, and related drugs

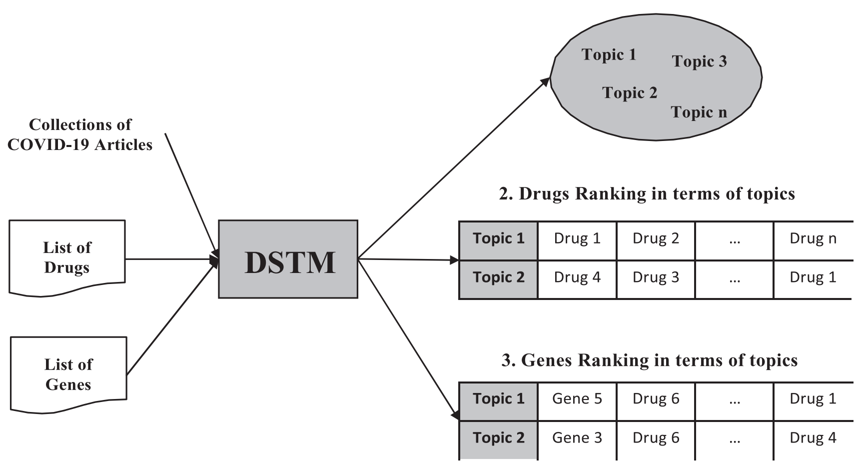

Fig. 6 Domain-Specific Topic Model (DSTM) deployed on OnTimeEvidence works as an analysis engine to discover the relationships among research topics, drugs, and genes.

and genes mentioned in the publication. The goal of DSTM is to learn the relationships among the research topics, drugs, and genes using an unsupervised machine learning approach. To train the DSTM, we only need to input a collection of COVID-19 articles, a list of drugs, and a list of genes. After completion of the training phrase, the DSTM can help users to analyze the most popular research topics in the articles as well as help rank the most commonly investigated drugs or genes based on each topic. Our DSTM can also effectively help scientists query COVID-19 relevant drugs and genes based on their research topics, or search relevant COVID-19 research topics based on specific drugs and genes.

4 OnTimeEvidence COVID-19 case study

Herein, we demonstrate the utility of the OnTimeEvidence platform for COVID-19-related data analytics in the form of a case study. As part of the data access management in OnTimeEvidence, we created multiple user profiles, assigned profiles to users, and ran tests related to data access based on the defined profiles. As shown in Fig. 7, we validated basic security checks, e.g., the web interface for the access request form to create user credentials with necessary validations. Thus we ensured that we prevent SQL injection attacks and fake malicious user(s) creation. We remark that the web

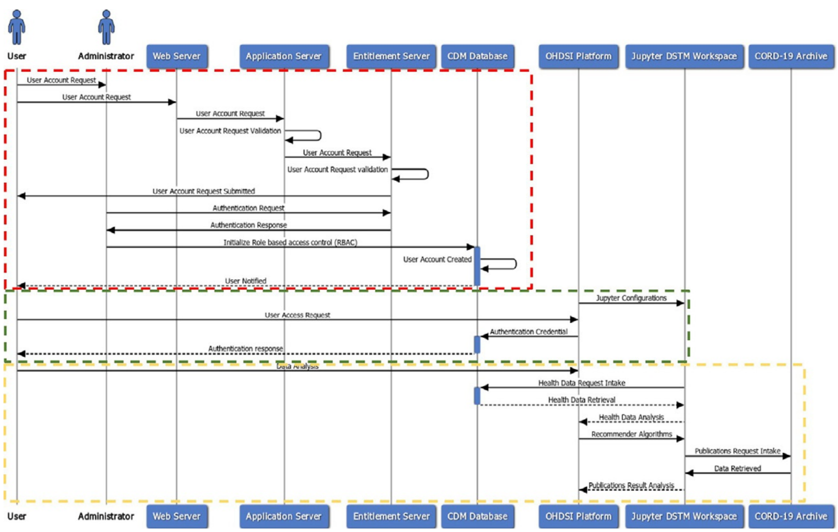

Fig. 7 Sequence diagram of the OnTimeEvidence steps to allow secure access to the analytics workspaces.

interface can be configured to be more secure by having a user community domain in the AWS certificate manager (ACM) service.

4.1 Secure access

The OnTimeEvidence web interface-related files are hosted on a S3 bucket with security options to allow exclusive subdomains accessibility. Traffic can be routed from our custom domain to this S3 bucket on a SSL channel to securely transfer user data and facilitate user communications. The filled user data in the forms are stored in an user entitlement database server (built with MongoDB). The entitlement dB is on a three-cluster sharded server hosted on AWS for high availability, and has services available such as field level encryption to encrypt the data at rest. Sharding is also useful for improving responsiveness and capacity of the entitlement dB. *Nodejs* is used in the application server that connects the OnTimeEvidence web interface (created using *Angularjs*) to the entitlement dB. The database credentials are hidden on the application server and are not accessible to nonauthorized users. Further, we use the *bcryptjs* library to encrypt the administrator's password so that the password is secured even at rest in the entitlement dB, i.e., a third party having access to the encrypted administrator password will not be able to access the data. We connect to the CDM on AWS Redshift through our *nodejs* powered application server. Such a setup allows us to run database query commands on the CDM model to create users with requested credentials and then assign them into certain groups (associated with their roles) for access on certain CDM schema elements in the entitlement dB (Figs. 8 and 9).

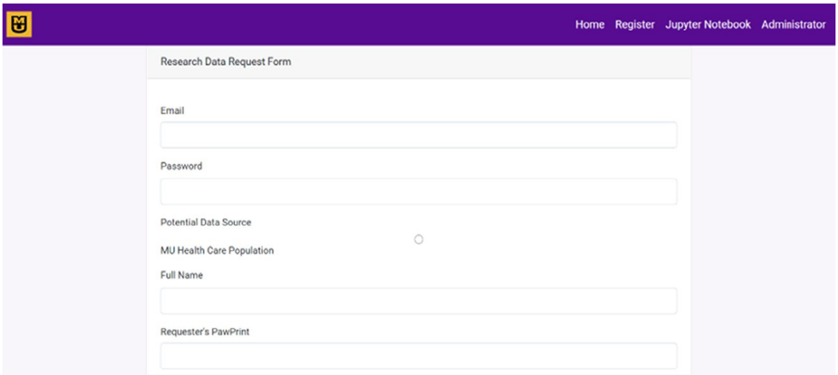

Fig. 8 Administrator web interface to authenticate, authorize, and manage users in the data processing pipeline.

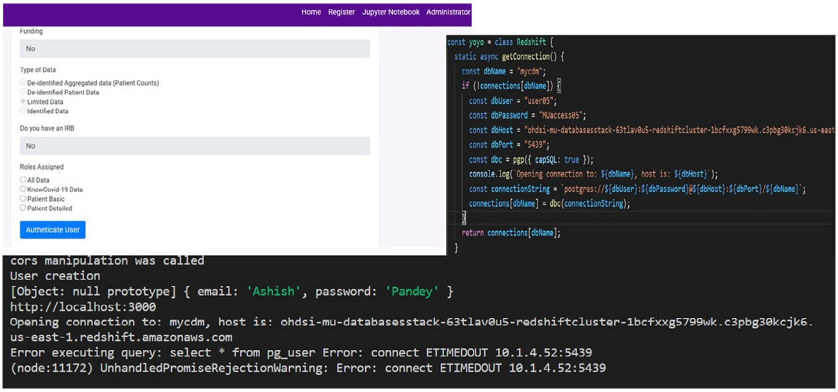

Fig. 9 Role assignment process for restricting and/or allowing users to access functionalities of the data processing pipeline.

4.2 System login and data request

Once the users are provisioned with the proper access roles, a researcher can access the OHDSI environment with the assigned OnTimeEvidence platform credentials. By default, the user will be taken to the JupyterLab environment where data request forms are available (as shown in Fig. 10) to

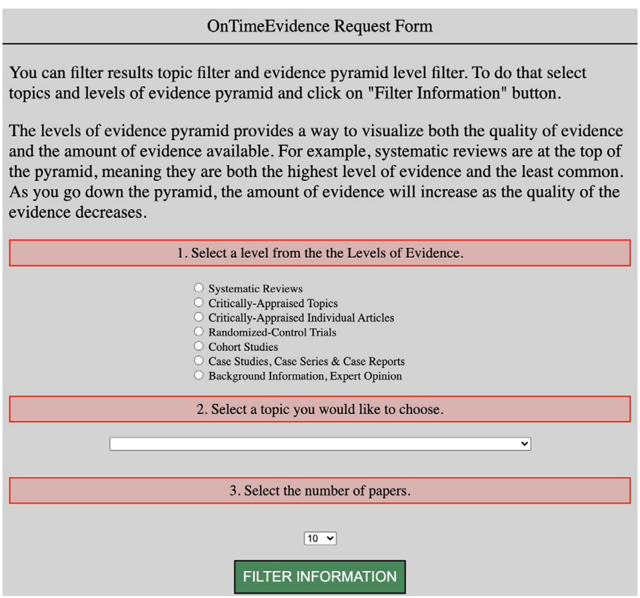

Fig. 10 Data access interface form integrated in the JupyterLab workspace for healthcare data consumers.

allow the submission of a particular CORD-19 request or SynPUF healthcare data request. Based on the user's role and the particular data request elements submitted by the researcher, the access management module, integrated as an extension to the OHDSI platform in OnTimeEvidence, will evaluate if the data elements requested are authorized for access by the related role's permissions. Accordingly, it will either authorize or deny the request. Upon authorization of the data request, the user will be able to retrieve and utilize the data within a Jupyter Notebook and employ the analytic tools available in the workspace using the two configuration modes described below.

4.3 OnTimeEvidence analytics workspace

Once the health data request is received, the platform provides the health consumer Jupyter workspace where the user can conduct an analysis over the SynPUF data stored in the CDM or the CORD-19 dataset stored on an Amazon Redshift cluster. In the following, we detail the process in which researchers can configure their workspace analysis through SynPUF data and DSTM embedded in the Jupyter Notebook for conducting a real-time analysis over publication archives.

In the first configuration mode, the researcher uses a Jupyter Notebook to retrieve and analyze patients' data from the SynPUF dataset. Such a configuration allows researchers to fill a data request, define population selection criteria elements (i.e., patient's year of birth, gender, race, diagnosed conditions, drug treatments), and the data domain elements to be extracted (conditions, drugs, procedures) for the required dataset. Upon submission, the request is processed by a customized JavaScript embedded in the request form, and the data elements included in the request are interpreted and a SQL statement is generated based on the standard schema of the CDM. The SQL statement is then embedded into a new Jupyter Notebook and this notebook instance is made available to the user within the JupyterLab environment. The user can subsequently use the Jupyter Notebook to execute Python code that runs the SQL statement against the CDM repository and retrieves the related dataset. With the data available in the Jupyter Notebook, users can use the analytic Python libraries to run data analyses. Results from this process can be saved in the Jupyter Notebook for further analysis or can be shared with collaborators. Further, the Jupyter Notebook can also be used in conjunction with results generated by the CORD-19 configuration described below to find correlations and gain insights.

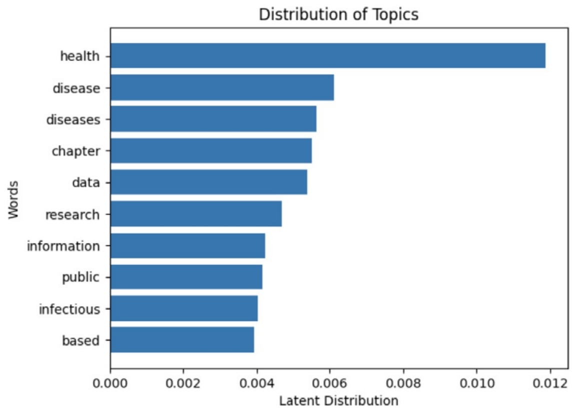

Fig. 11 Distribution chart of topics generated by the DSTM tool from the CORD-19 dataset; this chart is part of the visualizations presented in the analytics workspace.

The second configuration mode includes integrating the DSTM in the Jupyter Notebook to conduct an experiment over the CORD-19 data. Such a configuration allows health-care data consumers to filter their queries according to the Levels of Evidence Pyramid structure to obtain high-quality information from publication archives. The OnTimeEvidence data request form for COVID-19 allows users to select a level (e.g., background information to systematic reviews and meta analyses) and choose a topic from the DSTM that generates a Dirichlet distribution of words within each latent topic that was observed. Once the topic and level of choice are selected by data consumers in the request form, they are used as query parameters on the Jupyter Notebook. Fig. 11 shows a distribution of words for a given topic along with the frequency of that word in the CORD-19 dataset. Health-care data consumers can thus leverage this information to find the latest trends among topics that are pertinent to their research tasks related to the pandemic response.

5 Conclusion—What we have learnt?

Cloud-based platforms are critical for sharing and analyzing the rapidly increasing COVID-19 datasets in a scalable, standard, and secure manner, while also utilizing AI-based tools to automate the data pipeline processing for health-care data consumers (e.g., immunologists, clinical researchers). Our proposed OnTimeEvidence is an exemplar and leverages the OHDSI on the AWS environment to provide users with a scalable platform integrated with a standards-compliant data repository integrated with AI-based data analytics tools. To comply with the privacy requirements for health-care data, we adopted a role-based access control and authorization implementation to define and limit the access to the proper level of data for each user. To expedite the data request processing, OnTimeEvidence includes a data request form that guides users to select the data domain and data identifier elements to retrieve a particular dataset from the OHDSI repository required for a research task. To reduce the randomness found in existing approaches used to extract relevant information, we integrated the DSTM tool that uses the Gibbs sampling algorithm internally to generate a reliable set of results related to COVID-19 publication analytics. Consequently, OnTimeEvidence helps users to discover relationships, e.g., between drugs and genes within a large text corpus of medical research journals. Further, our OnTimeEvidence helps users to customize Jupyter workspaces included in the OHDSI deployment to perform COVID-19-related data retrieval and drill-down analytics to rapidly respond to the pandemic response issues.

References

Abdel-Basset, M., Chang, V., Hawash, H., Chakrabortty, R. K., & Ryan, M. (2021). Fss-2019-ncov: A deep learning architecture for semi-supervised few-shot segmentation of covid-19 infection. *Knowledge-Based Systems*, *212*, 106647.

Abdel-Basset, M., Chang, V., & Mohamed, R. (2020). 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*, *95*, 1066423.

Abdel-Basset, M., Chang, V., & Nabeeh, N. A. (2020). An intelligent framework using disruptive technologies for covid-19 analysis. *Technological Forecasting and Social Change*, *163*, 120431.

Ashraf, M. U., Hannan, A., Cheema, S. M., Ali, Z., Alofi, A., et al. (2020). Detection and tracking contagion using IoT-edge technologies: Confronting COVID-19 pandemic. In *2020 international conference on electrical, communication, and computer engineering (ICECCE)* (pp. 1–6). IEEE.

Barik, R. K., Dubey, H., & Mankodiya, K. (2017). Soa-fog: Secure service-oriented edge computing architecture for smart health big data analytics. In *2017 IEEE global conference on signal and information processing (GlobalSIP)* (pp. 477–481). IEEE.

Bayat, A., Szul, P., O'Brien, A. R., Dunne, R., Hosking, B., Jain, Y., et al. (2020). Variant-spark: Cloud-based machine learning for association study of complex phenotype and large-scale genomic data. *GigaScience*, *9*, giaa077.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Borton, J., Yu, A., Crego, A., Singh, A., Davern, M., & Hair, E. (2010). Data entrepreneurs' synthetic puf: A working puf as an alternative to traditional synthetic and non-synthetic pufs. In *JSM proceedings, survey research methods section*.

Cohen, A., & Nissim, N. (2018). Trusted detection of ransomware in a private cloud using machine learning methods leveraging meta-features from volatile memory. *Expert Systems with Applications*, *102*, 158–178.

Dai, B., Ding, S., Wahba, G., et al. (2013). Multivariate Bernoulli distribution. *Bernoulli*, *19*, 1465–1483.

Dinakarrao, S. M. P., Sayadi, H., Makrani, H. M., Nowzari, C., Rafatirad, S., & Homayoun, H. (2019). Lightweight node-level malware detection and network-level malware confinement in IoT networks. In *2019 design, automation & test in Europe conference & exhibition (DATE)* (pp. 776–781). IEEE.

Ekin Eren, M., Solovyev, N., Raff, E., Nicholas, C., & Johnson, B. (2020). *COVID-19 Kaggle literature organization*. arXiv e-prints, arXiv–2008.

Friedman, A. A., Letai, A., Fisher, D. E., & Flaherty, K. T. (2015). Precision medicine for cancer with next-generation functional diagnostics. *Nature Reviews Cancer*, *15*, 747–756.

García, Á. L., De Lucas, J. M., Antonacci, M., Zu Castell, W., David, M., Hardt, M., et al. (2020). A cloud-based framework for machine learning workloads and applications. *IEEE Access*, *8*, 18681–18692.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*, 5228–5235.

Hörandner, F., Krenn, S., Migliavacca, A., Thiemer, F., & Zwattendorfer, B. (2016). Credential: A framework for privacy-preserving cloud-based data sharing. In *2016 11th international conference on availability, reliability and security (ARES)* (pp. 742–749). IEEE.

Hossain, M. S., Muhammad, G., & Guizani, N. (2020). Explainable AI and mass surveillance system-based healthcare framework to combat COVID-I9 like pandemics. *IEEE Network*, *34*, 126–132.

Hripcsak, G., Duke, J. D., Shah, N. H., Reich, C. G., Huser, V., Schuemie, M. J., et al. (2015). Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Studies in Health Technology and Informatics*, *216*, 574.

Ioannidis, J. P., Salholz-Hillel, M., Boyack, K. W., & Baas, J. (2020). *The rapid, massive infection of the scientific literature and authors by covid-19*. bioRxiv.

Kricka, L. J., Polevikov, S., Park, J. Y., Fortina, P., Bernardini, S., Satchkov, D., et al. (2020). Artificial intelligence-powered search tools and resources in the fight against covid-19. *EJIFCC*, *31*, 106.

Makadia, R., & Ryan, P. B. (2014). Transforming the premier perspective® hospital database into the observational medical outcomes partnership (OMOP) common data model. *Egems*, *2*(1), 1110.

Matos, D. R., Pardal, M. L., Adao, P., Silva, A. R., & Correia, M. (2018). Securing electronic health records in the cloud. In *Proceedings of the 1st workshop on privacy by design in distributed systems* (pp. 1–6).

Milken Institute. (2020). *COVID-19 treatment and vaccine tracker*. URL: https://covid-19tracker.milkeninstitute.org.

Murad, M. H., Asi, N., Alsawas, M., & Alahdab, F. (2016). New evidence pyramid. *BMJ Evidence-Based Medicine*, *21*, 125–127.

Otoom, M., Otoum, N., Alzubaidi, M. A., Etoom, Y., & Banihani, R. (2020). An IoT-based framework for early identification and monitoring of covid-19 cases. *Biomedical Signal Processing and Control*, *62*, 102149.

Pickett, B. E., Sadat, E. L., Zhang, Y., Noronha, J. M., Squires, R. B., Hunt, V., et al. (2012). ViPR: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Research*, *40*, D593–D598.

Sackett, D. L. (1997). Evidence-based medicine. In *Seminars in perinatology* (pp. 3–5). Elsevier.

Sharma, S., Chen, K., & Sheth, A. (2018). Toward practical privacy-preserving analytics for IoT and cloud-based healthcare systems. *IEEE Internet Computing*, *22*, 42–51.

Simmhan, Y., Aman, S., Kumbhare, A., Liu, R., Stevens, S., Zhou, Q., et al. (2013). Cloud-based software platform for big data analytics in smart grids. *Computing in Science & Engineering*, *15*, 38–47.

Tuli, S., Tuli, S., Tuli, R., & Gill, S. S. (2020). Predicting the growth and trend of covid-19 pandemic using machine learning and cloud computing. *Internet of Things*, *11*, 100222.

Wiggins, J. (2018). *Create data science environments on AWS for health analysis using OHDSI*. URL: https://aws.amazon.com/blogs/big-data/creating-data-science-environments-on-aws-for-health-analysis-using-ohdsi/.

Yue, X., Wang, H., Jin, D., Li, M., & Jiang, W. (2016). Healthcare data gateways: Found healthcare intelligence on blockchain with novel privacy risk control. *Journal of Medical Systems*, *40*, 1–8.

Zhang, Y., Calyam, P., Joshi, T., Nair, S., & Xu, D. (2018). Domain-specific topic model for knowledge discovery through conversational agents in data intensive scientific communities. In *2018 IEEE international conference on big data (big data)* (pp. 4886–4895). IEEE.