JMIR Publications
The Leading eHealth Publisher

The journey of data within a Global Data Sharing Initiative: A federated three-layer data analysis pipeline to scale up multiple sclerosis research

Ashkan Pirmani, Edward De Brouwer, Lotte Geys, Tina Parciak, Yves Moreau, Liesbet M. Peeters

Submitted to: JMIR Medical Informatics
on: April 20, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript **5**
Supplementary Files **32**
Figures **33**
Figure 1 ..... **34**
Figure 2 ..... **35**
Figure 3 ..... **36**
Figure 4 ..... **37**
Figure 5 ..... **38**
Figure 6 ..... **39**
Figure 7 ..... **40**
Figure 8 ..... **41**
Multimedia Appendixes **42**
Multimedia Appendix 0 ..... **43**

Preprint JMIR Publications

The journey of data within a Global Data Sharing Initiative: A federated three-layer data analysis pipeline to scale up multiple sclerosis research

Ashkan Pirmani[1, 2, 3, 4] MSc; Edward De Brouwer[1] PhD; Lotte Geys[3, 4, 2] PhD; Tina Parciak[3, 4, 2] MSc; Yves Moreau[1*] Prof Dr; Liesbet M. Peeters[3, 4, 2*] Prof Dr

[1]ESAT-STADIUS, KU Leuven Leuven BE
[2]Data Science Institute (DSI), Hasselt University Diepenbeek BE
[3]Biomedical Research Institute (BIOMED), Hasselt University Diepenbeek BE
[4]University MS Center (UMSC), Hasselt University Diepenbeek BE
[*]these authors contributed equally

Corresponding Author:
Liesbet M. Peeters Prof Dr
Data Science Institute (DSI), Hasselt University
Agoralaan Building D
Diepenbeek
BE

Preprint
JMIR Publications

Abstract

Background: Investigating low-prevalence diseases, such as multiple sclerosis (MS), is challenging because of limited patient availability and the scattering of real-world data (RWD) across numerous sources. These obstacles impair data integration, standardization, and analysis, which negatively impacts the generation of significant meaningful clinical evidence.

Objective: This study aims to present a comprehensive, research question-agnostic, multistakeholder-driven end-to-end data analysis pipeline that accommodates three prevalent data sharing streams: individual data sharing, core dataset sharing, and federated data sharing.

Methods: A demand-driven methodology is employed for preprocessing and standardization, followed by three tiers of data acquisition, a data integration procedure, and a concluding analysis stage to fulfill RWD sharing requirements. The pipeline's effectiveness was demonstrated through its successful implementation in the COVID-19 and MS Global Data Sharing Initiative (GDSI).

Results: The GDSI has yielded multiple scientific publications and provided extensive worldwide guidance for the MS community. The pipeline facilitated the gathering of pertinent data from various sources, accommodating distinct sharing streams and assimilating them into a unified dataset for subsequent statistical analysis or secure data examination. Our pipeline contributed to the assembly of the largest dataset of people with MS infected with COVID-19.

Conclusions: The proposed data analysis pipeline exemplifies the potential of global stakeholder collaboration and underlines the significance of evidence-based decision-making. It serves as a paradigm for how data sharing initiatives can propel advancements in healthcare, emphasizing its adaptability and capacity to address diverse research inquiries.

(JMIR Preprints 20/04/2023:48030)
DOI: https://doi.org/10.2196/preprints.48030

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Preprint JMIR Publications

Original Manuscript

Preprint
JMIR Publications

Original Paper

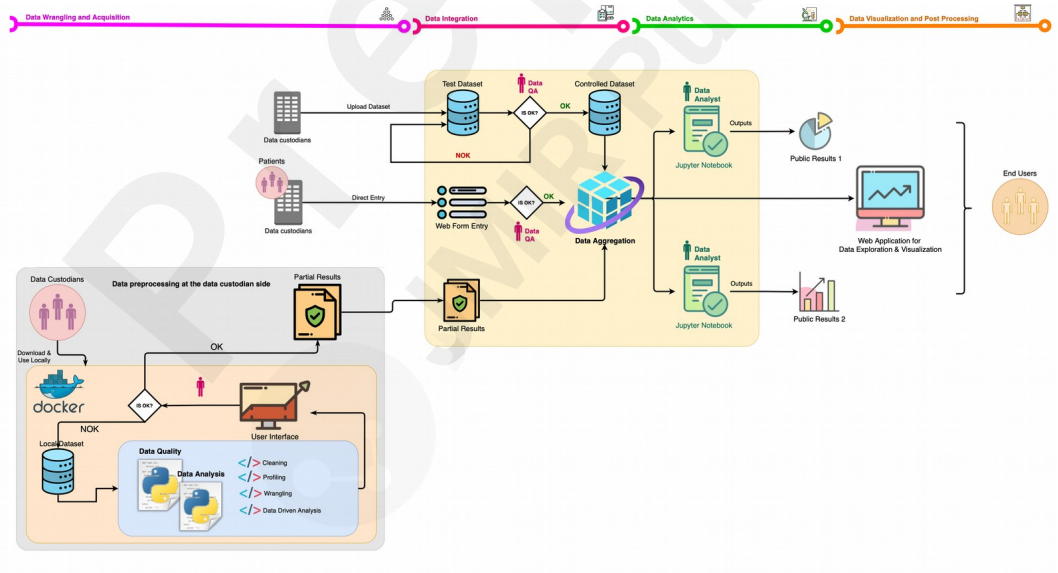
Ashkan Pirmani[a,b,c,d] MSC, Edward De Brouwer[a,e] PhD, Lotte Geys[b,c,d] PhD, Tina Parciak[b,c,d] MSC, Yves Moreau[a,*] PhD, Liesbet M. Peeters[b,c,d,*] PhD

[a] *ESAT-STADIUS, KU Leuven, Belgium*
[b] *University MS Center (UMSC), Belgium*
[c] *Biomedical Research Institute (BIOMED), Uhasselt, Belgium*
[d] *Data Science Institute (DSI), Uhasselt, Belgium*
[e] *Department of Genetics, Yale University, USA*

[*]These authors contributed equally to this work as senior authors. LMP is the corresponding author.
Liesbet.peeters@uhasselt.be; +3211269216; Agoralaan , 3590 Diepenbeek, Belgium

The journey of data within a Global Data Sharing Initiative: A federated three-layer data analysis pipeline to scale up multiple sclerosis research

Graphical Abstract



Graphical abstract: The COVID-19 and MS Global Data Sharing Initiative Analysis pipeline

Abstract

Background: Investigating low-prevalence diseases, such as multiple sclerosis (MS), is challenging because of rather small number of individuals affected by the disease and the scattering of real-world

data (RWD) across numerous data sources. These obstacles impair data integration, standardization, and analysis, which negatively impacts the generation of significant meaningful clinical evidence.
Objective: This study aims to present a comprehensive, research question-agnostic, multistakeholder-driven end-to-end data analysis pipeline that accommodates three prevalent data sharing streams: individual data sharing, core dataset sharing, and federated model sharing.
Methods: A demand-driven methodology is employed for standardization, followed by three streams of data acquisition, a data quality enhancement process, a data integration procedure, and a concluding analysis stage to fulfill RWD sharing requirements. The pipeline's effectiveness was demonstrated through its successful implementation in the COVID-19 and MS Global Data Sharing Initiative (GDSI).
Results: GDSI yielded multiple scientific publications and provided extensive worldwide guidance for the MS community. The pipeline facilitated gathering pertinent data from various sources, accommodating distinct sharing streams and assimilating them into a unified dataset for subsequent statistical analysis or secure data examination. Our pipeline contributed to the assembly of the largest dataset of people with MS infected with COVID-19.
Conclusions: The proposed data analysis pipeline exemplifies the potential of global stakeholder collaboration and underlines the significance of evidence-based decision-making. It serves as a paradigm for how data sharing initiatives can propel advancements in healthcare, emphasizing its adaptability and capacity to address diverse research inquiries.

Keywords: Data analysis pipeline; Federated model sharing; Real-world data; Evidence-based decision making; End-to-end pipeline; Multiple sclerosis

Introduction

Chronic diseases such as Multiple Sclerosis (MS) [1] present significant obstacles for research, primarily because of their limited prevalence, resulting in smaller study populations [2]. The scarcity of affected individuals is reinforced when considering the dispersion of Real-World Data (RWD) across diverse repositories. This scarce RWD, sourced during routine clinical care [3], [4], further coupled with heterogeneity in formats, quality standards, and regulatory guidelines, make the comprehensive collection and extraction of the meaningful clinical insights even more challenging [5], [6].

Despite these challenges, well-managed RWD has the potential to reveal significant patterns concerning diseases, patient experiences, and treatment outcomes [7], [8]. For instance, during the early stages of the COVID-19 pandemic, innovative data acquisition strategies overcame data scarcity, unlocking the potential of RWD for meaningful analysis [9]–[11]. These specific instances underline the broader concern: the RWD landscape is rife with challenges that are often understated.

Current literature tends to oversimplify the intricate processes involved in managing RWD. These include standardization, acquisition, quality enhancement, integration, storage, governance, visualization, and eventual analysis and interpretation. While these facets are crucial, they are often treated as isolated components rather than integral parts of an interconnected system, with certain areas occasionally overlooked [5].

Recent studies on COVID-19 bring this gap into sharper focus. Khalid and colleagues [12] focused on building analytical models using observational health data through machine learning but did not fully emphasize the vital aspect of data acquisition in the pipeline management. By contrast, Nishimwe and colleagues [13] concentrated on data integration, gathering data from various

hospitals, but did not delve into thorough in-depth data analysis. A study by Junior and colleagues [14] aimed to cover the whole data analytics pipeline but primarily focused on standardizing data from two different countries, giving less attention to crucial parts of RWD management, such as data acquisition, preprocessing, quality enhancement and analysis.
This fragmented focus points to the need for a more comprehensive strategy that neither compromises nor overlooks any part of the RWD management process. The absence of a holistic framework, coupled with the growing diversity and volume of RWD sources, intensifies the challenges in healthcare data-sharing and the conversion of RWD into actionable evidence, underscoring the need for standardized management [6].

In light of these challenges, the Global Data Sharing Initiative (GDSI) emerges as an exemplary solution that addresses multiple facets of RWD management, specifically in the context of COVID-19 and MS. Prompted by the urgent need to understand COVID-19's effects on people with MS, GDSI was launched [15]. By integrating data from over 80 countries, GDSI generated globally relevant insights [7], [16]–[20]. This large-scale effort resulted in the formation of the most extensive international cohort of COVID-19 cases among people with MS. In addition to enriching our understanding of the COVID-19 and MS interaction, GDSI showcased the enormous potential of large-scale international collaboration. Furthermore, the initiative set a methodological standard in global health research by introducing a data analysis pipeline with applications beyond MS.

This paper delves deep into GDSI's comprehensive RWD analysis pipeline, offering an end-to-end approach that spans from introducing a data dictionary to meticulous data acquisition, and ultimately, to deriving insightful clinical interpretations. One distinguishing aspect of our study lies in the pragmatic execution of this intricate, end-to-end analysis pipeline. As depicted in Figure 1, we have implemented a hybrid three-layer data acquisition architecture, all in strict compliance with the legal and ethical standards that govern data collection and dissemination. Designed for versatility and inclusivity, this architecture aimed to capture every data point possible. Concurrently, an astute approach to data integration was used, whereby these diverse data streams were seamlessly unified. This robust, unified dataset was then readied for further analysis and exploration.

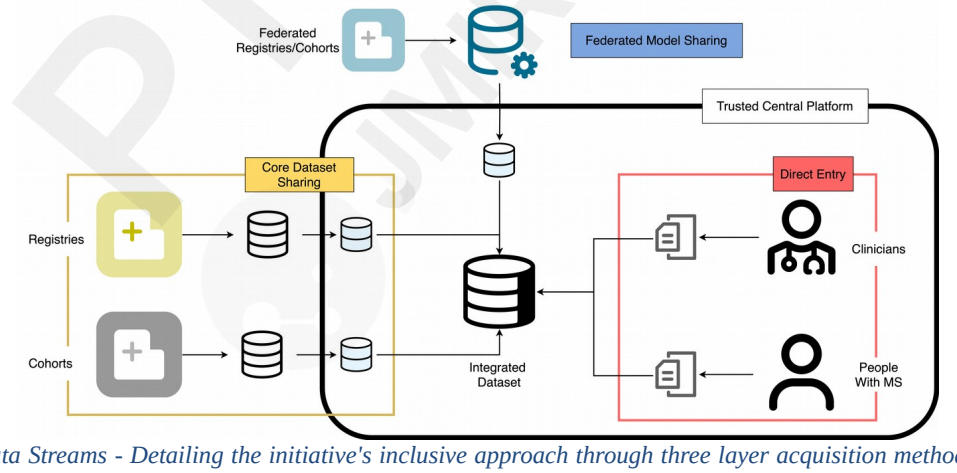


Figure 1: GDSI Data Streams - Detailing the initiative's inclusive approach through three layer acquisition methods: 1) Direct entry, where individuals upload their data via an online form; 2) Core dataset sharing, where registries upload patient-level data to the central platform under signed data transfer agreements and ethical approvals; 3) Federated model sharing, allowing registries with restrictive policies to participate without directly submitting patient-level data to the central platform.

Methods

The robustness and scale of GDSI's endeavor were mirrored by its foundational approach. As depicted in Figure 2, GDSI's RWD analysis pipeline provided the essential framework for comprehensive data management and analysis.
Centralized around a core platform, this pipeline progressed through five key stages (1) introducing a specialized data dictionary to standardize the data; (2) data acquisition, which details the methods used to gather the data; (3) an integral step for enhancing data quality; (4) data integration, responsible for aggregating various sources; and (5) the final stage, where the consolidated data is analyzed. In this section, we delve into each of these components.

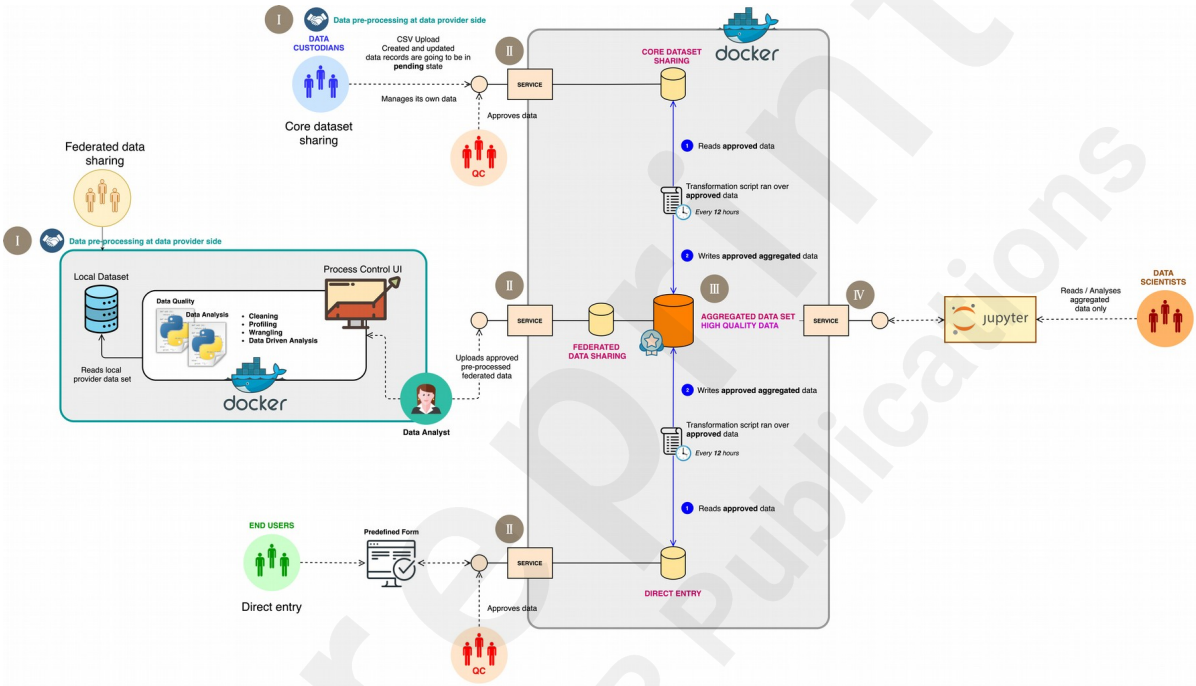


Figure 2: GDSI's end-to-end RWD analysis pipeline. Step I illustrate the standardization process, which serves as the foundation of this architecture. In this phase, data custodians are requested to map their data to the "COVID-19 in MS core dataset" (here referred to as the data dictionary). This process applies only to the core dataset and federated model sharing registries, as direct entry is already embedded with a data dictionary via the web form. Step II involves the data acquisition pipeline, featuring distinct levels of data acquisition that depend on the data holder's willingness and internal policies, all conducted in line with ethical and legal standards. Direct entry, core dataset sharing, and federated model sharing constitute the three data stream levels. The first two levels interact directly with a central platform, whereas federated registries necessitate additional steps before submitting outcomes to the central platform. To incorporate federated registries into the pipeline, predefined queries are dispatched alongside Docker containers to the local side of the registries. The results of these containers are then shipped back to the central platform. In Step III, data from different data holders are stored in separate layers to facilitate the next data integration process. Data integration, the subsequent step in the pipeline, entails consolidating data from distinct layers into a comprehensive dataset. Step IV emphasizes the use of the integrated dataset for statistical analysis, which is linked to the Jupyter Notebooks of the trusted central platform.

Data dictionary

A data dictionary serves as a guide that details the attributes of components within an information database, ensuring consistent terminology [21]–[23]. In the context of GDSI, this tool has proven invaluable for mitigating challenges posed by diverse languages and structures. A task force of domain experts, including epidemiologists, neurologists, and data scientists, reached a consensus on establishing the "COVID-19 in MS Core Dataset" data dictionary. This guide served as a keystone for harmonizing data from various sources. To tackle interoperability, data custodians used it as a reference, enabling them to standardize their datasets and streamline the extract, transform, load

process. A detailed overview of the variables employed in GDSI is provided by Simpson-Yap and colleagues [7], and a full list is accessible via the GitHub repository [24] and presented in Appendix A, Table 3.

Data acquisition

Recognizing the value of diverse data sources for research outcomes [25], [26], GDSI developed a hybrid data acquisition architecture. This framework consists of three distinct data-sharing streams: direct entry, core dataset sharing, and federated model sharing. Each stream is designed to accommodate specific data environments, ensuring a comprehensive and multifaceted collection approach. The primary distinguishing factor among these data sharing streams was the extent of willingness to share clinical records with the central server. In practice, confining data collection to a singular stream would drastically reduce the data volume, making the transition of all contributors to one mode unattainable. GDSI's strength was rooted in its adaptability, effortlessly accommodating these three sharing streams and fusing them into a unified dataset.

Direct entry: This stream prioritized direct engagement with both clinicians and patients. Patients provided their clinical records through structured questionnaires, while clinicians offered their observations after acquiring the necessary permissions. A unique characteristic of this stream was its rapid data entry mechanism. The predefined structure, designed to align with a condensed version of the data dictionary, ensured smooth data integration. Data was submitted via a web-based form on the centralized platform. Importantly, this form upheld patient privacy, excluding specific identifiers and enforcing stringent measures against website cookies and trackers.

Core dataset sharing: Adhering to the conventional approach for clinical data sharing, data providers contributed a subset of their dataset to the central platform. While this mechanism excelled in handling extensive data, it grappled with challenges related to data collaboration agreements and complex regulatory stipulations. The heterogeneity in data format further compounded these challenges. However, the architecture of core dataset sharing was designed to necessitate the schema of the data dictionary. As a result, data custodians needed to standardize their data format according to the data dictionary to upload their data using this stream. Upon achieving this congruence, custodians used the central platform's interface—a secure bridge that connected the local infrastructure of the data partners to the main platform—for data upload. For enhanced data security, once uploaded, data extraction is restricted. Additional security measures, such as user activity monitoring and stringent access policies were further implemented, ensuring registry members can only view their specific records, thus preserving data confidentiality.
As the pandemic progressed, the registries were periodically invited to contribute their core datasets to the central platform.

Federated model sharing: Addressing challenges such as strict internal policies that deterred or hindered some registries from sharing clinical records with the central platform, the federated model sharing was introduced. This decentralized solution brings a pivotal shift to regular data sharing streams. Central to this model is the principle of querying data directly at its source, thus eliminating the need to transfer patient-level data.
Instead of navigating the nuances of individual patient data, this strategy consolidates multivariable categories into aggregated “buckets”. These “buckets” are grouped categories where similar data is combined together rather than stored separately. By adopting this approach, potential risks linked to transferring patient-level data are mitigated, and the complexities tied to strict data-sharing

agreements are streamlined. A detailed examination of the “buckets” computation methodology can be viewed in Table 4 of Appendix A and within the associated GitHub repository [27].

Despite its advantages, the federated model sharing stream introduces its own challenges, especially when remote query executions result in inconsistencies across diverse systems. To compute the buckets, scripts were run locally using Docker [28] containers. Docker containers are self-contained software environments that promote standardization, which helps alleviate the typical technical challenges in such processes. These containers, referred to here as the federated pipelines, were deployed on each registry's infrastructure and were mounted with data that had already been standardized and aligned with the data dictionary, facilitating seamless execution. After computing these buckets, they were transferred to the central platform. Multiple versions of these Docker containers were used to distribute scripts across the federated model sharing registries.
The architecture of the most recent federated pipeline is presented in Figure 3. Associated resources, including a demonstrative video walkthrough, operational scripts, and the Docker image, can be found in references [29]–[32]. Furthermore, the entire source code has been made publicly accessible on GitHub [33]. This provides a thorough toolkit for those interested in understanding, replicating, or refining the framework of the federated pipeline. A comprehensive analysis of the various iterations of the federated pipeline is presented in Appendix B.

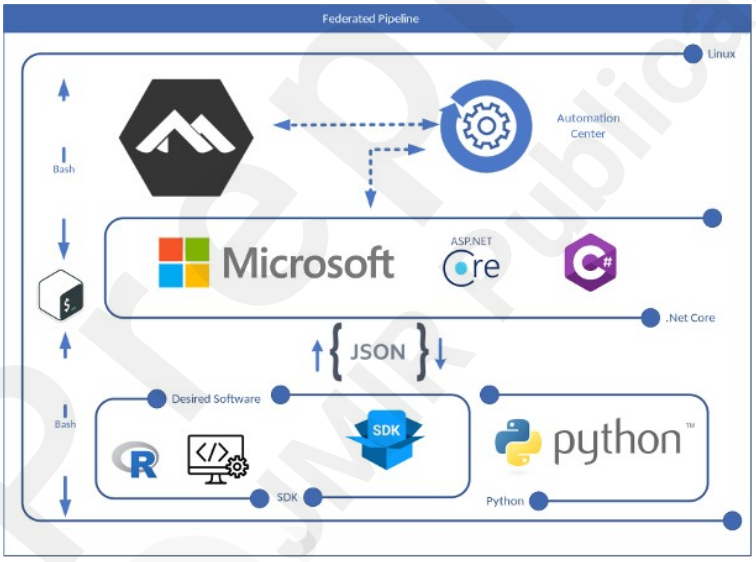


Figure 3: The Latest Federated Pipeline. The latest Federated Pipeline is a container composed of three primary components. The first component is the base image, which forms the bedrock of the infrastructure. This base image uses Alpine Linux as its underlying operating system. The remaining two components, the backend and frontend, are constructed on top of this base image. The backend consists of a suite of Python scripts, which are tasked with data quality assessment, enhancement, cleaning, and analysis. These scripts collaboratively process the incoming mapped data, preparing it for subsequent analysis. By contrast, the frontend was crafted using Microsoft's ASP.NET Core framework and the C# programming language. Within this pipeline, there is a customizable Automation Center module. This module can be adapted to meet the specific needs and requests of data partners. It also integrates Crontab, a tool that automates predefined tasks and outlines complex pipelines for execution at various intervals. The Automation Center module also links the container to the version control systems GitHub and DockerHub. This connection ensures the use of the most recent scripts and codes published by data analysts.

Data quality assessment and enhancement

The integrity of the acquired dataset was upheld through a rigorous data enhancement and quality evaluation process, which was integrated seamlessly into the central platform. In this process, each data variable was scrutinized against a binary criterion: PASS or FAIL. If a specific data point met

the pre-established benchmarks of quality and precision, it was accorded a “PASS”; otherwise, it was categorized under “FAIL”.

The input format for Direct entry eliminated the need for additional quality checks, as validation was directly integrated into the online form. For the Core dataset sharing approach, uploaded data was immediately assessed for quality, and a real-time feedback mechanism alerted contributors to any issues, allowing for immediate corrective action.

Conversely, within the federated model sharing approach, quality checks were conducted at the data source prior to aggregation.

The criteria are summarized in Table 1. In the PASS/FAIL column of this table, variables are flagged differently according to the data quality check. PASS is the flag for an accepted variable, FAIL is the flag for a dismissed variable, and EMPTY is the flag for each variable that is *missing/null*.

Note that FAIL does not necessarily mean that the data gets excluded, just that it is flagged as erroneous—it can also be adapted in some cases for analysis. FAIL means the following action: “Set the FAIL variable to missing, flag the variable, and keep the patient entry (row)”.

Additionally, specific rules were applied to dates in the data; more specifically, dates cannot be in the future (i.e., if anydate > date reporting; then the variable is flagged as FAIL) or before a person's birthdate (i.e., if anydate YEAR < (year_reporting - age), then the variable is flagged as FAIL). The COVID-19-related dates also must be later than the MS baseline dates (onset and diagnosis). A comprehensive version of this table is available on GitHub [34].

Table 1: Data quality assessment and enhancement — PASS and FAIL criterion (some highlighted examples)

Variable	format	Interdependency	PASS and FAIL
covid19_date_reporting	YYYY-MM-DD	None	if covid19_date_reporting < 2019 then fail else pass
covid19_has_symptoms	single choice (yes/no)	covid19_sympt_fever covid19_sympt_dry_cough covid19_sympt_fatigue covid19_sympt_pain covid19_sympt_sore_throat covid19_sympt_shortness_breath covid19_sympt_nasal_congestion covid19_sympt_loss_smell_taste covid19_sympt_pneumonia	if covid19_ has_symptoms = null then check the covid19_sympt_xx for ‘yes if any covid19_symptxx = ‘yes’ then covid19 has symptoms = ‘yes’ (for the analysis data) strict: if covid19_has_symptoms = ‘no’ AND any covid19_symp_xx = ‘yes’ then fail derivation: covid19_has_symptoms is secondary to covid19_ sympt _xx if any of the single symptoms are ‘yes’ the empty(!) covid19_has_symptoms will be set to ‘yes’ and vice versa (all symptoms = ‘no’, covid19_has_symptoms is set to ‘no’)
covid19_sympt_fever	single choice (yes/no)	covid19_has_symptoms	see covid19_has_symptoms
covid19_sympt_fatigue	single choice (yes/no)	covid19_has_symptoms	see covid19_has_symptoms
covid19_admission_hospital	single choice (yes/no)	None	if covid19_admission_hospital = ‘yes’ AND covid19_confirmed_case = ‘no’ then fail
age_years	Integer	None	if age_years < 0 OR age_years > 110 then fail else pass
ms_onset_date	YYYY-MM-DD	ms_diagnosis_date covid19_suspected_onset	if (ms_onset_date > ms_diagnosis_date) OR (ms_onset_date > covid19_suspected_onset) then fail else pass
edss_value	Number (0.0, 10.0)	None	if edss_value < 0 OR edss_value > 10 then fail
type_dmt	Single choice	None	if type_dmt = null AND type_dmt_other = null AND current_dmt = ‘yes’ then fail else pass
has_comorbidities	single choice (yes/no)	None	if has_comorbidities = null AND any_com_xx = ‘yes’ then set has_comorbidities = ‘yes’ (for analysis)
+ 67 more additional checks			

Data integration

The quality checked data acquired within each stream is stored distinctly, emphasizing the discrete nature of their origins. Consequently, the challenge emerges not just from the acquisition but notably from the critical task of integrating these separate datasets. To derive comprehensive insights, there was a paramount need to coalesce these distinct datasets into a singular, unified structure. In the ensuing sections, we outline the process employed to achieve this integration and present a harmonized analytical framework.

Consider $x_i=[x_{i,1},\ldots,x_{i,K}]$ to be the list of control and response variables of patient i used in the downstream statistical analysis. N indicates the total number of patient records and K represents the number of variables of interest. For each variable type, we define a list of non-overlapping ranges $\sum_k ¿\sigma_k^1,\sigma_k^2,\ldots,\sigma_k^{j_k}$ that partition the domain of each variable into distinct categories. That is, each variable $x_{i,k}$ can be categorized into a variable $y_{i,k}$ by defining $y_{i,k}=j\equiv x_{i,k}\in\sigma_k^j$ with $j\in\{1,\ldots,j_k\}$.
The data extracted from the federated model sharing registries was then a multivariate contingency table (S) of the patient counts for all combinations of all variables—that is, $S=\{(\sigma_0,\sigma_1,\ldots,\sigma_K,c):\sigma_K\in\sum_{k,c}¿\sum_{i=1}^{N} I[x_{i,0}\in\sigma_0,x_{i,1}\in\sigma_1,\ldots,x_{i,k}\in\sigma_K]\}$, where $I[\cdot]$ is the indicator function.
This set is conveniently represented as a table by considering each element of the set as a row and the columns consisting of the different variable names and patient counts. This table was subsequently stored on the central platform. The same computation performed on the direct entry and core dataset, as the raw data was available on the central platform, resulting in their specific binned count tables.
Finally, all data sources were aggregated by combining their respective binned counts representation S. The aggregation was performed by adding the patient counts of each data source for each subgroup of variables. Then, the aggregate set was expanded into a more extended table by repeating each row several times equal to the patient count of that specific row, which resulted in a table $X\in R^{N\times K}$, with K the number of variables used in the analysis and N the number of patients.

Data analysis

Following the careful integration of a diverse dataset, the pivotal challenge lay in deriving actionable insights. The GDSI analysis pipeline was uniquely engineered to remain agnostic to specific clinical research inquiries. Its versatile design permitted any statistical analysis to be executed based on the variables outlined in the data integration table. The efficacy of this approach is exemplified by its application to various research questions, as highlighted in [7], [17], [18], [20].
The analytical approach implemented by Simpson-Yap, De Brouwer, and colleagues [7] was adopted for the purpose of this paper. A multilevel mixed-effects logistic regression was employed to analyze the aggregated data table. This was carried out to assess the association between DMT and several outcomes, including hospitalization, ICU admission, ventilation, and death, while adjusting for variables such as age, sex, MS phenotype, and disability score. The goal of this evaluation was to determine the impact of MS-specific therapies on the severity of COVID-19. This statistical model provided a fine understanding of the complex relationship between these therapies and disease outcomes.

Results

Data acquisition

Using the pragmatic three-layer approach of the GDSI, we obtained the largest cohort of people with MS with COVID-19. The data was collected from a total of 80 countries, with the top 10 contributing countries being the United States (n=3157, 28%), Australia (n=1639, 14.5%), Spain (n=949, 8.4%), Sweden (n=949, 8.4%), Germany (n=765, 6.8%), Argentina (n=525, 4.7%), Brazil (n=451, 4%), Turkey (n=424, 3.8%), Denmark (n=201, 1.8%), and the United Kingdom (n=190, 1.7%), which accounted for over 80% of the total number of records.
Via direct entry, data was collected from a total of 67 countries, with Spain contributing the largest number of records (n=758, 54%), followed by the Netherlands (n=95, 6.9%), the United Kingdom (n=80, 5.8%), the United States (n=53, 3.8%), Australia (n=40, 2.9%), and 62 other countries (n=357, 25.8%), resulting in a total of 1,383 records. Data was collected from 18 different registries worldwide. Fourteen of these participated in core dataset sharing, contributing 6,374 records. Meanwhile, four used the federated model sharing approach, contributing an additional 3,527 records. Error: Reference source not found Table 2 enumerates these data sources. Figure 4 summarizes the number of records acquired at each stream of the data acquisition pipeline. Data acquired through Direct Entry has been released and is accessible through the associated PhysioNet repository [35].

Table 2: Data acquisition summary in GDSI

Method of Data Sharing	Counts	Percentage
Direct entry	1383	12,26 %
Core dataset sharing	6374	56,49 %
Federated model sharing	3527	31,26 %
Total counts:11284		

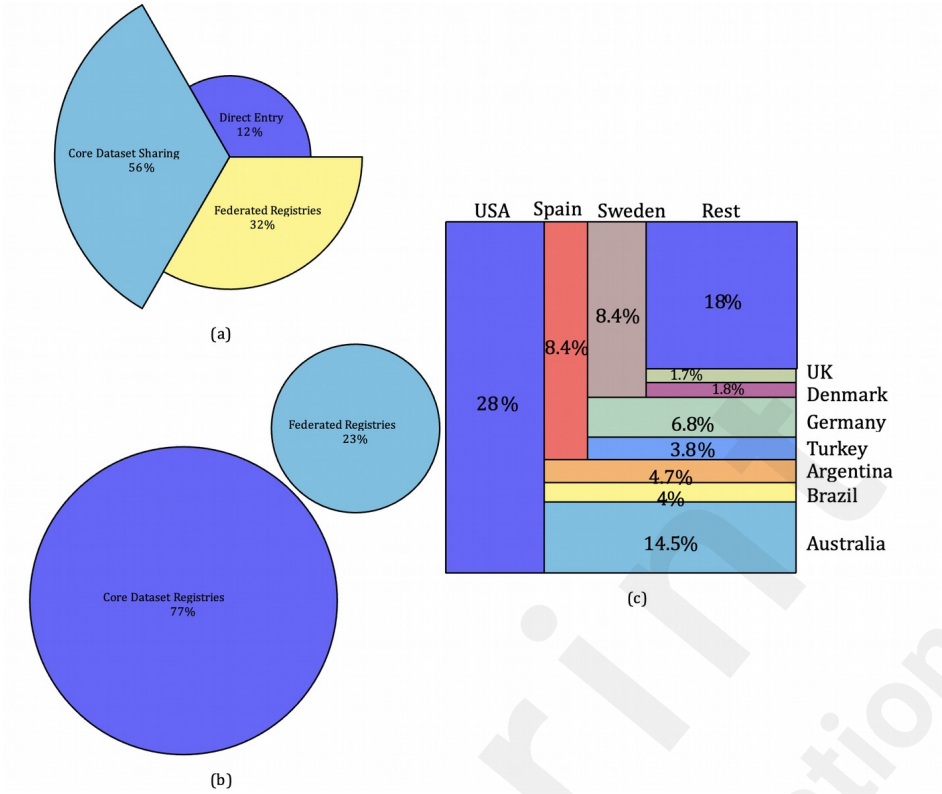


Figure 4: Summary of the data acquired by implementing the three-layer data acquisition. Federated registries (3a) contribute 32% of the data, while core dataset sharing accounts for 56%. Only 23% of the registries participated as federated registries (3b). Figure 3c provides a summary of the top 10 countries contributing data.

Data analysis

Within the data analysis conducted to assess the impact of different DMTs on COVID-19 severity [7] random effects were grouped by data sources. The following variables were used: age, sex, MS phenotype, disability score, DMTs, and COVID-19 severity. Age was categorized with the following range Σ_{age} = ((18,50),(50,70),(70,)). Sex was binarized into male and female. MS phenotype was binarized into relapsing–remitting MS and progressive MS. Disability score was dichotomized into Σ_{EDSS} = ((0,6),(6,10)). DMTs were categorized into Σ_{DMT} = (Untreated, Alemtuzumab, Cladribine, Dimethyl Fumarate, Fingolimod, Glatiramer Acetate, Interferon, Natalizumab, Ocrelizumab, Rituximab, Teriflunomide, and Other). COVID-19 severity was categorized into $\Sigma_{severity}$ = (Hospitalization, ICU admission, Ventilation, Death).
Compared to all other DMTs, patients using rituximab had a higher risk of hospitalization (aOR: 2.76, 95% CI: 1.87–4.07), ICU admission (aOR: 4.32, 95% CI: 2.27–8.23), and artificial ventilation (aOR: 6.15, 95% CI: 3.09–12.27). Ocrelizumab showed similar trends for hospitalization (aOR: 1.75, 95% CI 1.29–2.38) and ICU admission (aOR: 2.55, 95% CI: 1.49–4.36), but not ventilation (aOR: 1.60, 95% CI: 0.82– 3.14). Neither rituximab (aOR: 1.72, 95% CI: 0.58–5.10) nor ocrelizumab (aOR: 0.73, 95% CI: 0.32–1.70) were significantly associated with the risk of death. A comprehensive report of these findings can be found in Simpson-Yap, De Brouwer et al. [7].

Discussion

The COVID-19 pandemic underscored a pressing need to understand its effect on people with MS. Recognizing the criticality of solid evidence for disease management, a global strategy involving neurologists, patients, and registries was adopted. This collaborative approach paved the way for the GDSI's formation and the development of an end-to-end RWD analysis pipeline. Through this effort,

GDSI emerged as the most comprehensive federated international cohort of people with MS impacted by COVID-19, becoming an invaluable resource for informed decision-making.

Nevertheless, deriving conclusions from such data initiatives requires careful consideration of the inherent limitations of observational study designs. These studies provide unparalleled real-world insights, but it remains essential to situate the data within the confines of each study's specific limitations, especially when drawing from post hoc analyses based on existing information [36].

While GDSI showcased significant advancements, challenges inherent to its structure and execution were encountered. In this section, these challenges are delineated, encompassing aspects from data collection and analysis to concerns of interoperability, data quality, governance, data sharing, and privacy. By exploring these areas, insights are provided to optimize future initiatives and fully harness the potential of RWD in the context of global collaborative learning.

Challenges and solutions in data interoperability, quality, and governance

Interoperability and handling heterogenous data formats presented significant hurdles for GDSI. To counteract these challenges, a study-specific data dictionary was created. However, more advanced standardization methods such as a common data model [37], including FHIR [38] and OMOP [39], could further enhance standardization, making it more generalized and disease-agnostic [40], [41].

Building on the necessity for standardization, the significance of *data quality* has been universally recognized in healthcare as also highlighted by various studies [42]–[45].
In tandem with standardization efforts, GDSI integrated an automated data quality assessment framework into its data acquisition process. However, adoption of a generalized framework such as [46] can serve as a blueprint for enhancing data quality across various healthcare contexts, providing a more structured format to ensure reliability and precision.

As GDSI confronted challenges related to interoperability and data quality, the initiative also had to navigate the complex landscape of *regulatory compliance*. Implementing a federated governance model, GDSI effectively addressed the existing needs but simultaneously revealed a gap for a data governance model in healthcare—namely, the absence of implementations specifically tailored for a federated framework [47]. A more universal data governance model, such as the one proposed by Peregrina and colleagues [48], could potentially fill this gap, enhancing both organizational efficiency and the quality of analytical models.

Embracing federated model sharing and privacy concerns

While federated model sharing offers a unique approach to draw insights from patient-level data, it is worth acknowledging that even the impersonal shared statistics inherently encode some information [49]. However, these potential risks are managed under the strict supervision of GDSI, which operates within a rigorously regulated and controlled environment with trusted partners. To further mitigate risks, the data custodians in the federated model sharing underwent a formal assessment of privacy risks after running the script and before sharing the aggregated data with the central platform. This additional layer of scrutiny ensured that any potential privacy concerns were addressed prior to data dissemination.

Potential risks and their mitigation strategies were transparently communicated to all data providers via a clear analysis plan, striking a robust balance between efficient data use and strict privacy and security standards.

Although GDSI's federated model sharing has proven successful, it falls short in one crucial area: iterative asynchronous communication. This oversight leads to the introduction of “federated learning” [50], a methodology wherein a machine learning algorithm extracts knowledge from a variety of locally stored data without the need to transfer raw data enabling deploying sophisticated analysis [51].
Nonetheless, it is vital to recognize the associated risks and challenges. Federated learning or in general federated model sharing is not invulnerable to attacks [52], [53] or privacy breaches [49].

Considering these risks, it might be necessary to reevaluate the GDSI's current assumptions of trustworthiness, inquisitiveness, and non-antagonistic behavior among all participants for a wider scope of application. Incorporating privacy-preserving algorithms, such as differential privacy [54] and homomorphic encryption [55], can bolster security measures, though potentially affecting analytical performance or necessitating extensive computational resources [56]. Despite these challenges, federated learning has shown promise in a range of studies [57]–[61]. However, most of these analyses were tailored to specific use cases. There remains a need for a more generalized federated learning pipeline that can be applied broadly, rather than being limited to project-specific applications [62] .

Recognizing the inherent risks in the federated approach, GDSI took proactive steps to ensure privacy and build trust within the entire pipeline. In response to concerns regarding *privacy* and *tool reliability*, GDSI adopted privacy-by-design principles and utilized certified toolboxes that underwent third-party verification. This approach emphasizes the continual need for assessment and evaluation of privacy safeguards.

Enhancing collaboration: user engagement in the GDSI pipeline

As GDSI delved deeper into privacy and security measures, it became evident that an improved user experience was pivotal for the pipeline's success. The intricate nature of the RWD analysis pipeline, coupled with its limited visualization capabilities and an initial oversight in stakeholder inclusion, gave rise to a “black box” perception. Recognizing the urgent need for better communication and more user-friendly tools, GDSI instituted a dedicated task force. This team took charge from the study's inception through to the formulation of evidence-based guidelines, guaranteeing that every stage aligned with the multifaceted needs of all stakeholders. By doing so, GDSI not only fostered trust and collaboration but also strongly resonated with the project's overarching principles of engagement and transparency.

The deployment of GDSI's user-centric interactive web application, complemented by detailed documentation and illustrative visuals, helped demystify the pipeline's complexity. By offering accessible and user-friendly tools, this approach fostered a more nuanced stakeholder engagement, bridging the divide between intricate operations and approachability. The effectiveness of visualization in healthcare is supported by various studies [63]–[66]. Tools like Jaspersoft [67], Tableau [68], Looker [69], Domo [70], Tibco Spotfire [71], and Power BI [71] offer a business-level data analytics platform, underscoring the significance of converting intricate datasets into comprehensible visuals.

Pragmatism in GDSI: balancing innovation and adaptation

In the conceptualization and development of GDSI, striking a balance between advanced innovation and practical inclusivity was paramount. This principle was clearly manifested in the design of the data acquisition architecture. Typically, healthcare frameworks gravitate toward a federated or centralized model. Yet, GDSI embraced a hybrid strategy, seeking to cater to a broad spectrum of users and registries. This novel approach marked a significant departure from the norm, merging technological advancement with operational flexibility.

However, with innovation come challenges. Although GDSI's analysis pipeline presents a viable technical solution for collaborative healthcare learning, it also grappled with broader societal challenges. One salient example was the containerization strategy. Initially promising, it met resistance from certain federated model sharing registries because of their internal policies. Such challenges underscore the ever-present demand for adaptability amidst rapid technological shifts. However, GDSI responded proactively, making the source code available and bolstering it with a comprehensive manual and robust support. Such measures exemplify GDSI's commitment to reconciling groundbreaking advancements with real-world constraints.

This commitment extended beyond technical challenges. The global reach of GDSI emphasized the importance of resource efficiency, especially in regions with limited internet connectivity. In striving for a global impact, GDSI reiterated its pledge to balance technological progress with practical considerations across diverse geographies.
In light of these experiences, one thing becomes clear for the success of initiatives like GDSI: continuous education, proactive stakeholder engagement, and evidence-based demonstrations in controlled environments are not just beneficial—they are essential.

GDSI as a blueprint for data sharing initiatives in biomedical research

GDSI emerged as a pragmatic blueprint for interdisciplinary biomedical research. The meticulous planning and systematic execution of the initiative showcased how strategic processes can serve as foundational guides for upcoming biomedical consortia. The open-source resources GDSI provided [24], [27], [30], [31], [33]–[35], [72]–[74] can be directly leveraged and adjusted after thorough assessment and evaluation. These resources bifurcate into two main categories: disease-agnostic and disease-specific components.

Within the context of disease-agnostic components, architecture of GDSI's end-to-end data analysis pipeline stands out, highlighting its modularity and adaptability. This pipeline's design facilitates significant customization, catering to various data acquisition streams. The hybrid nature of the data acquisition module allows initiatives to choose one or a combination of data collection methods based on their distinct needs and policies. Additionally, GDSI's data integration framework plays a crucial role in amalgamating these diverse data sources into a unified and comprehensive dataset. Together, these components offer a versatile foundation that other biomedical initiatives can adapt and leverage according to their specific requirements.

Turning to disease-specific components, aspects like the data dictionary and data quality assessments were designed primarily for the research question centered around MS and COVID-19. Even though these components are specialized, they act as guiding principles for other research ventures. The data dictionary, augmented by its metadata, provides a robust foundation for the next phases of the pipeline. It offers a detailed account of acquisition variables and sets clear data quality criteria. A

significant point to note is that the data dictionary aids in determining the rules for data quality assessments, presenting a methodical approach to data validation. This thorough approach emphasizes the importance of precise planning and specificity when delving into disease-focused research questions, setting an example for other initiatives to follow.

To conclude, the flexibility and adaptability inherent in GDSI's comprehensive data analysis pipeline, coupled with its disease-specific components, meld to present a versatile tool for crafting sturdy data architectures across a spectrum of biomedical research landscapes. Those seeking a deeper understanding and guidance on harnessing and replicating GDSI's capabilities are directed to Appendix C. This section offers a detailed roadmap based on GDSI's experiences and insights, a flowchart tracing the initiative from its inception to its research question resolutions, and guidance on replicating GDSI's federated model sharing infrastructure.

Conclusion

GDSI had a substantial impact that extended beyond its initial focus on COVID-19 and MS. It contributed to numerous scientific publications and played a pivotal role in shaping global guidelines for the MS community [7], [9], [17]–[20]. This underscores the vast potential of data-driven, collaborative efforts to yield improved healthcare outcomes.

A cornerstone of GDSI's success was its RWD analysis pipeline. Crafted to navigate technical, epidemiological, and sociological challenges, this pipeline facilitated the seamless integration of varied data streams into a single dataset. This cohesive strategy enabled large-scale collaborative research and offered the flexibility to accommodate the diverse policies, regulations, and needs of various data providers.

Serving as a practical blueprint, GDSI addressed not only current healthcare challenges but also laid the groundwork for future initiatives. Its hybrid approach to data acquisition and analysis provided a scalable framework applicable to other healthcare sectors. In doing so, GDSI stands as a compelling example of how data sharing and collaborative learning can meaningfully advance healthcare research, going beyond the specific challenges of MS and COVID-19.

Acknowledgments

The author(s) have disclosed that they received financial support for the research, authorship, and/or publication of this article from the following sources: the operational costs associated with this study were funded by the Multiple Sclerosis International Federation (MSIF) and the Multiple Sclerosis Data Alliance (MSDA) operating under the European Charcot Foundation. The MSDA is a global non-for-profit multi-stakeholder collaboration acting under the umbrella of the European Charcot Foundation, financially supported by a combination of industry partners including Novartis, Merck, Biogen, Janssen, Bristol-Myers Squibb and Roche. Additionally, this work was supported by the Flemish Government through the Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen program and the Research Foundation Flanders for ELIXIR Belgium. QMENTA provided the central platform, while Amazon supplied the computational resources utilized in this work. The statistical analysis was conducted at CORe, The University of Melbourne, with support from NHMRC (1129189 and 1140766). The authors wish to extend their sincere appreciation to Nikola Lazovski for his invaluable guidance and collaboration throughout the GDSI project, especially concerning the central platform. They are also profoundly grateful to Dr. Ilse Vermeulen for her unwavering support and encouragement throughout the various stages of drafting and conceptualizing the manuscript.

Ashkan Pirmani and Yves Moreau are funded by (1) VLAIO PM: Augmenting Therapeutic Effectiveness through Novel Analytics (HBC.2019.2528); Research Council KU Leuven: Symbiosis 4 (C14/22/125), Symbiosis3 (C14/18/092; CELSA - Active Learning (CELSA/21/019). Ashkan Pirmani, and Yves Moreau are affiliated to Leuven.AI and received funding from the Flemish Government (AI Research Program, FWO SBO (S003422N). Edward De Brouwer was funded by an FWO-SB grant and has no conflicts of interest to disclose. Liesbet M. Peeters has no personal pecuniary interests to disclose, other than being the chair of the MSDA, which received income from a range of corporate sponsors, recently including Biogen, Bristol Myers Squibb, Janssen Pharmaceuticals, Merck, Novartis, and Roche. Lotte Geys has no other conflicts of interest to disclose than that she is funded by the Flemish Government under the “Onderzoeksprogramma Artificiele Intelligentie Vlaanderen”. Tina Parciak has no other conflicts of interests to disclose than that she is funded by the Flemish Government under the “Special Research Fund (Bijzonder Onderzoeksfonds, BOF)”.

Author contributions:

AP played a major role in acquiring the data, designing the federated infrastructures, and drafting and revising the manuscript for content. LG and TP contributed to the manuscript drafting and revision for content. EDB contributed to the manuscript drafting and revision, played a major role in acquiring the data, and contributed to the study concept or design as well as the analysis and/or interpretation of data. YM and LMP provided overall supervision and coordination of the study and contributed to the manuscript drafting and revision for content, including writing for content.

Abbreviations

MS: multiple sclerosis
RWD: real-world data
DMTs: disease-modifying therapies
GDSI: Global data sharing initiative

References

[1] A. A. Mitani and S. Haneuse, “Small data challenges of studying rare diseases,” *JAMA Netw Open*, vol. 3, no. 3, pp. e201965–e201965, 2020.

[2] C. Walton *et al.*, “Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition,” *Mult Scler*, vol. 26, no. 14, pp. 1816–1821, Dec. 2020, doi: 10.1177/1352458520970841.

[3] “Real-World Evidence | FDA.” https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence (accessed Mar. 30, 2023).

[4] M. Burcu *et al.*, “Real-world evidence to support regulatory decision-making for medicines: Considerations for external control arms,” *Pharmacoepidemiol Drug Saf*, vol. 29, no. 10, p. 1228, Oct. 2020, doi: 10.1002/PDS.4975.

[5] V. A. Rudrapatna and A. J. Butte, “Opportunities and challenges in using real-world data for health care,” *J Clin Invest*, vol. 130, no. 2, pp. 565–574, 2020, doi: 10.1172/JCI129197.

[6] K. Hiramatsu, A. Barrett, and Y. Miyata, “Current Status, Challenges, and Future Perspectives of Real-World Data and Real-World Evidence in Japan,” *Drugs Real World Outcomes*, vol. 8, no. 4, pp. 459–480, Dec. 2021, doi: 10.1007/S40801-021-00266-3/TABLES/8.

[7] S. Simpson-Yap *et al.*, "Associations of Disease-Modifying Therapies With COVID-19 Severity in Multiple Sclerosis," *Neurology*, vol. 97, no. 19, pp. e1870–e1885, 2021, doi: 10.1212/WNL.0000000000012753.

[8] V. B. Katkade, K. N. Sanders, and K. H. Zou, "Real world data: An opportunity to supplement existing evidence for the use of long-established medicines in health care decision making," *J Multidiscip Healthc*, vol. 11, pp. 295–304, 2018, doi: 10.2147/JMDH.S160029.

[9] L. M. Peeters *et al.*, "COVID-19 in people with multiple sclerosis: A global data sharing initiative," *Multiple Sclerosis Journal*, vol. 26, no. 10, pp. 1157–1162, 2020, doi: 10.1177/1352458520941485.

[10] V. Antoniou, E. Vassilakis, and M. Hatzaki, "Is Crowdsourcing a Reliable Method for Mass Data Acquisition? The Case of COVID-19 Spread in Greece During Spring 2020," *ISPRS International Journal of Geo-Information 2020, Vol. 9, Page 605*, vol. 9, no. 10, p. 605, Oct. 2020, doi: 10.3390/IJGI9100605.

[11] C. S. Yu *et al.*, "A COVID-19 Pandemic Artificial Intelligence–Based System With Deep Learning Forecasting and Automatic Statistical Data Acquisition: Development and Implementation Study," *J Med Internet Res 2021;23(5):e27806 https://www.jmir.org/2021/5/e27806*, vol. 23, no. 5, p. e27806, May 2021, doi: 10.2196/27806.

[12] S. Khalid *et al.*, "A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data," *Comput Methods Programs Biomed*, vol. 211, Nov. 2021, doi: 10.1016/J.CMPB.2021.106394.

[13] A. Nishimwe *et al.*, "Leveraging artificial intelligence and data science techniques in harmonizing, sharing, accessing and analyzing SARS-COV-2/COVID-19 data in Rwanda (LAISDAR Project): study design and rationale," *BMC Med Inform Decis Mak*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/S12911-022-01965-9.

[14] E. P. P. Junior *et al.*, "Integrating real-world data from Brazil and Pakistan into the OMOP common data model and standardized health analytics framework to characterize COVID-19 in the Global South," *J Am Med Inform Assoc*, vol. 30, no. 4, pp. 643–655, Apr. 2023, doi: 10.1093/JAMIA/OCAC180.

[15] L. M. Peeters *et al.*, "Multiple Sclerosis Data Alliance – A global multi-stakeholder collaboration to scale-up real world data research," *Mult Scler Relat Disord*, vol. 47, 2021, doi: 10.1016/j.msard.2020.102634.

[16] S. Simpson-Yap *et al.*, "Associations of Disease-Modifying Therapies with COVID-19 Severity in Multiple Sclerosis," *Neurology*, vol. 97, no. 19, 2021, doi: 10.1212/WNL.0000000000012753.

[17] S. Simpson-Yap *et al.*, "1298Associations of DMT therapies with COVID-19 severity in multiple sclerosis," *Int J Epidemiol*, vol. 50, no. Supplement_1, Sep. 2021, doi: 10.1093/IJE/DYAB168.604.

[18] S. Simpson-Yap *et al.*, "Updated Results of the COVID-19 in MS Global Data Sharing Initiative," *Neurology - Neuroimmunology Neuroinflammation*, vol. 9, no. 6, 2022, doi: 10.1212/NXI.0000000000200021.

[19] "An update from the MS Global Data Sharing Initiative – Thank You! - MS International Federation." https://www.msif.org/news/2022/04/06/an-update-from-the-ms-global-data-sharing-initiative-thank-you/ (accessed Mar. 30, 2023).

[20] S. Simpson-Yap *et al.*, "Severity of COVID19 infection among patients with multiple sclerosis treated with interferon-β," *Mult Scler Relat Disord*, vol. 66, 2022, doi: 10.1016/j.msard.2022.104072.

[21] A. McCabe *et al.*, "Development and validation of a data dictionary for a feasibility analysis of emergency department key performance indicators," *Int J Med Inform*, vol. 126, pp. 59–64, 2019, doi: 10.1016/j.ijmedinf.2019.01.015.

[22] S. Lin, L. J. Morrison, and S. C. Brooks, "Development of a data dictionary for the Strategies for Post Arrest Resuscitation Care (SPARC) network for post cardiac arrest research," *Resuscitation*, vol. 82, no. 4, pp. 419–422, 2011, doi: 10.1016/j.resuscitation.2010.12.006.

[23] E. Moss, "The National Health Data Dictionary," *Health Information Management*, vol. 24, no. 1,

pp. 26–29, 1994, doi: 10.1177/183335839402400112.
[24] "COVID19-GDSI/Data Dictionary.pdf at main · MS-DATA-ALLIANCE/COVID19-GDSI · GitHub." https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI/blob/main/Data%20Dictionary.pdf (accessed Apr. 09, 2023).
[25] J. A. Cook and G. S. Collins, "The rise of big clinical databases," *British Journal of Surgery*, vol. 102, no. 2, pp. e93–e101, Jan. 2015, doi: 10.1002/BJS.9723.
[26] V. B. Katkade, K. N. Sanders, and K. H. Zou, "Real world data: An opportunity to supplement existing evidence for the use of long-established medicines in health care decision making," *J Multidiscip Healthc*, vol. 11, pp. 295–304, 2018, doi: 10.2147/JMDH.S160029.
[27] "COVID19-GDSI/Buckets for federated registries.pdf at main · MS-DATA-ALLIANCE/COVID19-GDSI." https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI/blob/main/Buckets%20for%20federated%20registries.pdf (accessed Apr. 11, 2023).
[28] C. Anderson, "Docker [software engineering]," *IEEE Softw*, vol. 32, no. 3, pp. 102-c3, 2015.
[29] "MDSA Federated Pipeline COV 2.1 - YouTube." https://www.youtube.com/watch?v=d-QuCNDbHKc&ab_channel=AshkanPirmani (accessed Mar. 31, 2023).
[30] "GitHub - MS-DATA-ALLIANCE/COVID19-GDSI." https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI/tree/main (accessed Apr. 09, 2023).
[31] "GitHub - MS-DATA-ALLIANCE/COVID19-GDSI2021: MSDA Toolkit For Federated Registries participating in the COVID19 in MS Global Data Sharing initiative (GDSI)." https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI2021 (accessed Mar. 31, 2023).
[32] "msdaalliance/covid19gdsi-ui - Docker Image | Docker Hub." https://hub.docker.com/r/msdaalliance/covid19gdsi-ui (accessed Mar. 31, 2023).
[33] "GitHub - MS-DATA-ALLIANCE/COVID19-GDSI2021-UI: MSDA Toolkit For Federated Registries participating in the COVID19 in MS Global Data Sharing initiative (GDSI)." https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI2021-UI (accessed Apr. 11, 2023).
[34] "COVID19-GDSI/Data quality assessment and enhancement pipeline.pdf at main · MS-DATA-ALLIANCE/COVID19-GDSI · GitHub." https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI/blob/main/Data%20quality%20assessment%20and%20enhancement%20pipeline.pdf (accessed Apr. 09, 2023).
[35] "Patient-level dataset to study the effect of COVID-19 in people with Multiple Sclerosis v1.0.0." https://physionet.org/content/patient-level-data-covid-ms/1.0.0/ (accessed Aug. 13, 2023).
[36] K. Benson and A. J. Hartz, "A Comparison of Observational Studies and Randomized, Controlled Trials," *New England Journal of Medicine*, vol. 342, no. 25, pp. 1878–1886, 2000, doi: 10.1056/NEJM200006223422506.
[37] E. A. Voss *et al.*, "Feasibility and utility of applications of the common data model to multiple, disparate observational health databases," *Journal of the American Medical Informatics Association*, vol. 22, no. 3, pp. 553–564, Mar. 2015, doi: 10.1093/JAMIA/OCU023.
[38] E. R. Pfaff *et al.*, "Fast Healthcare Interoperability Resources (FHIR) as a Meta Model to Integrate Common Data Models: Development of a Tool and Quantitative Validation Study," *JMIR Med Inform 2019;7(4):e15199 https://medinform.jmir.org/2019/4/e15199*, vol. 7, no. 4, p. e15199, Oct. 2019, doi: 10.2196/15199.
[39] N. Ahmadi, Y. Peng, M. Wolfien, M. Zoch, and M. Sedlmayr, "OMOP CDM Can Facilitate Data-Driven Studies for Cancer Prediction: A Systematic Review," *Int J Mol Sci*, vol. 23, no. 19, p. 11834, Oct. 2022, doi: 10.3390/IJMS231911834/S1.
[40] Y. Yu *et al.*, "Developing an ETL tool for converting the PCORnet CDM into the OMOP CDM to facilitate the COVID-19 data integration," *J Biomed Inform*, vol. 127, p. 104002, Mar. 2022, doi: 10.1016/J.JBI.2022.104002.
[41] R. Makadia and P. B. Ryan, "Transforming the Premier Perspective® Hospital Database into the Observational Medical Outcomes Partnership (OMOP) Common Data Model," *eGEMs*, vol. 2, no. 1, p. 15, Nov. 2014, doi: 10.13063/2327-9214.1110.

[42] Y. Chao-Gan and Z. Yu-Feng, “DPARSF: A MATLAB toolbox for ‘pipeline’ data analysis of resting-state fMRI,” *Front Syst Neurosci*, vol. 4, p. 1377, May 2010, doi: 10.3389/FNSYS.2010.00013/BIBTEX.

[43] G. A. Van der Auwera *et al.*, “From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline,” *Curr Protoc Bioinformatics*, vol. 43, no. 1110, 2013, doi: 10.1002/0471250953.BI1110S43.

[44] G. V. Uritskiy, J. Diruggiero, and J. Taylor, “MetaWRAP - A flexible pipeline for genome-resolved metagenomic data analysis 08 Information and Computing Sciences 0803 Computer Software 08 Information and Computing Sciences 0806 Information Systems,” *Microbiome*, vol. 6, no. 1, pp. 1–13, Sep. 2018, doi: 10.1186/S40168-018-0541-1/FIGURES/7.

[45] H. Chen, M. C. Lau, M. T. Wong, E. W. Newell, M. Poidinger, and J. Chen, “Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline,” *PLoS Comput Biol*, vol. 12, no. 9, p. e1005112, Sep. 2016, doi: 10.1371/JOURNAL.PCBI.1005112.

[46] K. Lee, N. Weiskopf, and J. Pathak, “A Framework for Data Quality Assessment in Clinical Research Datasets,” *AMIA Annual Symposium Proceedings*, vol. 2017, p. 1080, 2017, Accessed: Jul. 12, 2023. [Online]. Available: /pmc/articles/PMC5977591/

[47] J. A. Peregrina Pérez, G. Ortiz Bellot, and C. Zirpins, “Data Governance for Federated Machine Learning in secure Web-Based Systems,” *Actas de las Jornadas de Investigación Predoctoral en Ingeniería Informática: Proceedings of the Doctoral Consortium in Computer Science (JIPII 2021), 2021, ISBN 978-84-89867-47-5, pages 36-39*, pp. 36–39, 2021, Accessed: Jul. 30, 2023. [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=8403181

[48] J. A. Peregrina, G. Ortiz, and C. Zirpins, “Towards Data Governance for Federated Machine Learning,” *Communications in Computer and Information Science*, vol. 1617 CCIS, pp. 59–71, 2022, doi: 10.1007/978-3-031-23298-5_5/COVER.

[49] R. Nasirigerdeh, R. Torkzadehmahani, J. Baumbach, and D. B. Blumenthal, “On the Privacy of Federated Pipelines,” *SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1975–1979, Jul. 2021, doi: 10.1145/3404835.3462996.

[50] H. Brendan McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, Feb. 2016, Accessed: Mar. 31, 2023. [Online]. Available: https://arxiv.org/abs/1602.05629v4

[51] N. Rieke *et al.*, “The future of digital health with federated learning,” *NPJ Digit Med*, vol. 3, no. 1, Dec. 2020, doi: 10.1038/S41746-020-00323-1.

[52] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, “How To Backdoor Federated Learning,” *Proc Mach Learn Res*, vol. 108, pp. 2938–2948, 2020.

[53] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, “Data poisoning attacks against federated learning systems,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12308 LNCS, pp. 480–501, 2020, doi: 10.1007/978-3-030-58951-6_24/TABLES/6.

[54] C. Dwork, “Differential privacy,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4052 LNCS, pp. 1–12, 2006, doi: 10.1007/11787006_1/COVER.

[55] A. Wood, K. Najarian, and D. Kahrobaei, “Homomorphic Encryption for Machine Learning in Medicine and Bioinformatics,” *ACM Computing Surveys (CSUR)*, vol. 53, no. 4, Aug. 2020, doi: 10.1145/3394658.

[56] J. L. Raisaro *et al.*, “Protecting Privacy and Security of Genomic Data in i2b2 with Homomorphic Encryption and Differential Privacy,” *IEEE/ACM Trans Comput Biol Bioinform*, vol. 15, no. 5, pp. 1413–1426, Sep. 2018, doi: 10.1109/TCBB.2018.2854782.

[57] T. M. Deist *et al.*, “Distributed learning on 20 000+ lung cancer patients -- The Personal Health

Train," *Radiotherapy and Oncology*, vol. 144, pp. 189–200, 2020.
[58] L. Huang, A. L. Shea, H. Qian, A. Masurkar, H. Deng, and D. Liu, "Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records," *J Biomed Inform*, vol. 99, p. 103291, 2019, doi: 10.1016/j.jbi.2019.103291.
[59] A. Jochems *et al.*, "Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries," *Int J Radiat Oncol Biol Phys*, vol. 99, no. 2, pp. 344–352, 2017.
[60] J. Li *et al.*, "A multicenter random forest model for effective prognosis prediction in collaborative clinical research network," *Artif Intell Med*, vol. 103, p. 101814, 2020, doi: 10.1016/j.artmed.2020.101814.
[61] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," *Med Image Anal*, vol. 65, p. 101765, 2020, doi: 10.1016/j.media.2020.101765.
[62] A. Pirmani, E. De Brouwer, Y. Moreau, and L. M. Peeters, "Federated learning for everyone (FL4E)," *F1000Res*, vol. 12, May 2023, doi: 10.7490/F1000RESEARCH.1119405.1.
[63] D. Gotz and D. Borland, "Data-Driven Healthcare: Challenges and Opportunities for Interactive Visualization," *IEEE Comput Graph Appl*, vol. 36, no. 3, pp. 90–96, May 2016, doi: 10.1109/MCG.2016.59.
[64] J. G. Stadler, K. Donlon, J. D. Siewert, T. Franken, and N. E. Lewis, "Improving the Efficiency and Ease of Healthcare Analysis Through Use of Data Visualization Dashboards," *Big Data*, vol. 4, no. 2, pp. 129–135, Jun. 2016, doi: 10.1089/BIG.2015.0059.
[65] A. Menon, M. S. Aishwarya, A. Maria Joykutty, A. Y. Av, and A. Y. Av, "Data Visualization and Predictive Analysis for Smart Healthcare: Tool for a Hospital," *TENSYMP 2021 - 2021 IEEE Region 10 Symposium*, Aug. 2021, doi: 10.1109/TENSYMP52854.2021.9550822.
[66] G. Battineni, M. Mittal, and S. Jain, "Data visualization in the transformation of healthcare industries," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 64, pp. 1–23, 2021, doi: 10.1007/978-981-16-0538-3_1/FIGURES/4.
[67] "Reporting and embedded business intelligence software | Jaspersoft." https://www.jaspersoft.com/ (accessed Jul. 31, 2023).
[68] "Business Intelligence and Analytics Software | Tableau." https://www.tableau.com/ (accessed Jul. 31, 2023).
[69] "Looker business intelligence platform embedded analytics | Google Cloud." https://cloud.google.com/looker (accessed Aug. 11, 2023).
[70] "Discover the Domo Data Experience Platform | Domo." https://www.domo.com/ (accessed Aug. 11, 2023).
[71] "Data Visualization | Microsoft Power BI." https://powerbi.microsoft.com/en-us/ (accessed Aug. 11, 2023).
[72] "jupyter/datascience-notebook - Docker Image | Docker Hub." https://hub.docker.com/r/jupyter/datascience-notebook (accessed Mar. 31, 2023).
[73]. "NET | Build. Test. Deploy." https://dotnet.microsoft.com/en-us/ (accessed Jul. 31, 2023).
[74] "Crontab.guru - The cron schedule expression editor." https://crontab.guru/ (accessed Jul. 31, 2023).
[75] "index | Alpine Linux." https://www.alpinelinux.org/ (accessed Jul. 22, 2023).
[76] "Snyk | Developer security | Develop fast. Stay secure. | Snyk." https://snyk.io/ (accessed Mar. 31, 2023).

Appendix A

In this part, we offer a detailed analysis of the data dictionary employed within the GDSI. Additionally, our discussion broadens to encompass the architectural schema of the federated model sharing in relation to the sharing of variables, as well as the computational methods linked to this process.
By elucidating both the data dictionary and the federated model sharing schema, our objective is to impart a thorough understanding of the systematic organization, interrelationships, and manipulation techniques of data within the GDSI framework.

Error: Reference source not found A summary of the dictionary is also shown in Table 3. The variables of the data dictionary are clustered into six groups: (1) COVID-19 incidence, (2) COVID-19 severity, (3) demographics, (4) MS history and severity, (5) DMT information, and (6) comorbidities.

Table 3: Summary of the data dictionary. The collected variables are clustered in six small subsets, and five different features are collected for each variable. Variable Name is the first feature, and it explains which data was collected. Variable ID is how information is stored in the database. Data Type shows the data format, and Value states which values this feature can take. The last feature represents whether that variable was collected by the data partner or not.

Variable Name	Variable ID	Data Type	Value	Variable Collected
COVID-19 Incidence				
Date of Visit/Reporting	covid19_date_reporting	YYYY-MM-DD	Date	Yes/No
COVID-19 Symptoms	covid19_has_symptoms	Single Choice	Binary	Yes/No
Duration of self-isolation (in days)	covid19_self isolation_duration	Number	Integer	Yes/No
country COVID-19 symptoms occurred	covid19 country	Text	Text	Yes/No
+17 more fields				
COVID-19 Severity				
Admission in Hospital of COVID-19	covid19_admission_hospital	Single Choice	Binary	Yes/No
Admission date	covid19_admission_hospital_date	YYYY-MM-DD	Date	Yes/No
Ventilation needed during hospital stay?	covid19_ventilation	Single Choice	Binary	Yes/No
Did the patient receive ECMO?	covid19_ecmo	Single Choice	Binary	Yes/No
+10 more fields				
Demographics				
Age (years)	Age_years	Number	Integer	Yes/No
Sex	Sex	Single Choice	Options	Yes/No
Current Smoker	Current_smoker	Single Choice	Binary	Yes/No
Weight (in kg)	Weight	Number	Integer	Yes/No
+4 more field				
MS history and severity				
MS Type	Ms_type	Single Choice	Options	Yes/No
MS onset	Ms_onset_date	YYYY-MM-DD	Date	Yes/No
EDSS value	Edss_value	Number	Float [0.0,10.0]	Yes/No
Date of evaluation	Edss_date_diagnosis	YYYY-MM-DD	Date	Yes/No
+7 more fields				
DMT Information				
DMT current usage	Current_dmt	Single Choice	Options	Yes/No
Type of last/current DMT	Type_dmt	Single Choice	Options	Yes/No
Glucocorticoid during the past months	Dmt_glucocorticoid	Single Choice	Binary	Yes/No

DMT Start Date	Dmt_start_date	YYYY-MM-DD	Date	Yes/No
+8 more fields				

Comorbidities

Comorbidities	Has_comorbidities	Single Choice	Binary	Yes/No
Cardiovascular disease	Com_cardiovascular_disease	Single Choice	Binary	Yes/No
Hypertension	Com_hypertension	Single Choice	Binary	Yes/No
Diabetes	Com_diabetes	Single Choice	Binary	Yes/No
+7 more fields				

Table 4 illustrates how buckets were calculated for the federated registries.

Table 4: Data computed for the federated model sharing registries

Title	Variable ID	How Calculated
Diagnosis of COVID-19	covid19_diagnosis	Covid19_diagnosis= “confirmed”; if covid19_confirmed_case= “yes” Covid19_diagnosis= “suspected”; if covid19_suspected_case= “yes”
Age categories	age_in_cat	age_in_cat = 0 if 0 < age_years <18 age_in_cat = 1; if 18≤ age_years ≤50 age_in_cat = 2; if 50< age_years ≤70 age_in_cat = 3; if 70< age_years
Type of MS	ms_type2	ms_type2 = “relapsing_remitting” if ms_type= “RRMS” ms type2= “other” if – ms_type= “CIS” - or ms_type = “not_sure” - or ms_type = “” ms_type2= “progressive_MS” if - ms_type= “SPMS” - or ms_type = “PPMS”
EDSS Categories	edss_in_cat2	edss_in_cat2= “zero” if 0≤ edss ≤6 edss_in_cat2= “one” if 6< edss
DMT Type	dmt_type_overall	dmt_type overall= “No information on DMT use” if current_dmt= missing or type_dmt = missing dmt_type_overall= “currently not using any DMT” if current_dmt= “no” OR “no, but was in the past” dmt_type_overal= “currently on interferon” if current_dmt = “yes” AND type_dmt = “interferons” dmt_type_overall= “currently on glatiramer” if current_dmt = “yes” AND type_dmt = “glatiramer” dmt_type_overall= “currently on natalizumab” if current_dmt = “yes” AND type_dmt = “natalizumab” dmt_type_overall= “currently on fingolimod” if current_dmt = “yes” AND type_dmt = “fingolimod” dmt_type_overall = “currently on dimethyl fumarate” if current_dmt = “yes” AND type_dmt = “dimethyl fumarate” dmt_type_overall = “currently on teriflunomide” if current_dmt = “yes” AND type_dmt = “teriflunomide” dmt_type_overall = “currently on alemtuzumab” if current_dmt = “yes” AND type_dmt = “alemtuzumab” dmt_type_overall= “currently on cladribine” if current_dmt = “yes” AND type_dmt = “cladribine” dmt_type_overall = “currently on siponimod” if current_dmt = “yes” AND type_dmt = “siponimod” dmt_type_overall = “currently on rituximab” if current_dmt = “yes” AND type_dmt = “rituximab” dmt_type_overall = “currently on ocrelizumab”

		if current_dmt = “yes” AND type_dmt = “ocrelizumab” dmt_type_overall = “currently on another drug not listed” if current_dmt “yes” AND type_dmt other is not missing
+15 more variables		

Appendix B

During the progression of the GDSI project, three iterative pipelines were developed—Federated Pipeline COV1.0, COV2.0, and COV2.1—each building upon the features and capabilities of the previous version. These pipelines were designed to cater to the needs of federated model sharing registries, continually improving over time. In the following sections, we will first highlight the key aspects of these pipelines and then provide a detailed overview of the infrastructure, functionalities, and modules associated with them, showcasing their evolution and enhancement over time.
Key infrastructural highlights encompass the following:

1. **Federated Pipeline COV 1.0**: Containerization with employing the Jupyter Data Science Notebook's[75] official Docker image, facilitating a consistent environment for data custodians and ensuring seamless deployment across diverse computing infrastructures.

2. **Federated Pipeline COV 2.0**: Implementation of a language-agnostic base image with Ubuntu Linux as the foundation, a user-friendly web application utilizing the ASP.NET Core framework [76], and the incorporation of an Automation Center module with a job scheduler (Crontab [77]) and seamless connection to version control systems (GitHub and DockerHub).

3. **Federated Pipeline COV 2.1**: Emphasis on privacy and security enhancements through the adoption of Alpine Linux [78], as the container's base image, the execution of multiple hardening processes, and a significant reduction in the container's compressed image size.

The initial iteration, Federated Pipeline COV 1.0, aimed to tackle the complexities associated with computing infrastructure compatibility in federated model sharing. By leveraging containerization and utilizing the official Docker image of the Jupyter Data Science Notebook [75], the pipeline created a consistent, encapsulated environment for data custodians. This approach facilitated compatibility and streamlined deployment across various computing infrastructures, contributing to the pipeline's scalability and adaptability. Upon completion of the analysis, data custodians compiled the results into a compressed zip file and shared them with the data analyst team, which was then manually uploaded to the platform. This process ensured the privacy and security of patient-level data at the local data custodian level while transmitting only aggregated outcomes. Consequently, the first version of the federated pipeline was established, as illustrated in Figure 5.Error: Reference source not found

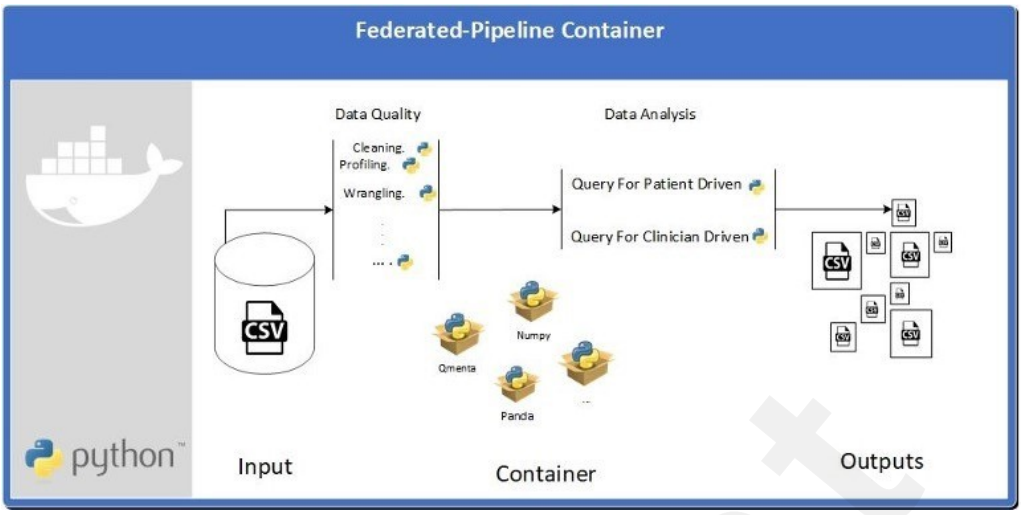


Figure 5: The federated pipeline COV 1.0 architecture.
This infrastructure encompasses a series of scripts written in Python. These scripts are divided into two parts: (1) data quality scripts that are responsible for enhancing and cleaning the data and (2) a predefined query for data analysis.

To enhance the user experience, Federated Pipeline COV 2.0 was developed. The improved architecture featured Ubuntu Linux as the underlying operating system, which enabled the creation of a web service application using the lightweight, cross-platform ASP.NET Core framework. This development offered data custodians a more transparent and user-friendly interface by incorporating different interactive infographics and plots as a result of executing the analyses, reducing the black-box approach associated with federated registries. The backend consisted of a suite of Python scripts for data quality assessment, enhancement, cleaning, and analysis. Key improvements in COV 2.0 included the customizable *Automation Center* module, which could be tailored to the needs and requests of data partners, and the integration of Crontab to automate predefined tasks and define complex pipelines for the execution at different intervals. Additionally, the automation center module connected the container to the version control systems GitHub and DockerHub, ensuring the use of the latest available scripts and codes published by data analysts. Another benefit of the refined architecture was the language-agnostic infrastructure, which allowed data analyst teams to utilize their preferred programming language, eliminating the need for translation. Figure 6 illustrates the updated architecture of the pipeline.

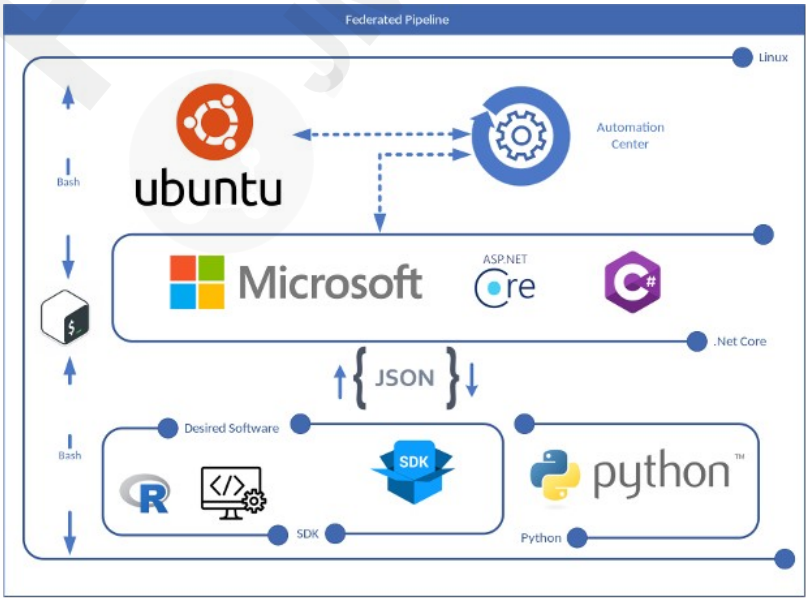


Figure 6: The Federated Pipeline COV 2.0 architecture. This container is comprised of three main components. The first component is

the base image, which serves as the foundation of this infrastructure. This base image uses Ubuntu Linux as the underlying operating system. The second and third components, the backend and frontend, are built upon the base image. The backend is composed of a series of scripts written in Python, which are responsible for data quality assessment, enhancement, cleaning, and analysis. These scripts work in tandem to process the incoming mapped data and prepare it for further analysis. The frontend, by contrast, was developed using Microsoft's ASP.NET Core framework and the C# programming language.

The final iteration, Federated Pipeline COV 2.1, placed significant emphasis on enhancing privacy and security measures because of the sensitive nature of patient-level data interaction. To achieve this, the base image of the container transitioned to Alpine Linux which is renowned for its robust security and lightweight size. Several hardening processes were implemented, including the removal of unnecessary accounts, dangerous commands, and utilities, as well as the restriction of system file and directory permissions. A third-party tool, Snyk [79], was used to evaluate vulnerabilities and privacy risks within the container. The findings of this assessment are summarized in Table 5 Error: Reference source not found.
The transition from Ubuntu to Alpine Linux, combined with the use of minimal resources, specifically in the frontend, resulted in a drastically reduced container image size from 1.31 GB in the initial version to a mere 120 MB in the latest version, representing a decrease exceeding 90%. This led to improved resource efficiency.

The design of the federated model sharing infrastructure was governed by four main principles: privacy, security, efficiency, and user-friendliness. These guiding principles ensured a balanced approach towards efficient data utilization while maintaining stringent standards of privacy, security and facilitating a user-friendly interface.

Table 5: Snyk vulnerability assessment report. This report was exported for the federated infrastructures. The counts are extracted from Snyk. Severity levels are used to evaluate the potential risk posed to an application. A high severity level indicates that an attacker may be able to gain unauthorized access to sensitive data within the application. A medium severity level suggests that under certain circumstances, an attacker could access sensitive data within the application. A low severity level means that the application may expose some data, which could be used alongside other vulnerabilities to launch an attack on the application. These numbers correspond to the privacy and security notice scanned within this platform.

	Federated Pipeline COV 1.0	**Federated Pipeline COV 2.0**	**Federated Pipeline COV 2.1**
High	113	19	0
Medium	51	12	0
Low	228	131	0
Total	**392**	**162**	**0**

Table 6 provides a high-level summary of the infrastructure used in the federated model sharing registries, along with a comparison of the different components within the infrastructure.

Table 6: Overview of the federated infrastructures deployed in the GDSI

	COV 1.0	**COV 2.0**	**COV 2.1**	**Remark**
Containerization	Docker	Docker	Docker	
Base Image	Jupyter	Ubuntu	Alpine	
Interactive Interface	Yes	Yes	Yes	ASP .NET Core web app
Interface Technology	Notebook	ASP .NET Core	ASP .NET Core	
Open-source	Yes	Yes	Yes	Under different licenses
Language agnostic	No	Yes	Yes	Executes multiple languages
Debugging Panel	Yes	Yes	Yes	Notebook interface and Linux shell
Customizable	No	Yes	Yes	Based on user requests
Version Control	No	No	Yes	Handled by Docker Hub
GitHub Repository	No	Yes	Yes	Full development process committed in the repository
Job Scheduler	No	Yes	Yes	Scheduling using Cron module
Resource Efficiency	1.31 GB	334.91 MB	120.46 MB	91% improvement

Appendix C

This section aims to offer a comprehensive guide on leveraging the GDSI framework for other research questions, emphasizing the advantages for emerging data-sharing consortia. To tailor GDSI to other disease domains, it is recommended to follow this roadmap:

1. **Establish a Taskforce:** Convene a multidisciplinary team of experts from fields such as data science, clinical research, and epidemiology to provide oversight and guidance throughout the adaptation process.
2. **Analysis Plan Formation:** Develop a robust analysis plan that outlines the study design, objectives, scope, comprehensive research questions, and expected deliverables. The analysis plan should be approved by the taskforce to ensure alignment with the initiative's goals.
3. **Data Dictionary or Harmonization Method:** Create or adapt a comprehensive data dictionary under the supervision of the taskforce. This dictionary should cover all required variables and include metadata like acceptable ranges and input types. Alternatively, integrate a sophisticated data harmonization method if a data dictionary is not feasible.
4. **Customization of Data Acquisition Modules:** Leverage the modularity of GDSI's data acquisition architecture to select, drop, or tailor specific data streams based on the project's scale, objectives, and requirements.
5. **Quality Control Adaptation:** Implement data quality assessments informed by the newly crafted or adapted data dictionary. The metadata in the data dictionary should serve as a cornerstone for defining quality control metrics. Logical controls should also be tailored to align with the objectives outlined in the analysis plan.
6. **Data Integration:** In case of using federated model sharing, use the schema of the particular data stream as the reference point for integration. If federated model sharing is not used, establish core dataset sharing, to serve as the base schema for integrating additional data streams.
7. **Data Visualization and Dashboarding:** Design real-time dashboards and reporting tools for ongoing monitoring. These tools should facilitate sanity checks among stakeholders and the taskforce to ensure the integrity of the collected data and the progress of the pipeline.
8. **Data Analysis**: Conduct data analytics as per the guidelines defined in the analysis plan. Note that the integrated dataset should allow the flexibility for unplanned, exploratory analyses.
9. **Pilot Testing**: Validate the system's performance and adaptability through pilot testing, which can be performed using available benchmark datasets or synthetic data.
10. **Review and Iteration**: Based on feedback and outcomes from the pilot testing, make any necessary refinements to prepare the pipeline for broader-scale deployment.

Figure 7 elucidates the roadmap of GDSI by outlining the critical steps from the initiative's inception to the realization of its intended outcomes.

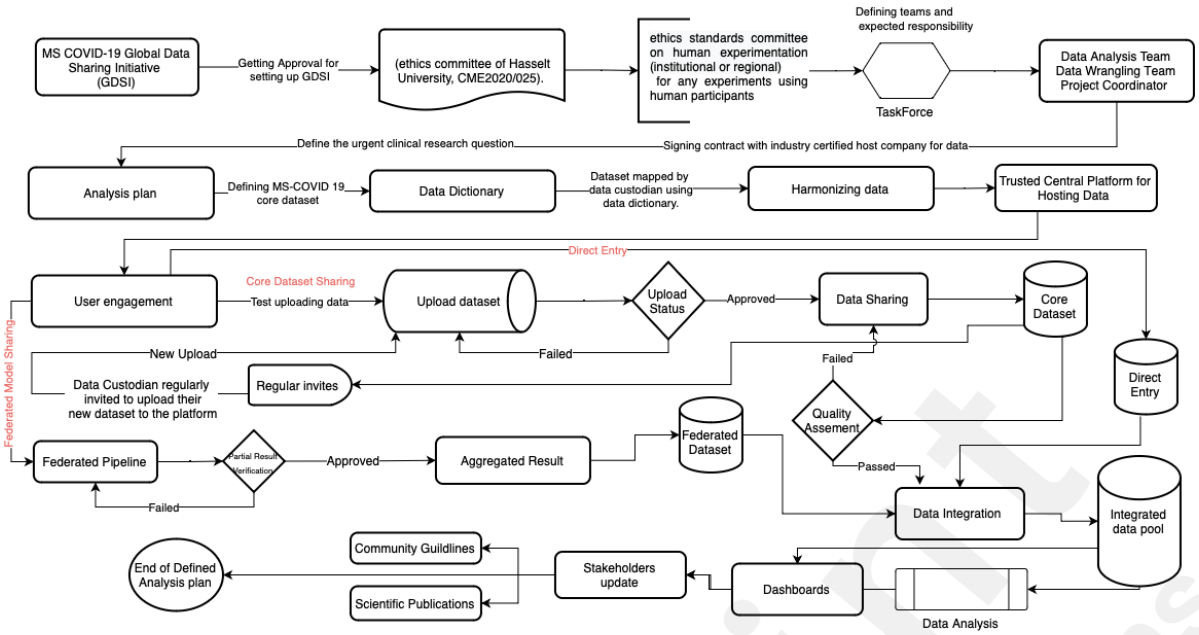


Figure 7: GDSI Flowchart: A Comprehensive Roadmap Detailing the Key Steps from Initiative Inception to Final Outcomes, Serving as a Blueprint for Adapting GDSI Principles to Other Biomedical Data Sharing Consortia

In the absence of GDSI's central platform, a federated model sharing infrastructure alongside a direct entry dataset [35] can co-create a smaller scale version of GDSI. Moreover, a tunable mock data generator [74] has been provided to create datasets of various sizes and characteristics, acting as a synthetic registry. This mock dataset can be integrated into the federated model sharing infrastructure, and the aggregated results can be used to produce a unified dataset for various analyses, as demonstrated in Figure 8.

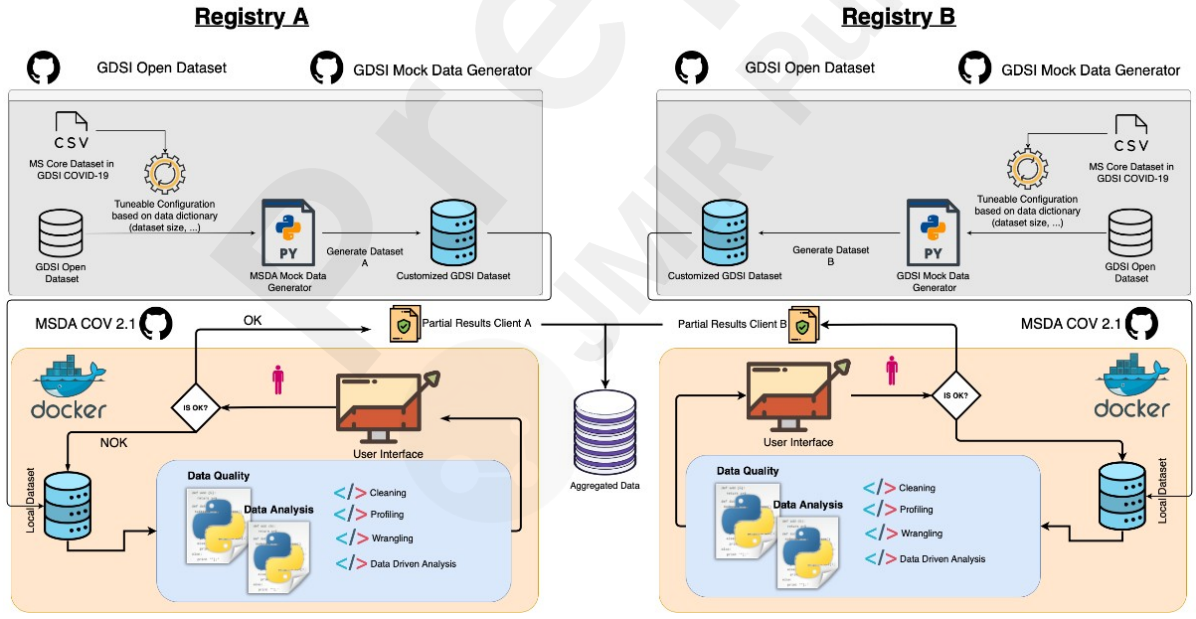


Figure 8: A Guide to Replicating Federated Model Sharing Using GDSI Principles

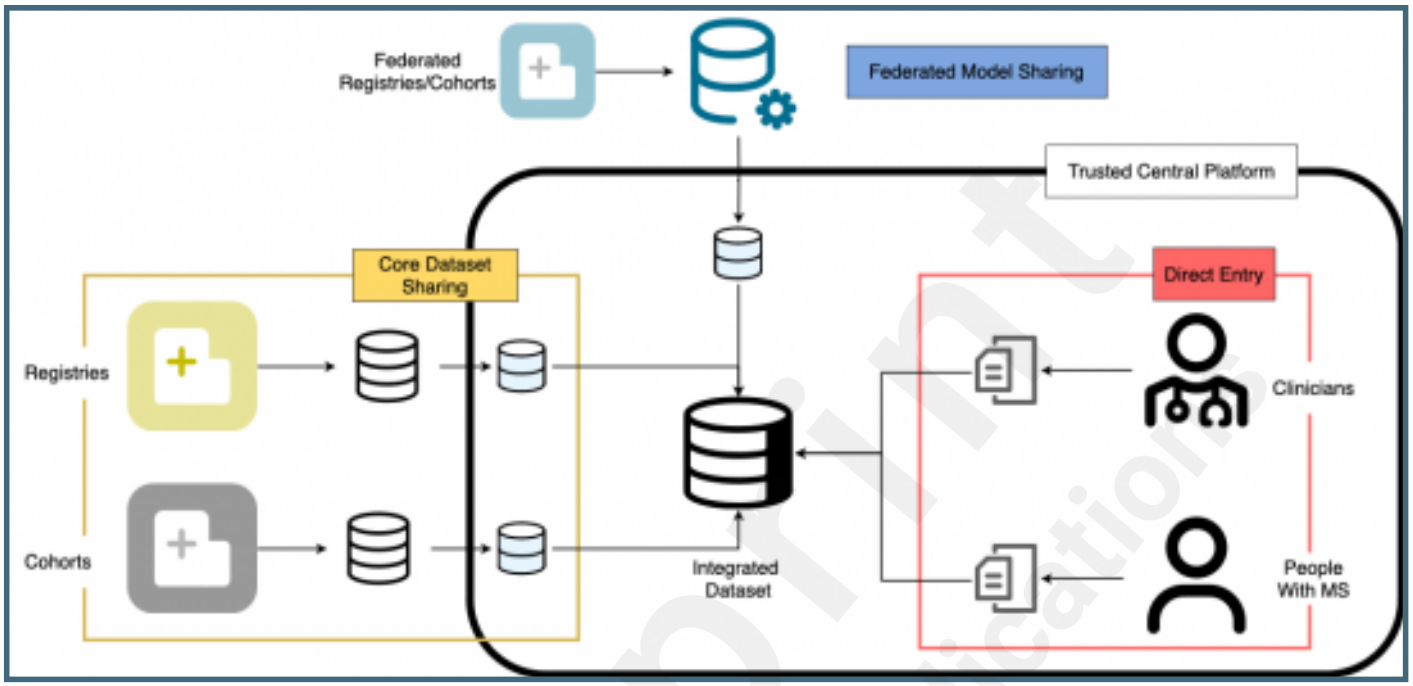
Supplementary Files

Preprint JMIR Publications

Figures

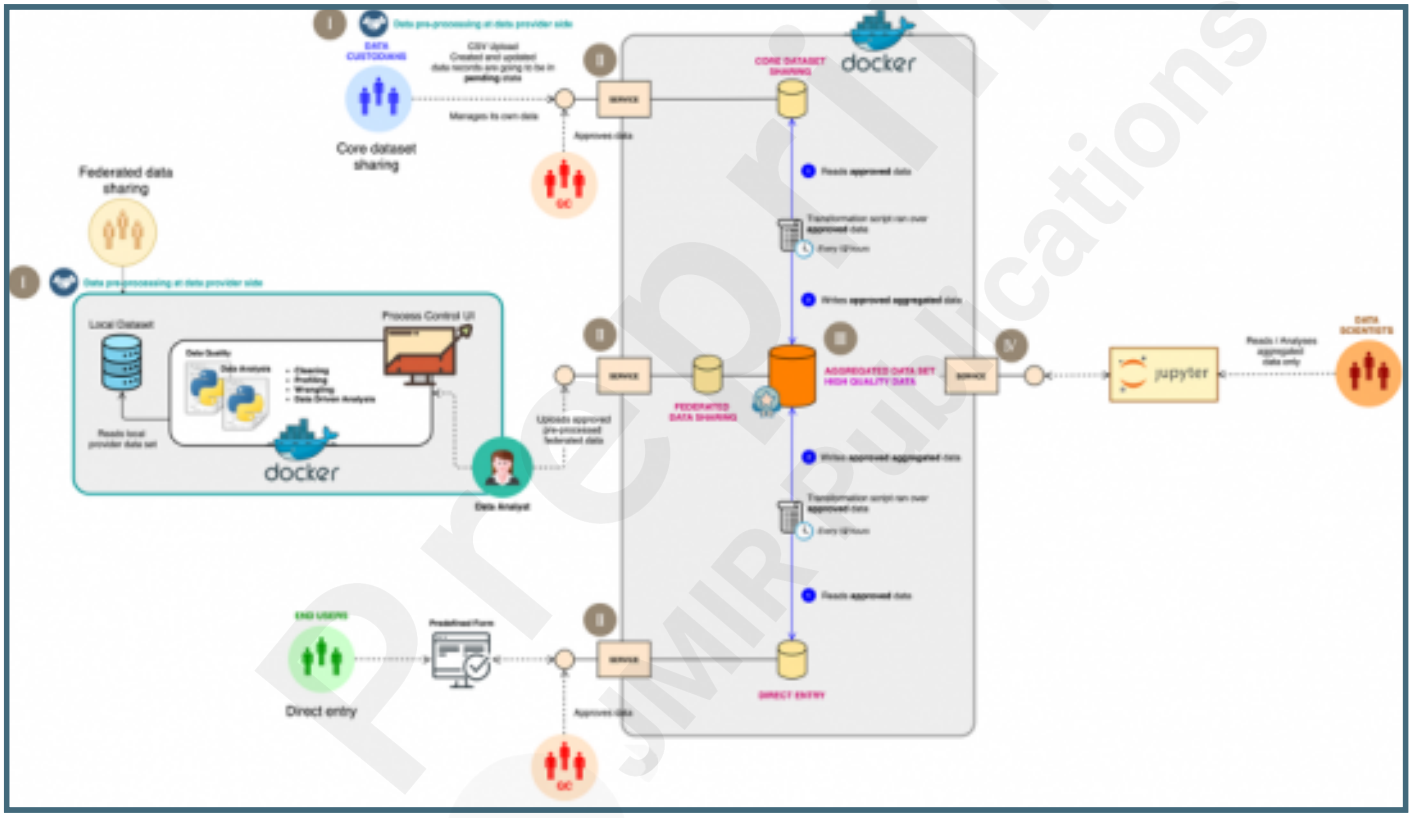
Preprint JMIR Publications

GDSI Data Streams - Detailing the initiative's inclusive approach through three layer acquisition methods: 1) Direct entry, where individuals upload their data via an online form; 2) Core dataset sharing, where registries upload patient-level data to the central platform under signed data transfer agreements and ethical approvals; 3) Federated model sharing, allowing registries with restrictive policies to participate without directly submitting patient-level data to the central platform.

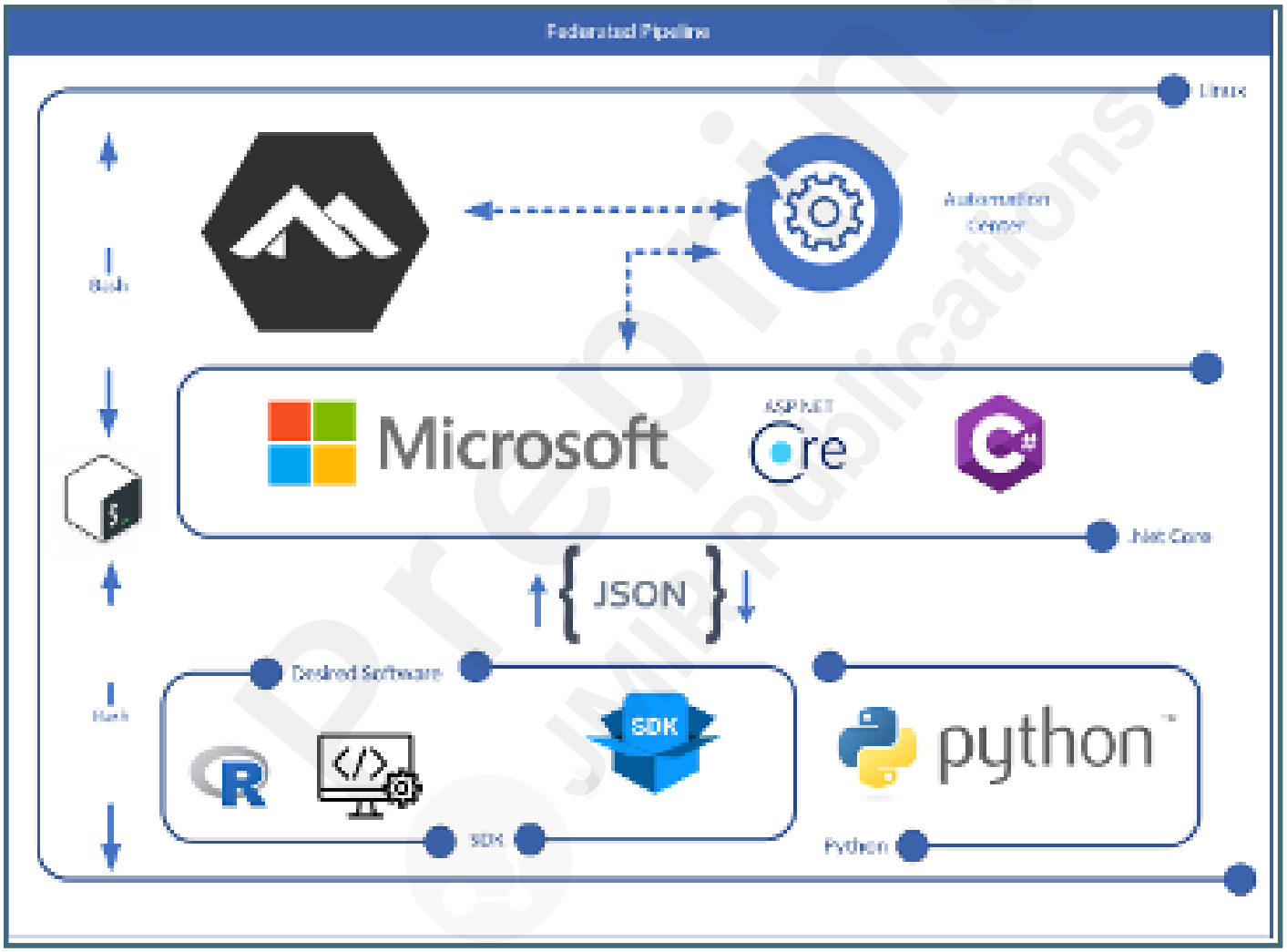


Preprint JMIR Publications

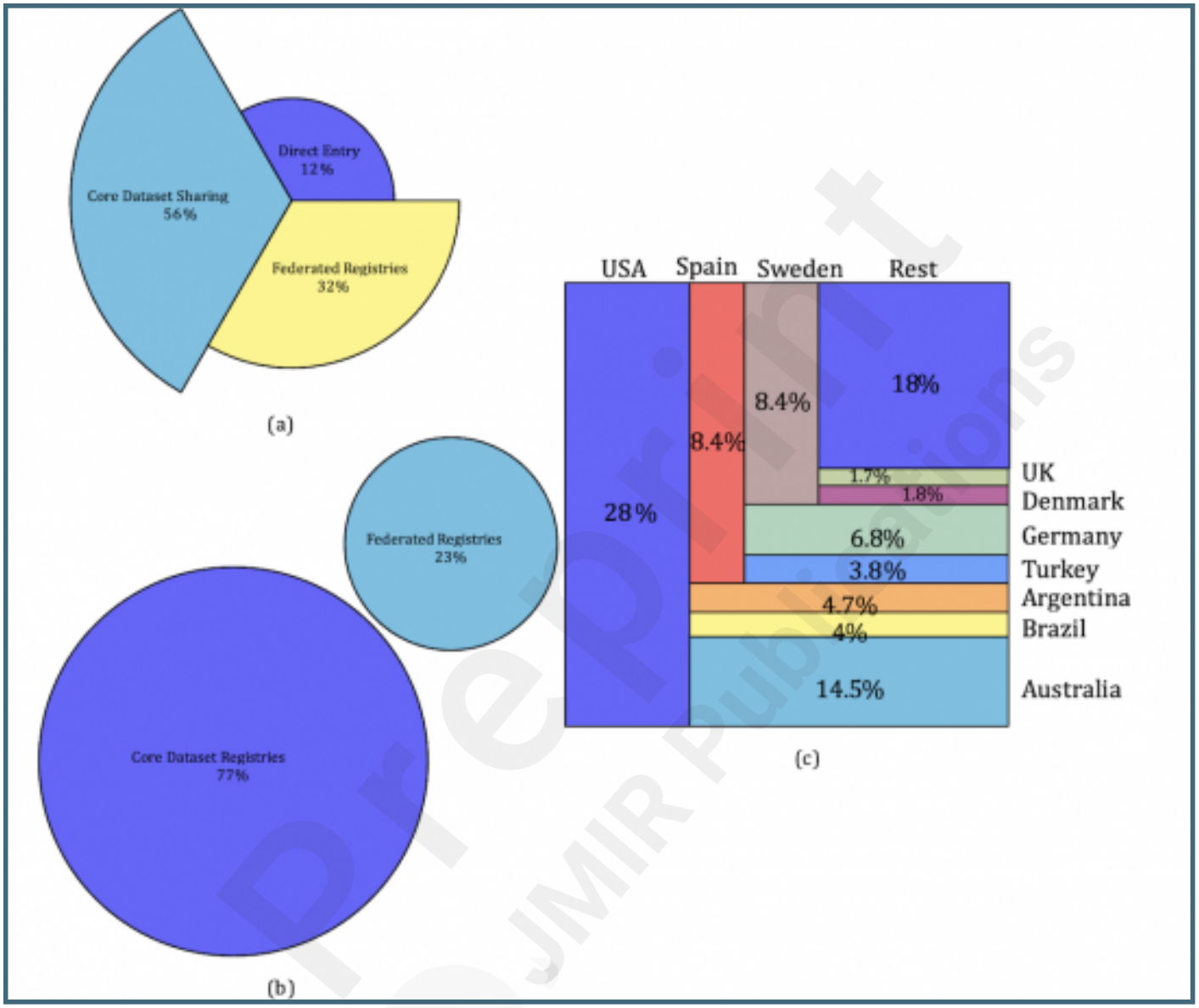
GDSI's end-to-end RWD analysis pipeline. Step I illustrate the standardization process, which serves as the foundation of this architecture. In this phase, data custodians are requested to map their data to the “COVID-19 in MS core dataset” (here referred to as the data dictionary). This process applies only to the core dataset and federated model sharing registries, as direct entry is already embedded with a data dictionary via the web form. Step II involves the data acquisition pipeline, featuring distinct levels of data acquisition that depend on the data holder's willingness and internal policies, all conducted in line with ethical and legal standards. Direct entry, core dataset sharing, and federated model sharing constitute the three data stream levels. The first two levels interact directly with a central platform, whereas federated registries necessitate additional steps before submitting outcomes to the central platform. To incorporate federated registries into the pipeline, predefined queries are dispatched alongside Docker containers to the local side of the registries. The results of these containers are then shipped back to the central platform. In Step III, data from different data holders are stored in separate layers to facilitate the next data integration process. Data integration, the subsequent step in the pipeline, entails consolidating data from distinct layers into a comprehensive dataset. Step IV emphasizes the use of the integrated dataset for statistical analysis, which is linked to the Jupyter Notebooks of the trusted central platform.



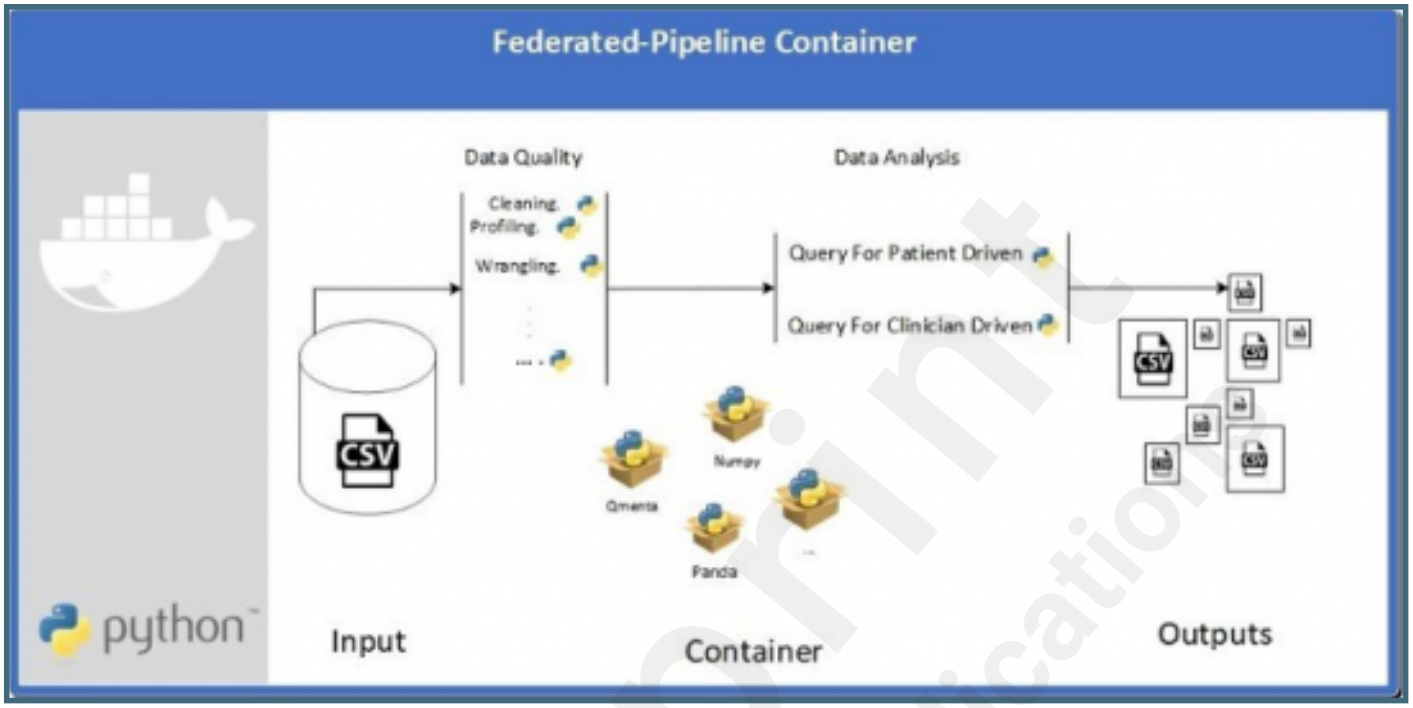
The Latest Federated Pipeline. The latest Federated Pipeline is a container composed of three primary components. The first component is the base image, which forms the bedrock of the infrastructure. This base image uses Alpine Linux as its underlying operating system. The remaining two components, the backend and frontend, are constructed on top of this base image. The backend consists of a suite of Python scripts, which are tasked with data quality assessment, enhancement, cleaning, and analysis. These scripts collaboratively process the incoming mapped data, preparing it for subsequent analysis. By contrast, the frontend was crafted using Microsoft's ASP.NET Core framework and the C# programming language. Within this pipeline, there is a customizable Automation Center module. This module can be adapted to meet the specific needs and requests of data partners. It also integrates Crontab, a tool that automates predefined tasks and outlines complex pipelines for execution at various intervals. The Automation Center module also links the container to the version control systems GitHub and DockerHub. This connection ensures the use of the most recent scripts and codes published by data analysts.



Summary of the data acquired by implementing the three-layer data acquisition. Federated registries (3a) contribute 32% of the data, while core dataset sharing accounts for 56%. Only 23% of the registries participated as federated registries (3b). Figure 3c provides a summary of the top 10 countries contributing data.

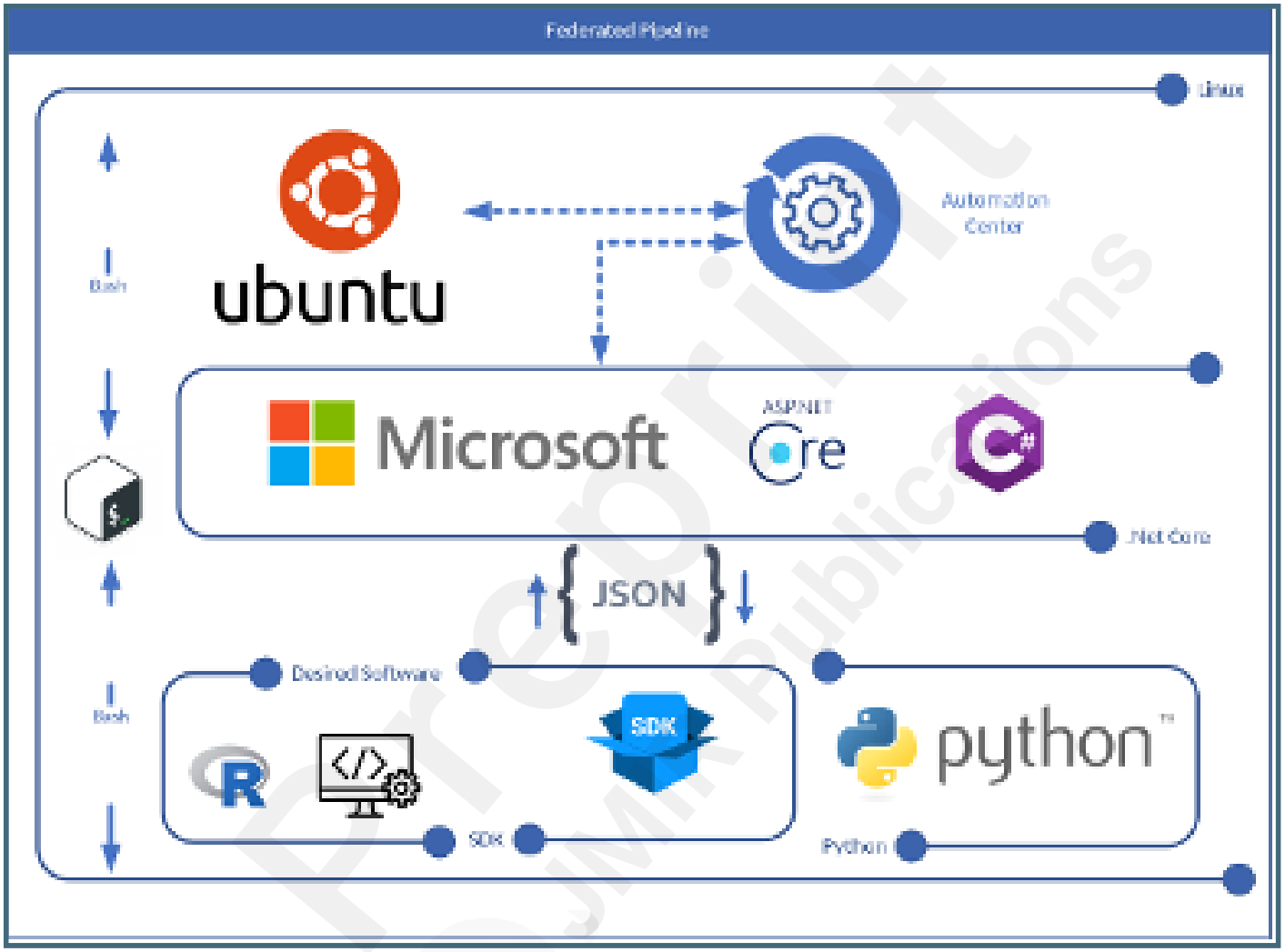


The federated pipeline COV 1.0 architecture. This infrastructure encompasses a series of scripts written in Python. These scripts are divided into two parts: (1) data quality scripts that are responsible for enhancing and cleaning the data and (2) a predefined query for data analysis.

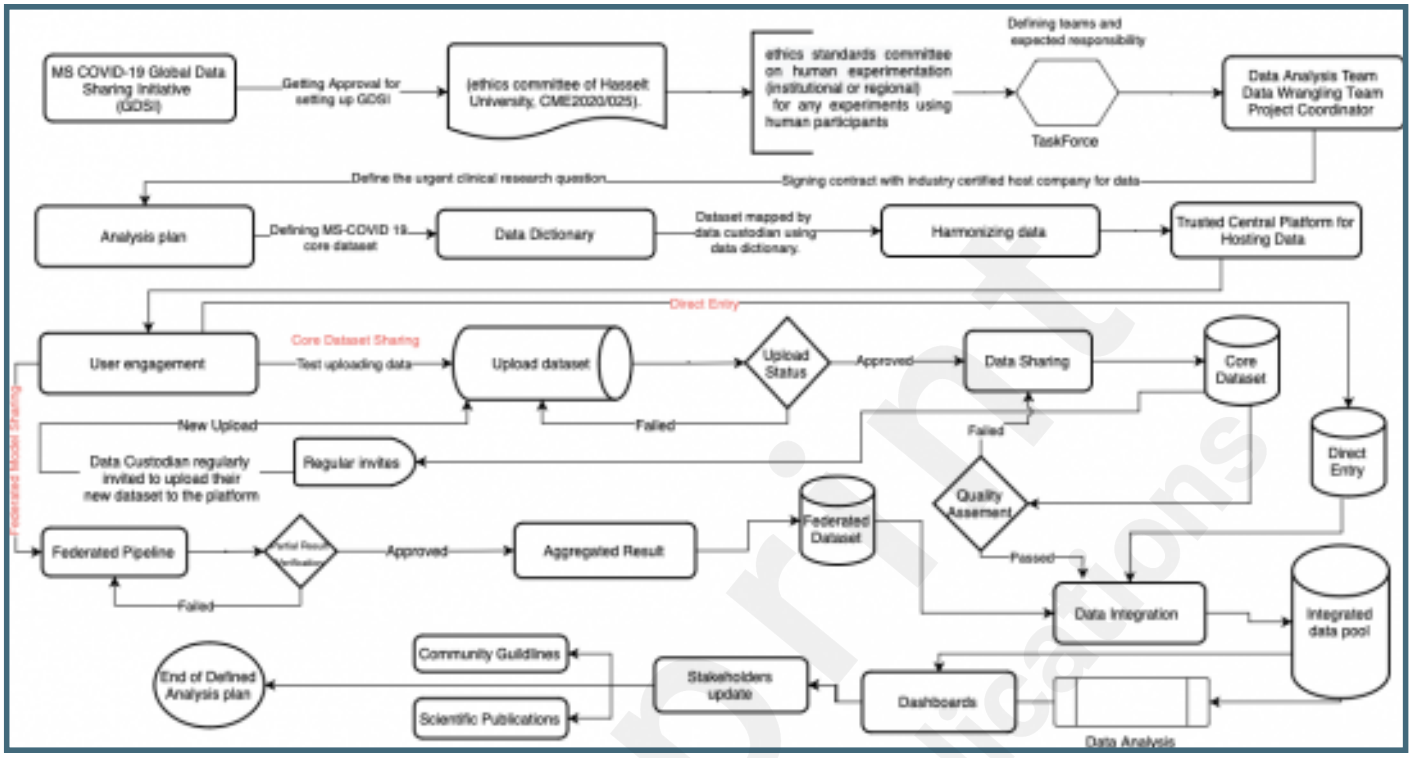


Preprint JMIR Publications

The Federated Pipeline COV 2.0 architecture. This container is comprised of three main components. The first component is the base image, which serves as the foundation of this infrastructure. This base image uses Ubuntu Linux as the underlying operating system. The second and third components, the backend and frontend, are built upon the base image. The backend is composed of a series of scripts written in Python, which are responsible for data quality assessment, enhancement, cleaning, and analysis. These scripts work in tandem to process the incoming mapped data and prepare it for further analysis. The frontend, by contrast, was developed using Microsoft's ASP.NET Core framework and the C# programming language.

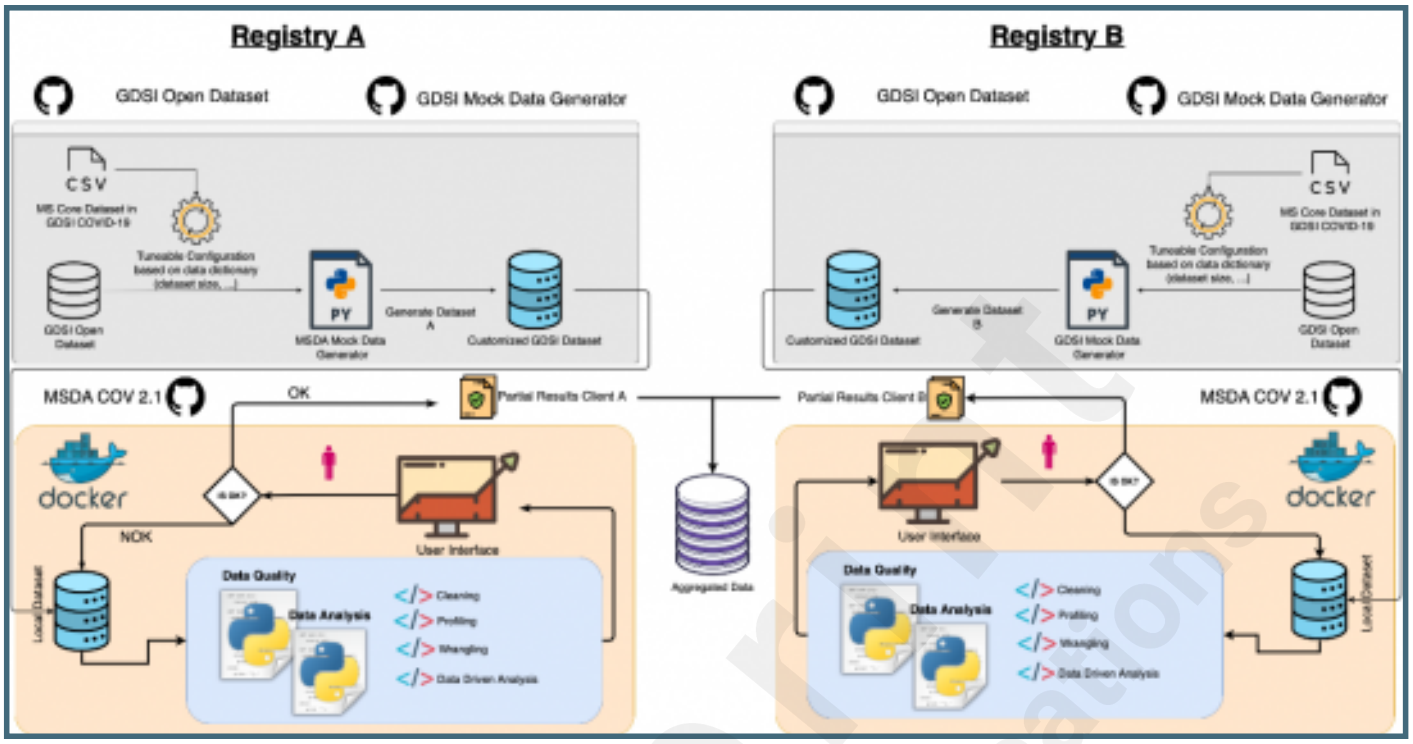


GDSI Flowchart: A Comprehensive Roadmap Detailing the Key Steps from Initiative Inception to Final Outcomes, Serving as a Blueprint for Adapting GDSI Principles to Other Biomedical Data Sharing Consortia.



Preprint JMIR Publications

A Guide to Replicating Federated Model Sharing Using GDSI Principles.



Preprint JMIR Publications

Multimedia Appendixes

Preprint JMIR Publications

Graphical Abstract.
URL: http://asset.jmir.pub/assets/380c94f50ee00c76cab645515aa1242c.png

Preprint JMIR Publications