Original Paper

The Journey of Data Within a Global Data Sharing Initiative: A Federated 3-Layer Data Analysis Pipeline to Scale Up Multiple Sclerosis Research

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

[1]ESAT, STADIUS, KU Leuven, Leuven, Belgium

[2]Biomedical Research Institute, Hasselt University, Diepenbeek, Belgium

[3]Data Science Institute, Hasselt University, Diepenbeek, Belgium

[4]University Multiple Sclerosis Center, Hasselt University, Diepenbeek, Belgium

[*]these authors contributed equally

Corresponding Author:
Liesbet M Peeters, Prof Dr
Biomedical Research Institute
Hasselt University
Agoralaan, Building C
Diepenbeek, 3590
Belgium
Phone: 32 11 26 92 05
Email: liesbet.peeters@uhasselt.be

Abstract

Background: Investigating low-prevalence diseases such as multiple sclerosis is challenging because of the rather small number of individuals affected by this disease and the scattering of real-world data across numerous data sources. These obstacles impair data integration, standardization, and analysis, which negatively impact 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 3 prevalent data-sharing streams: individual data sharing, core data set sharing, and federated model sharing.

Methods: A demand-driven methodology is employed for standardization, followed by 3 streams of data acquisition, a data quality enhancement process, a data integration procedure, and a concluding analysis stage to fulfill real-world data-sharing requirements. This pipeline's effectiveness was demonstrated through its successful implementation in the COVID-19 and multiple sclerosis global data sharing initiative.

Results: The global data sharing initiative yielded multiple scientific publications and provided extensive worldwide guidance for the community with multiple sclerosis. The pipeline facilitated gathering pertinent data from various sources, accommodating distinct sharing streams and assimilating them into a unified data set for subsequent statistical analysis or secure data examination. This pipeline contributed to the assembly of the largest data set of people with multiple sclerosis 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 health care, emphasizing its adaptability and capacity to address diverse research inquiries.

(JMIR Med Inform 2023;11:e48030) doi: 10.2196/48030

KEYWORDS

data analysis pipeline; federated model sharing; real-world data; evidence-based decision-making; end-to-end pipeline; multiple sclerosis; data analysis; pipeline; data science; federated; neurology; brain; spine; spinal nervous system; neuroscience; data sharing; rare; low prevalence

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 the 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 meaningful clinical insights even more challenging [5,6].

Despite these challenges, well-managed RWD have 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.

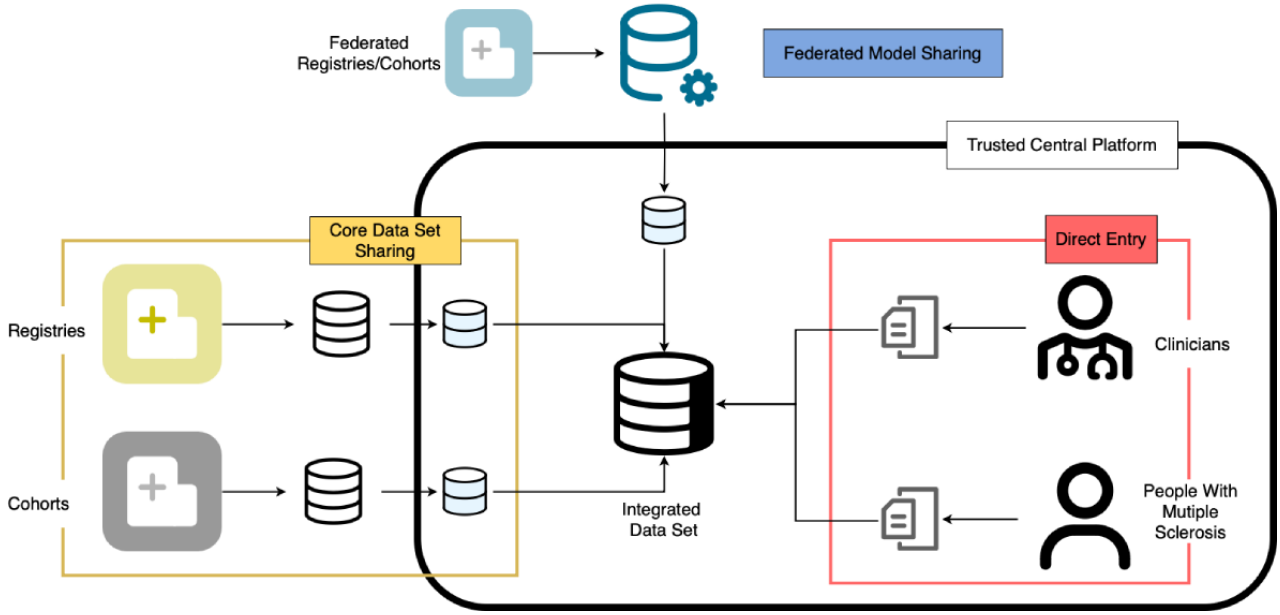
Current literature tends to oversimplify the intricate processes involved in managing RWD. These include standardization, acquisition, quality enhancement, integration, storage, governance, visualization, and eventual analysis and interpretation. Although these facets are crucial, they are often treated 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 by 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 2 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 health care 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 [7,16-19]. 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 3-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 data set was then readied for further analysis and exploration.

Figure 1. The global data sharing initiative data streams detailing the initiative's inclusive approach through a hybrid 3-layer data acquisition architecture: (1) direct entry, where individuals upload their data via a web-based form; (2) core data set sharing, where registries upload patient-level data to the central platform under signed data transfer agreements and ethics approvals; and (3) federated model sharing, allowing registries with restrictive policies to participate without directly submitting patient-level data to the central platform.

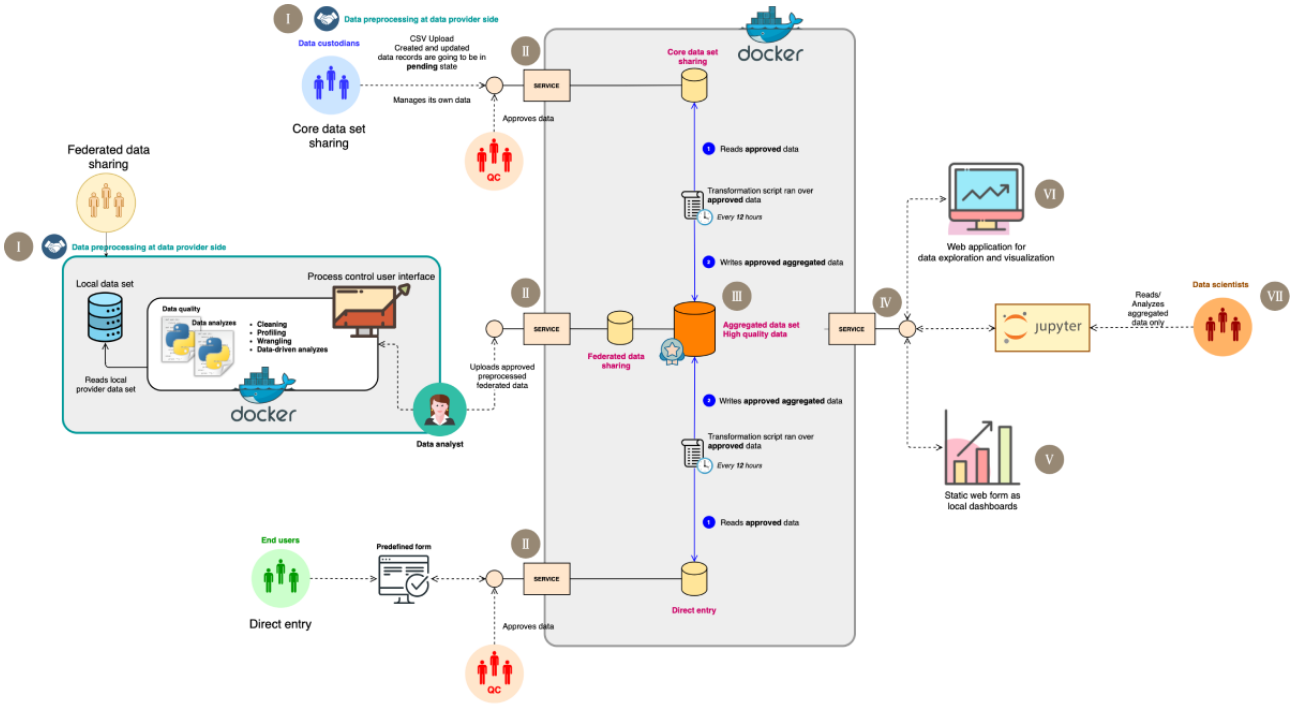


Methods

Overview of GDSI's Data Analysis Pipeline

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 5 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 are analyzed.

Figure 2. The global data sharing initiative's end-to-end real-world data analysis pipeline. Step I illustrates 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 multiple sclerosis core data set" (here referred to as the data dictionary). This process applies only to the core data set 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 data set sharing, and federated model sharing constitute the 3 data stream levels. The first 2 levels interact directly with a central platform where the core dataset is shared as static files, in this instance, Comma Separated Values (CSV), 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 data set. Step IV emphasizes the utilization of the integrated data set for further data exploration and analysis. Step V highlights the local dashboard, which serves as a quality check, enabling data providers to give feedback on their uploaded data as an additional sanity check. Step VI underscores the online dashboard that has been fed by the integrated data set, utilized by the taskforce during the development of the research questions to ascertain the feasibility of the study and to monitor the data being collected. In step VII, a Jupyter Notebook is provided to the data analysis team, securely connected to the integrated data set, facilitating statistical analysis.



Ethics Approval

This study received ethics approval from the ethics committee of Hasselt University (approval CME2020/025). For an in-depth discussion concerning ethical authorization, kindly refer to Simpson-Yap et al [17].

Data Dictionary

A data dictionary serves as a guide that details the attributes of components within an information database, ensuring consistent terminology [20-22]. 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 Data set" 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 data sets 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 [23] and presented in Table S1 of Multimedia Appendix 1.

Data Acquisition

Recognizing the value of diverse data sources for research outcomes [8,24], GDSI developed a hybrid data acquisition architecture. This framework consists of 3 distinct data sharing streams: direct entry, core data set 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 1 mode unattainable. GDSI's strength was rooted in its adaptability, effortlessly accommodating these 3 sharing streams and fusing them into a unified data set.

Data Sharing Streams

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 were 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 Data Set Sharing

Adhering to the conventional approach for clinical data sharing, data providers contributed a subset of their data set to the central platform. Although this mechanism excelled in handling extensive data, it grappled with challenges related to data collaboration agreements and complex regulatory stipulations. The heterogeneity in the data format further compounded these challenges. However, the architecture of core data set 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 that 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 data sets to the central platform.

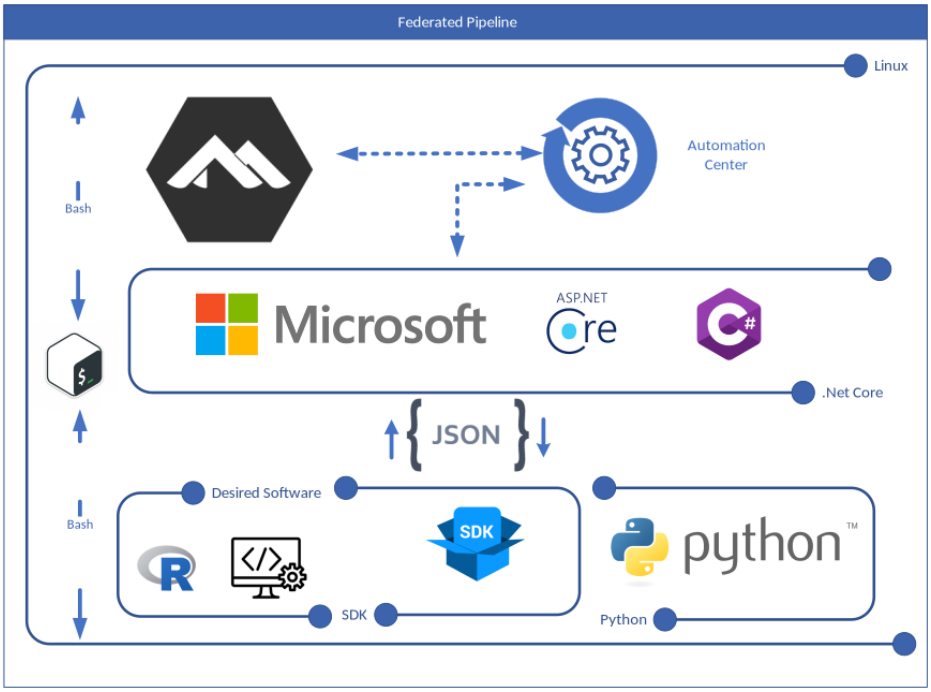
Federated Model Sharing

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 are 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 S2 of Multimedia Appendix 1 and within the associated GitHub repository [25].

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 [26] 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 [27]. Associated resources, including a demonstrative video walk-through, operational scripts, and the Docker image, can be found in [28-31]. Furthermore, the entire source code has been made publicly accessible on GitHub [32]. 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 Multimedia Appendix 2 [33].

Figure 3. The latest federated pipeline. This is a container composed of 3 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, which allows the container to be fine-tuned with other software development kits for further refinements and functionalities [27]. The remaining 2 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 GitHub and Docker Hub version control systems. This connection ensures the use of the most recent scripts and codes published by data analysts. SDK: software development kit.



Data Quality Assessment and Enhancement

The integrity of the acquired data set 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 web-based form. For the core data set sharing approach, uploaded data were 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 get excluded; it is 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 (ie, if any date > date reporting, then the variable is flagged as FAIL) or before a person's birth date (ie, if any date 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 criteria (some highlighted examples).[a]

Variable	Format	Interdependency	Pass and fail criteria
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, then 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[b]	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[c]	Number (0.0, 10.0)	None	if edss_value < 0 OR edss_value > 10, then fail
type_dmt[d]	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)

[a]67 more checks have been performed, but these checks are not presented in this table.

[b]MS: multiple sclerosis.

[c]EDSS: Expanded Disability Status Scale.

[d]DMT: disease-modifying therapy.

Data Integration

The quality-checked data acquired within each stream are 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 data sets. To derive comprehensive insights, there was a paramount need to coalesce these distinct data sets 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 nonoverlapping ranges $\sum_k = \sigma_k^1, \sigma_k^2, \ldots, \sigma_k^{jk}$ that partitions 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 were then converted into 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 $I[.]$ 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 different variable names and patient counts. This table was subsequently stored on the central platform. The same computation was performed on the direct entry and core data

set, as the raw data were 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* as the number of variables used in the analysis and *N* as the number of patients.

Data Analysis

Following the careful integration of a diverse data set, 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 [7,16,17,19]. The analytical approach implemented by Simpson-Yap 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 performed to assess the association between disease-modifying therapies (DMTs) and several outcomes, including hospitalization, intensive care unit 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

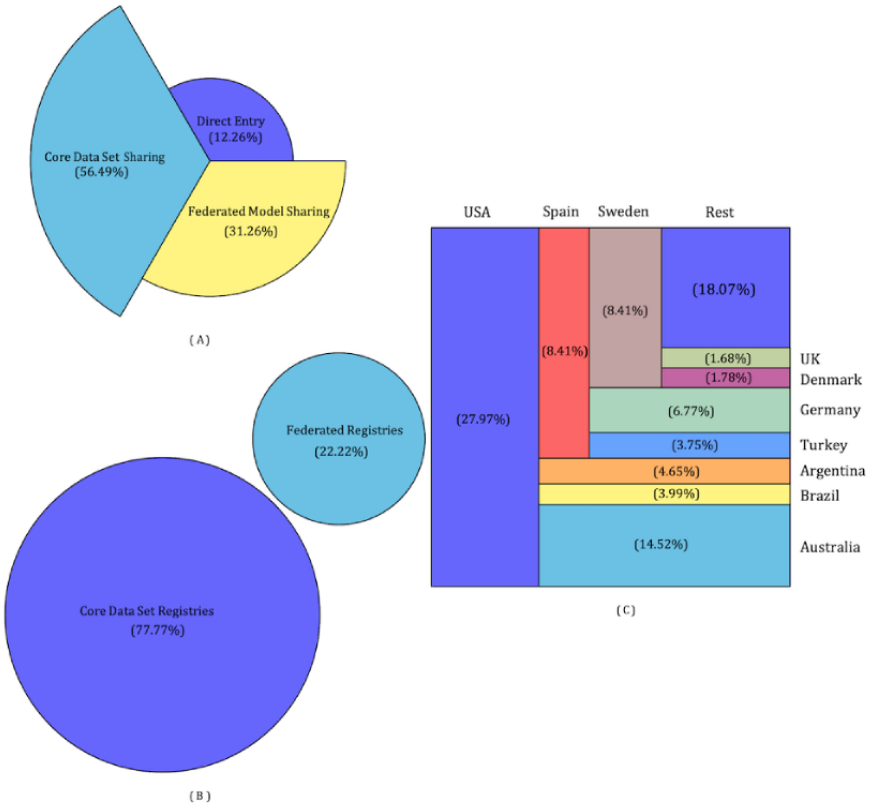
Data Acquisition

Using the pragmatic 3-layer approach of GDSI, we obtained the largest cohort of people with MS infected with COVID-19. The data were collected from 80 countries, with the top 10 contributing countries being the United States (3157/11,284, 27.97%), Australia (1639/11,284, 14.52%), Spain (949/11,284, 8.41%), Sweden (949/11,284, 8.41%), Germany (765/11,284, 6.77%), Argentina (525/11,284, 4.65%), Brazil (451/11,284, 3.99%), Turkey (424/11,284, 3.75%), Denmark (201/11,284, 1.78%), and the United Kingdom (190/11,284, 1.68%), which accounted for over 80% of the total number of records. Via direct entry, data were collected from 67 countries, with Spain contributing the largest number of records (758/1383, 54.80%), followed by the Netherlands (95/1383, 6.86%), United Kingdom (80/1383, 5.78%), United States (53/1383, 3.83%), Australia (40/1383, 2.89%), and 62 other countries (357/1383, 25.81%), resulting in a total of 1383 records. Data were collected from 18 different registries worldwide. Fourteen of these participated in core data set sharing, contributing to 6374 records. Meanwhile, 4 used the federated model sharing approach, contributing to an additional 3527 records. 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 have been released and are accessible through the associated PhysioNet repository [35].

Table 2. Data acquisition summary in the global data sharing initiative (N=11,284).

Method of data sharing	Values, n (%)
Direct entry	1383 (12.26)
Core data set sharing	6374 (56.49)
Federated model sharing	3527 (31.26)

Figure 4. Summary of the data acquired by implementing the 3-layer data acquisition. (A) Federated registries contribute to 31.26% (3527/11,284) of the data, while core data set sharing accounts for 56.49% (6374/11,284). (B) Only 22% (4/18) of the registries participated as federated registries. (C) 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 in the following ranges: Σ_{age} = (18-50 years, 50-70 years, >70 years). Sex was binarized into male and female. MS phenotype was binarized into relapsing-remitting MS and progressive MS. The disability score was dichotomized into $\Sigma_{\text{Expanded Disability Status Scale}}$ = (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, intensive care unit admission, ventilation, death). Compared to patients using all other DMTs, those using rituximab had a higher risk of hospitalization (adjusted odds ratio [aOR] 2.76, 95% CI 1.87-4.07), intensive care unit 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 intensive care unit 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 et al [7].

Discussion

Insights From the GDSI Study on MS and COVID-19

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 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]. Although 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 Fast Healthcare Interoperability Resources [38] and Observational Medical Outcomes Partnership [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 health care, 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, the adoption of a generalized framework such as [46] can serve as a blueprint for enhancing data quality across various health care 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 health care, 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 et al [48] could potentially fill this gap, enhancing both organizational efficiency and the quality of analytical models.

Embracing Federated Model Sharing and Privacy Concerns

Although 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, thereby 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 re-evaluate GDSI's current assumptions of trustworthiness, inquisitiveness, and nonantagonistic 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 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 health care is supported by various studies [63-66]. Tools such as 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 data sets 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, health care 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 comes challenges. Although GDSI's analysis pipeline presents a viable technical solution for collaborative health care 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 amid 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, but 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 provides [23,25,29,30,32,34,35,72-74] can be directly leveraged and adjusted after thorough assessment and evaluation. These resources bifurcate into 2 main categories: disease-agnostic and disease-specific components.

Within the context of disease-agnostic components, the 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 data set. 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 can refer to Multimedia Appendix 3, which 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. A graphical abstract delineating the high-level architecture of this study is presented in Multimedia Appendix 4, which provides additional insights regarding the architectural framework.

Conclusion

GDSI had 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 community with MS [7,9,16-19]. This underscores the vast potential of data-driven collaborative efforts to yield improved health care 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 data set. 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 health care challenges but also laid the groundwork for future initiatives. Its hybrid approach to data acquisition and analysis provided a scalable framework applicable to other health care sectors. In doing so, GDSI stands as a compelling example of how data sharing and collaborative learning can meaningfully advance health care 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, or publication of this paper from the following sources: the operational costs associated with this study were funded by the Multiple Sclerosis International

Federation and the Multiple Sclerosis Data Alliance (MSDA) operating under the European Charcot Foundation. The MSDA is a global not-for-profit multistakeholder 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 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 the Clinical Outcomes Research Unit, The University of Melbourne, with support from National Health and Medical Research Council (1129189 and 1140766). The authors wish to extend their sincere appreciation to Nikola Lazovski for his invaluable guidance and collaboration throughout the global data sharing initiative 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.

Authors' 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 manuscript drafting and revision for content. EDB contributed to manuscript drafting and revision, played a major role in acquiring data, and contributed to the study concept or design as well as the analysis or interpretation of data. YM and LMP provided overall supervision and coordination of the study and contributed to manuscript drafting and revision for content, including writing for content.

Conflicts of Interest

AP and YM are funded by VLAIO (Flanders Innovation and Entrepreneurship) PM: Augmenting Therapeutic Effectiveness through Novel Analytics (HBC.2019.2528) and Research Council Katholieke Universiteit Leuven: Symbiosis 4 (C14/22/125) and Symbiosis 3 (C14/18/092; CELSA - Active Learning; CELSA/21/019). AP and YM are affiliated to Leuven.AI and received funding from the Flemish government (Artificial Intelligence Research Program, Research Foundation Flanders [FWO] Strategic Basic [SB] Research; S003422N). EDB was funded by an FWO-SB grant. LMP is the chair of MSDA, which received income from a range of corporate sponsors, including Biogen, Bristol Myers Squibb, Janssen Pharmaceuticals, Merck, Novartis, and Roche. LG is funded by the Flemish government under the Onderzoeksprogramma Artificiele Intelligentie Vlaanderen. TP is funded by the Flemish government under the Bijzonder Onderzoeksfonds special research fund BOF22OWB01.

Multimedia Appendix 1

Analysis of the data dictionary employed within the global data sharing initiative.
[DOCX File , 20 KB-Multimedia Appendix 1]

Multimedia Appendix 2

The 3 iterative pipelines developed: federated pipeline COV1.0, COV2.0, and COV2.1.
[DOCX File , 139 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Roadmap of the global data sharing initiative.
[DOCX File , 427 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Graphical abstract.
[PNG File , 3416 KB-Multimedia Appendix 4]

References

1. Mitani AA, Haneuse S. Small data challenges of studying rare diseases. JAMA Netw Open 2020 Mar 02;3(3):e201965 [FREE Full text] [doi: 10.1001/jamanetworkopen.2020.1965] [Medline: 32202640]
2. Walton C, King R, Rechtman L, Kaye W, Leray E, Marrie RA, et al. Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS, third edition. Mult Scler 2020 Dec;26(14):1816-1821 [FREE Full text] [doi: 10.1177/1352458520970841] [Medline: 33174475]
3. Real-world evidence. US Food and Drug Administration. URL: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence [accessed 2023-03-30]
4. Burcu M, Dreyer NA, Franklin JM, Blum MD, Critchlow CW, Perfetto EM, et al. Real-world evidence to support regulatory decision-making for medicines: considerations for external control arms. Pharmacoepidemiol Drug Saf 2020 Oct;29(10):1228-1235 [FREE Full text] [doi: 10.1002/pds.4975] [Medline: 32162381]

5. Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. J Clin Invest 2020 Feb 03;130(2):565-574 [FREE Full text] [doi: 10.1172/JCI129197] [Medline: 32011317]
6. Hiramatsu K, Barrett A, Miyata Y, PhRMA Japan Medical Affairs Committee Working Group 1. Current status, challenges, and future perspectives of real-world data and real-world evidence in Japan. Drugs Real World Outcomes 2021 Dec;8(4):459-480 [FREE Full text] [doi: 10.1007/s40801-021-00266-3] [Medline: 34148219]
7. Simpson-Yap S, De Brouwer E, Kalincik T, Rijke N, Hillert JA, Walton C, et al. Associations of disease-modifying therapies with COVID-19 severity in multiple sclerosis. Neurology 2021 Nov 09;97(19):e1870-e1885 [FREE Full text] [doi: 10.1212/WNL.0000000000012753] [Medline: 34610987]
8. Katkade VB, Sanders KN, Zou KH. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. JMDH 2018 Jul;Volume 11:295-304 [doi: 10.2147/jmdh.s160029]
9. Peeters LM, Parciak T, Walton C, Geys L, Moreau Y, De Brouwer E, et al. COVID-19 in people with multiple sclerosis: a global data sharing initiative. Mult Scler 2020 Sep;26(10):1157-1162 [FREE Full text] [doi: 10.1177/1352458520941485] [Medline: 32662757]
10. Antoniou V, Vassilakis E, Hatzaki M. Is crowdsourcing a reliable method for mass data acquisition? The case of COVID-19 spread in Greece during spring 2020. ISPRS Int J Geo Inf 2020 Oct 14;9(10):605 [doi: 10.3390/ijgi9100605]
11. Yu C, Chang S, Chang T, Wu JL, Lin Y, Chien H, 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 May 20;23(5):e27806 [FREE Full text] [doi: 10.2196/27806] [Medline: 33900932]
12. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernández-Bertolín S, Kim C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. Comput Methods Programs Biomed 2021 Nov;211:106394 [FREE Full text] [doi: 10.1016/j.cmpb.2021.106394] [Medline: 34560604]
13. Nishimwe A, Ruranga C, Musanabaganwa C, Mugeni R, Semakula M, Nzabanita J, 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 2022 Aug 12;22(1):214 [FREE Full text] [doi: 10.1186/s12911-022-01965-9] [Medline: 35962355]
14. Junior EPP, Normando P, Flores-Ortiz R, Afzal MU, Jamil MA, Bertolin SF, 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 2023 Mar 16;30(4):643-655 [FREE Full text] [doi: 10.1093/jamia/ocac180] [Medline: 36264262]
15. Peeters LM, Parciak T, Kalra D, Moreau Y, Kasilingam E, van Galen P, et al. Multiple Sclerosis Data Alliance - A global multi-stakeholder collaboration to scale-up real world data research. Mult Scler Relat Disord 2021 Jan;47:102634 [FREE Full text] [doi: 10.1016/j.msard.2020.102634] [Medline: 33278741]
16. Simpson-Yap S, Brouwer ED, Kalincik T, et al. Associations of DMT therapies with COVID-19 severity in multiple sclerosis? Int J Epidemiol 2021 Sep 02:51 [doi: 10.1093/ije/dyab168.604]
17. Simpson-Yap S, Pirmani A, Kalincik T, De Brouwer E, Geys L, Parciak T, et al. Updated results of the COVID-19 in MS global data sharing initiative. Neurol Neuroimmunol Neuroinflamm 2022 Aug 29;9(6):e200021 [doi: 10.1212/nxi.0000000000200021]
18. An update from the MS global data sharing initiative - thank you!. MS International Federation. URL: https://www.msif.org/news/2022/04/06/an-update-from-the-ms-global-data-sharing-initiative-thank-you/ [accessed 2023-03-30]
19. Simpson-Yap S, Pirmani A, De Brouwer E, Peeters LM, Geys L, Parciak T, et al. Severity of COVID19 infection among patients with multiple sclerosis treated with interferon-β. Mult Scler Relat Disord 2022 Oct;66:104072 [FREE Full text] [doi: 10.1016/j.msard.2022.104072] [Medline: 35917745]
20. McCabe A, Nic An Fhailí S, O'Sullivan R, Brenner M, Gannon B, Ryan J, et al. Development and validation of a data dictionary for a feasibility analysis of emergency department key performance indicators. Int J Med Inform 2019 Jun;126:59-64 [doi: 10.1016/j.ijmedinf.2019.01.015] [Medline: 31029264]
21. Lin S, Morrison LJ, Brooks SC. Development of a data dictionary for the Strategies for Post Arrest Resuscitation Care (SPARC) network for post cardiac arrest research. Resuscitation 2011 Apr;82(4):419-422 [doi: 10.1016/j.resuscitation.2010.12.006] [Medline: 21276647]
22. Moss E. The national health data dictionary. Health Inf Manag 1994 Mar;24(1):26-29 [doi: 10.1177/183335839402400112] [Medline: 10141009]
23. COVID19-GDSI/data dictionary. GitHub. URL: https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI/blob/main/Data%20Dictionary.pdf [accessed 2023-04-09]
24. Cook JA, Collins GS. The rise of big clinical databases. Br J Surg 2015 Jan;102(2):e93-e101 [doi: 10.1002/bjs.9723] [Medline: 25627139]
25. COVID19-GDSI/Buckets for federated registries. GitHub. URL: https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI/blob/main/Buckets%20for%20federated%20registries.pdf [accessed 2023-04-11]
26. Anderson C. Docker [software engineering]. IEEE Softw 2015 May;32(3):102-1c3 [doi: 10.1109/ms.2015.62]
27. Alpine Linux. URL: https://www.alpinelinux.org/ [accessed 2023-07-22]

28. Demo presentation of the federated pipeline. MSDA Federated Pipeline COV 2.1 YouTube page. URL: https://www.youtube.com/watch?v=d-QuCNDbHKc&ab_channel=AshkanPirmani [accessed 2023-03-31]
29. MS-DATA-ALLIANCE/COVID19-GDSI. GitHub. URL: https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI/tree/main [accessed 2023-04-09]
30. MS-DATA-ALLIANCE/COVID19-GDSI2021: MSDA toolkit for federated registries participating in the COVID-19 in MS Global Data Sharing initiative (GDSI). GitHub. URL: https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI2021 [accessed 2023-03-31]
31. msdaalliance/covid19gdsi-ui. Docker Hub. URL: https://hub.docker.com/r/msdaalliance/covid19gdsi-ui [accessed 2023-03-31]
32. MS-DATA-ALLIANCE/COVID19-GDSI2021-UI: MSDA toolkit for federated registries participating in the COVID-19 in MS Global Data Sharing initiative (GDSI). GitHub. URL: https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI2021-UI [accessed 2023-04-11]
33. Snyk. URL: https://snyk.io/ [accessed 2023-03-31]
34. MS-DATA-ALLIANCE/COVID19-GDSI. GitHub. URL: https://github.com/MS-DATA-ALLIANCE/COVID19-GDSI/blob/main/Data%20quality%20assessment%20and%20enhancement%20pipeline.pdf [accessed 2023-04-09]
35. Patient-level dataset to study the effect of COVID-19 in people with multiple sclerosis v1. PhysionNet. URL: https://physionet.org/content/patient-level-data-covid-ms/1.0.0/ [accessed 2023-08-13]
36. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. N Engl J Med 2000 Jun 22;342(25):1878-1886 [doi: 10.1056/nejm200006223422506]
37. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc 2015 May;22(3):553-564 [FREE Full text] [doi: 10.1093/jamia/ocu023] [Medline: 25670757]
38. Pfaff ER, Champion J, Bradford RL, Clark M, Xu H, Fecho K, 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 Oct 16;7(4):e15199 [FREE Full text] [doi: 10.2196/15199] [Medline: 31621639]
39. Ahmadi N, Peng Y, Wolfien M, Zoch M, Sedlmayr M. OMOP CDM can facilitate data-driven studies for cancer prediction: a systematic review. Int J Mol Sci 2022 Oct 05;23(19):11834 [FREE Full text] [doi: 10.3390/ijms231911834] [Medline: 36233137]
40. Yu Y, Zong N, Wen A, Liu S, Stone DJ, Knaack D, 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 2022 Mar;127:104002 [FREE Full text] [doi: 10.1016/j.jbi.2022.104002] [Medline: 35077901]
41. Makadia R, Ryan PB. Transforming the premier perspective hospital database into the observational medical outcomes partnership (OMOP) common data model. EGEMS (Wash DC) 2014;2(1):1110 [FREE Full text] [doi: 10.13063/2327-9214.1110] [Medline: 25848597]
42. Chao-Gan Y, Yu-Feng Z. DPARSF: A MATLAB toolbox for "pipeline" data analysis of resting-state fMRI. Front Syst Neurosci 2010;4:13 [FREE Full text] [doi: 10.3389/fnsys.2010.00013] [Medline: 20577591]
43. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 2013;43(1110):11.10.1-11.10.33 [FREE Full text] [doi: 10.1002/0471250953.bi1110s43] [Medline: 25431634]
44. Uritskiy GV, DiRuggiero J, Taylor J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 2018 Sep 15;6(1):158 [FREE Full text] [doi: 10.1186/s40168-018-0541-1] [Medline: 30219103]
45. Chen H, Lau MC, Wong MT, Newell EW, Poidinger M, Chen J. Cytofkit: A bioconductor package for an integrated mass cytometry data analysis pipeline. PLoS Comput Biol 2016 Sep;12(9):e1005112 [FREE Full text] [doi: 10.1371/journal.pcbi.1005112] [Medline: 27662185]
46. Lee K, Weiskopf N, Pathak J. A framework for data quality assessment in clinical research datasets. AMIA Annu Symp Proc 2017;2017:1080-1089 [FREE Full text] [Medline: 29854176]
47. Perez JA, Bellot GO, Zirpins C. Data governance for federated machine learning in secure web-based systems. In: Minutes of the Predoctoral Research Conference in Computer Engineering: Proceedings of the Doctoral Consortium in Computer Science. Spain: Universidad de la Rioja; 2021:36-39
48. Peregrina JA, Ortiz G, Zirpins C. Towards data governance for federated machine learning. In: Advances in Service-Oriented and Cloud Computing. Switzerland: Springer Cham; 2022.
49. Nasirigerdeh R, Torkzadehmahani R, Baumbach J, et al. On the privacy of federated pipelines. 2021 Presented at: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval; July 11-15; Canada [doi: 10.1145/3404835.3462996]
50. McMahan B, Moore E, Ramage D, Hampson S, et al. Communication-efficient learning of deep networks from decentralized data. arXiv Preprint posted online on February 17, 2016. [doi: 10.48550/arXiv.1602.05629]
51. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med 2020;3:119 [FREE Full text] [doi: 10.1038/s41746-020-00323-1] [Medline: 33015372]
52. Bagdasaryan E, Veit A, Hua Y, et al. How to backdoor federated learning. arXiv Preprint posted online on July 2, 2018. [doi: 10.48550/arXiv.1807.00459]

53. Tolpegin V, Truex S, Gursoy ME, Liu L. Data poisoning attacks against federated learning systems. arXiv Preprint posted online on August 11, 2020. [doi: 10.1007/978-3-030-58951-6_24]
54. Cynthia D. Differential privacy. In: Lecture Notes in Computer Science. Switzerland: Springer Nature; 2006:1-12
55. Wood A, Najarian K, Kahrobaei D. Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Comput Surv 2020 Aug 25;53(4):1-35 [doi: 10.1145/3394658]
56. Raisaro JL, Choi G, Pradervand S, Colsenet R, Jacquemont N, Rosat N, et al. Protecting privacy and security of genomic data in i2b2 with homomorphic encryption and differential privacy. IEEE/ACM Trans. Comput. Biol. and Bioinf 2018:1-1 [doi: 10.1109/tcbb.2018.2854782]
57. Deist TM, Dankers FJWM, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, et al. Distributed learning on 20,000+ lung cancer patients - The personal health train. Radiother Oncol 2020 Mar;144:189-200 [FREE Full text] [doi: 10.1016/j.radonc.2019.11.019] [Medline: 31911366]
58. Huang L, Shea AL, Qian H, Masurkar A, Deng H, Liu D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. J Biomed Inform 2019 Nov;99:103291 [FREE Full text] [doi: 10.1016/j.jbi.2019.103291] [Medline: 31560949]
59. Jochems A, Deist TM, El Naqa I, Kessler M, Mayo C, Reeves J, et al. Developing and validating a survival prediction model for NSCLC patients through distributed learning across 3 countries. Int J Radiat Oncol Biol Phys 2017 Oct 01;99(2):344-352 [FREE Full text] [doi: 10.1016/j.ijrobp.2017.04.021] [Medline: 28871984]
60. Li J, Tian Y, Zhu Y, Zhou T, Li J, Ding K, et al. A multicenter random forest model for effective prognosis prediction in collaborative clinical research network. Artif Intell Med 2020 Mar;103:101814 [doi: 10.1016/j.artmed.2020.101814] [Medline: 32143809]
61. Li X, Gu Y, Dvornek N, Staib LH, Ventola P, Duncan JS. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Med Image Anal 2020 Oct;65:101765 [FREE Full text] [doi: 10.1016/j.media.2020.101765] [Medline: 32679533]
62. Pirmani A, De Brouwer E, Moreau Y, Peeters LM. Federated learning for everyone (FL4E). Elixir All Hands Meeting 2023:1 [doi: 10.7490/F1000RESEARCH.1119405.1]
63. Gotz D, Borland D. Data-driven healthcare: challenges and opportunities for interactive visualization. IEEE Comput Grap Appl 2016 May;36(3):90-96 [doi: 10.1109/mcg.2016.59]
64. Stadler JG, Donlon K, Siewert JD, Franken T, Lewis NE. Improving the efficiency and ease of healthcare analysis through use of data visualization dashboards. Big Data 2016 Jun;4(2):129-135 [doi: 10.1089/big.2015.0059] [Medline: 27441717]
65. Menon A, Aishwarya MS, Joykutty AM, Av AY. Data visualization and predictive analysis for smart healthcare: tool for a hospital. 2021 Presented at: 2021 IEEE Region 10 Symposium (TENSYMP); October 4, 2021; Jeju, Republic of Korea p. 1-8 [doi: 10.1109/TENSYMP52854.2021.9550822]
66. Battineni G, Mittal M, Jain S. Data visualization in the transformation of healthcare industries. In: Advanced Prognostic Predictive Modelling in Healthcare Data Analytics. Singapore: Springer; 2021:1-23
67. Reporting and embedded business intelligence software. Jaspersoft. URL: https://www.jaspersoft.com/ [accessed 2023-07-31]
68. Business intelligence and analytics software. Tableau. URL: https://www.tableau.com/ [accessed 2023-07-31]
69. Looker business intelligence platform embedded analytics. Google Cloud. URL: https://cloud.google.com/looker [accessed 2023-08-11]
70. Discover the Domo data experience platform. Domo. URL: https://www.domo.com/ [accessed 2023-08-11]
71. Data visualization. Microsoft Power BI. URL: https://powerbi.microsoft.com/en-us/ [accessed 2023-08-11]
72. jupyter/datascience-notebook. Docker Hub. URL: https://hub.docker.com/r/jupyter/datascience-notebook [accessed 2023-03-31]
73. .NET. Microsoft. URL: https://dotnet.microsoft.com/en-us/ [accessed 2023-07-31]
74. The cron schedule expression editor. Crontab. URL: https://crontab.guru/ [accessed 2023-07-31]

Abbreviations

aOR: adjusted odds ratio
DMT: disease-modifying therapy
GDSI: global data sharing initiative
MS: multiple sclerosis
RWD: real-world data

Edited by C Lovis; submitted 20.04.23; peer-reviewed by X Liu, D Blumenthal; comments to author 05.07.23; revised version received 25.08.23; accepted 30.09.23; published 09.11.23

Please cite as:
Pirmani A, De Brouwer E, Geys L, Parciak T, Moreau Y, Peeters LM
The Journey of Data Within a Global Data Sharing Initiative: A Federated 3-Layer Data Analysis Pipeline to Scale Up Multiple Sclerosis Research
JMIR Med Inform 2023;11:e48030
URL: https://medinform.jmir.org/2023/1/e48030
doi: 10.2196/48030
PMID: 37943585