Advancing Data Sovereignty in Africa: Deploying DataSHIELD to Federate Cross-Border Data Silos within the Data Science Without Borders Initiative

Agnes N. Kiragga

akiragga@aphrc.org

Data Science Program, Research Division, African Population and Health Research Center

Michael Ochola

Data Science Program, Research Division, African Population and Health Research Center

David Sarrat-González

Barcelona Institute for Global Health (ISGlobal)

Pauline Andeso

Data Science Program, Research Division, African Population and Health Research Center

Steve Cygu

Data Science Program, Research Division, African Population and Health Research Center

Reinpeter Momanyi

Data Science Program, Research Division, African Population and Health Research Center

Miranda Barasa

Data Science Program, Research Division, African Population and Health Research Center

David Amadi

London School of Hygiene and Tropical Medicine

Ousmane Diop

Institute for Health Research, Epidemiological Surveillance and Training (IRESSEF)

Anicet Onana

Douala General Hospital

Aminata Mboup

Institute for Health Research, Epidemiological Surveillance and Training (IRESSEF)

Samuel Iddi

Data Science Program, Research Division, African Population and Health Research Center

John M. Bwanika

Africa Centre for Applied Digital Health (CADH)

Bertrand Hugo Mbatchou

Douala General Hospital

Moussa Sarr

Institute for Health Research, Epidemiological Surveillance and Training (IRESSEF)

Rebecca Wilson

Institute of Population Health, Faculty of Health and Life Sciences, University of Liverpool

Juan R Gonzalez

Barcelona Institute for Global Health (ISGlobal)

Article

Keywords: DataSHIELD, Federated Analysis, Africa, Data Sovereignty, OMOP Common Data Model

Posted Date: March 31st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9149238/v1

Additional Declarations: No competing interests reported.

Abstract

Collaborative health research across Africa is constrained by data sovereignty concerns, heterogeneous regulatory frameworks, limited infrastructure, and persistent capacity gaps, which hinder equitable cross-border data sharing. These challenges limit the availability of large, harmonized datasets needed to address the continent's burden of infectious and non-communicable diseases while maintaining control over sensitive health data.

Within this context, the Data Science Without Borders in Africa (DSWB) project implemented a privacy-preserving federated analysis framework using DataSHIELD across four institutions in Senegal, Cameroon, Ethiopia, and Kenya. This study documents the design, deployment, and early outcomes of this implementation. The approach combined consortium-wide data governance harmonization, a structured capacity-building programme, and a standardized technical architecture integrating DataSHIELD with the OMOP Common Data Model via dsOMOP.

The deployment resulted in a functional federated network that enabled secure, in-situ analyses of harmonized clinical data. Multidisciplinary teams successfully executed federated descriptive and modelling analyses, managed Opal-based servers and implemented OMOP-based data harmonization workflows. Challenges related to connectivity, software heterogeneity, and institutional security were mitigated through server-side computation, standardized environments, and collaboration with institutional IT teams. Overall, this work demonstrates the feasibility of scalable, ethically robust federated health data analysis in diverse African research settings while preserving data sovereignty.

Introduction

Africa faces a disproportionate burden of infectious and non-communicable diseases, with over 100 disease outbreaks reported annually on the continent [1]. Effective responses to tuberculosis, malaria, HIV, Ebola, COVID-19, Marburg virus, and antimicrobial resistance require high-quality, comprehensive datasets that can inform evidence-based policymaking and enhance healthcare decision-making. A recent study reported more than 100 disease outbreaks annually in Africa, yet the scarcity of carefully curated, large-scale epidemiological data sources and limited analytical capacity severely constrain timely, effective responses to health emergencies [2]. The recent COVID-19 pandemic underscored the urgent need for transparent and timely continental data sharing across African borders to coordinate pandemic preparedness and response [3]. Cross-border data sharing in Africa is hampered by several inherent challenges, including a fragmented and evolving regulatory landscape shaped by the heterogeneity of data protection legislation. As of 2024, approximately 36 of 54 African countries have enacted some form of data protection legislation – a dramatic increase from just 12 countries in 2012 [4,5]. However, this legislative proliferation has created a patchwork of incompatible regulatory frameworks rather than a harmonized approach. There are profound differences across African countries in definitions for personal data, sensitive data, anonymization, and pseudonymization techniques to ensure confidentiality; differences in national legal consenting and data processing [6],

hence promoting differences in enforcement mechanisms for non-compliance to data protection laws, particularly to health and research data and materials. While other continents such as Europe, have developed continental data protection laws, including the General Data Protection Regulation (GDPR) and the EU Data Governance Act [7], the African Union developed the AU Convention on Cyber Security and Personal Data Protection [8] in June 2023, offering a framework for harmonized data sharing and protection across the continent. However, as of 2024, 15 AU member states had ratified the convention, while others, including South Africa, struggle to align it with existing national data protection guidelines. Furthermore, data management practices across the continent remain inconsistent and are often hampered by a lack of standardized platforms and interoperable systems. To date, the use of international common data models such as the OMOP remains very low in Africa [9]. Variability in software and hardware capabilities, and in adherence to data governance frameworks, further complicates collaborative

research efforts. The continent still grapples with a shortage of skilled professionals equipped to handle emerging data technologies and sophisticated analytic pipelines, exacerbating existing gaps in research capacity. For example, 1% of global AI talent is in Africa [10] - making it hard to promote advancement in data use. These technical and infrastructural challenges must be addressed in tandem with ethical and regulatory reforms to realize the full potential of data sharing for advancing scientific discovery in Africa.

Ethical considerations play a central role in the adoption and implementation of data-sharing practices across African research settings. The imperative to protect participant confidentiality and ensure responsible stewardship of sensitive information remains paramount, particularly in environments where regulatory frameworks for data protection are still evolving. Research Ethics Committees frequently grapple with balancing the potential benefits of open data with the need to safeguard individual rights, especially when dealing with datasets that can be re-identified or fall into ambiguous regulatory categories. Federated data analysis offers a promising solution to some of these ethical and technical dilemmas. By enabling researchers to analyze data across multiple sites without requiring centralization or the direct sharing of raw data, federated approaches mitigate the risks of privacy breaches and unauthorized data transfers. This model supports compliance with varying national and institutional regulations and respects local control over data assets, both of which are critical in contexts where trust, transparency, and equity are ongoing concerns. Federated analysis offers a transformative alternative: instead of moving data to the analysis, analyses are sent to the data. Individual-level data never leaves the custodian's server; only non-disclosive summary statistics are returned to researchers. Incorporating federated models into research collaboration not only strengthens data protection but also addresses long-standing issues of data sovereignty and participant autonomy. By maintaining data within institutional boundaries while still enabling meaningful scientific inquiry, federated systems can promote both ethical integrity and research excellence throughout Africa.

DataSHIELD is a scientifically mature privacy-preserving federated analysis framework enacted through a series of open-source software and infrastructure [11,12]. The DataSHIELD framework comprises four

foundational concepts that apply to all DataSHIELD analytic functions, regardless of data type or whether analysis is applied to data at one location or federated across multiple locations:

1. Processing of individual-level data (IPD) is conducted at the data custodian site "*taking the analysis to the data*".
2. Individual-level data is never viewed by the analyst.
3. Individual-level data is never shared or moved within the analysis network – it is maintained under the oversight of the data controller.
4. Only outputs that are not directly disclosive leave the data controller's environment. These outputs can be passed to the analyst or can be used in iterative algorithms that combine non-disclosive components from multiple locations in the analysis network.

In practice, DataSHIELD adheres to privacy-by-design principles that are embedded within computing infrastructure, analytical functionality, automated output checking, and data privacy methods. It is designed to operate in accordance with best-practice data governance for the use of sensitive health data, as outlined by [13]. This approach enables real-time federated analysis, resulting in faster time-to-results through automation of both the analysis coordination and disclosure protection. Moreover, the high level of control, customisation, and oversight afforded to data controllers when deploying DataSHIELD allows it to be readily tailored to the heterogeneous data governance requirements commonly encountered in cross-country research networks.

Since its inception in 2009, DataSHIELD has been co-designed through engagement with a wide range of stakeholders, including cross-disciplinary researchers (spanning computer science, medicine, epidemiology, bioethics, social sciences, law, and related fields), data controllers, research participants, and members of the public [14–17]. As a result, DataSHIELD represents a socially aligned approach that is explicitly designed to comply with contemporary data protection legislation, including the General Data Protection Regulation (GDPR). In addition, it aligns with recently developed frameworks governing the secure access and use of sensitive research data, including the Five Safes Framework [18,19] and the adoption of the FAIR (Findable, Accessible, Interoperable, and Reusable) principles [20].

Unique to DataSHIELD, two federated analytical approaches are provided for data modelling. The first is a one-stage individual participant data (IPD) meta-analysis, also referred to as a full-likelihood approach or virtual pooling. In this approach, distributed algorithms are executed simultaneously across multiple data nodes, operating directly on individual-level data without any transfer or sharing of such data between sites [11]. This enables the generation of a global analysis that is mathematically equivalent to an analysis conducted on physically pooled data [21,22].

The second approach is a two-stage IPD meta-analysis, also known as study-level meta-analysis. Here, algorithms are applied independently at each data node to produce study-specific summary estimates, which are subsequently combined using conventional meta-analytic methods, such as random-effects meta-analysis.

In addition, functions developed within the DataSHIELD federated analytics framework incorporate a range of disclosure prevention strategies. These strategies are tailored to the specific statistical functionality, data characteristics, and analytical context in which they are applied. Methods include the use of automated statistical disclosure control [13] and the application of data privacy methodologies, such as differential privacy. Irrespective of the disclosure control mechanism employed, the configuration of thresholds and parameters remains under the sole authority of the data controller, reflecting the local data context and acceptable level of disclosure risk.

The technical maturity of DataSHIELD is evidenced by its active and open community [23], which is supported by a well-defined governance structure, shared infrastructure and dedicated training resources and events to support contributors and users. Since its initial development, DataSHIELD has expanded substantially beyond its foundational components and now comprises more than 20 community-developed packages within its broader ecosystem [23]. This includes extended statistical functionality, including advanced modelling, time-to-event analysis, Bayesian approaches, compositional data analysis, and machine and federated learning. Additional packages expand the application across different infrastructures and data types, including tabular, omics, image, and routine healthcare data.

The relevance and scientific maturity of DataSHIELD have been formally recognised in *The Goldacre Review*, an independent evaluation of the use of UK healthcare data for research commissioned by the UK Secretary of State for Health and Social Care. The review identifies DataSHIELD as a widely adopted open-source platform enabling federated analysis of health data [24]. It has been integrated with Trusted Research Environment (TRE) architectures, supporting the federated analysis of routinely collected healthcare data [12,25] and is now being deployed within the NHS England Secure Data Environment network, enabling federated cross-regional analyses of UK routine healthcare data. Comparable national initiatives are underway in Germany, where DataSHIELD is implemented as part of NFDI4Health, the National Research Data Infrastructure for Personal Health Data [26].

DataSHIELD has a long-standing application within networks of European longitudinal studies addressing healthy obesity, environmental health, exposome and omics research, diabetes and child development [27–30]. Alongside this, it supports federated analytics for significant pan-European academic-industry partnerships including IMI-RHAPSODY for precision therapy in diabetes and IMI-SOPHIA to change the future of obesity care [31], and pan-European COVID-19 research [32].

Unlike traditional data-sharing approaches, DataSHIELD allows statistical analysis to be performed directly on distributed datasets without requiring the movement or centralization of sensitive information, thereby preserving local data autonomy and ensuring compliance with stringent ethical and regulatory frameworks for personal data protection. Its embedded multi-layer disclosure control mechanisms, aligned with the *Five Safes* principles, ensure that all analytical outputs are rigorously checked to prevent any risk of re-identification or unintended data release [13]. This approach supports scalable, flexible, and trustworthy research collaboration, even across institutions with heterogeneous

infrastructures and governance models. This federated and privacy-preserving approach is particularly significant for the African research landscape, where health and biomedical data are often fragmented across countries and institutions, governed by heterogeneous legal frameworks, and managed within infrastructures that vary greatly in maturity and security. DataSHIELD supports a broad range of federated analytical capabilities, including big data [33] and multi-omics [34], enabling researchers to perform descriptive statistics, regression modelling, and survival analysis directly on distributed data while maintaining strict privacy protection. Beyond these core methods, the platform has expanded into advanced biomedical analytics, including privacy-preserving machine learning approaches such as multi-task learning through the dsMTL package, which allows simultaneous identification of shared and site-specific patterns in complex, high-dimensional datasets [35]. For multi-omics applications, extensions like dsOmics empower secure federated analysis of genomic, transcriptomic, and other molecular data across multiple cohorts without exposing individual-level information [34]. Additionally, interoperability with established clinical data standards is facilitated through dsOMOP, which enables seamless integration of the OMOP Common Data Model into federated workflows, enhancing harmonization and large-scale reuse of healthcare data [36]. Collectively, these tools position DataSHIELD as a mature and versatile analytical ecosystem capable of supporting collaborative research across diverse infrastructures while safeguarding data confidentiality and upholding regulatory and ethical requirements.

In summary, by enabling complex statistical analyses to be performed across multiple sites while keeping sensitive data securely behind institutional firewalls, DataSHIELD effectively mitigates the risks of re-identification and data misuse [13]. At the same time, it promotes equitable participation in multi-centre research by allowing institutions with diverse infrastructure capacities to contribute valuable data and insights without compromising local governance autonomy. Accordingly, DataSHIELD addresses critical ethical, technical, and regulatory challenges inherent in cross-border data sharing and provides a trusted pathway for African research institutions to engage in global health data collaborations while maintaining control over their data assets and protecting the rights of the populations they serve.

Methods

Ethics Statement

The project received ethical approval to conduct the research from Strathmore University (Reference No. SU-ISERC2367/24). Furthermore, administrative clearance from each Pathfinder's administration. All methods were performed in accordance with the relevant guidelines and regulations. Written informed consent for participation in the study was obtained voluntarily from the research participants. This was after understanding the study's purpose. We maintained participants' confidentiality by keeping identifying information (telephone numbers) under a key and lock, and by coding using privacy-enhancing methods. The data was analysed and published in aggregate form to avoid the identification

of individual participants. The data is currently stored under key and lock for the five years after publication.

Project Design and Scope

Data Science Without Borders (DSWB), inaugurated in February 2022, is a three-year initiative funded by Wellcome [37]. It aims to collaboratively design strategies to enhance data systems and the application of data science tools to improve data utilization for evidence generation in Africa. DSWB commenced its operations in collaboration with three distinct health institutions: the Armauer Hansen Research Institute (AHRI) in Ethiopia, the Douala General Hospital (DGH) in Cameroon, and the Institute for Health Research, Epidemiological Surveillance, and Training (IRESSEF) in Senegal. While all three countries had national data protection laws that govern data sharing, the data governance practices differed amongst the institutions, with the existence of institutional data protection frameworks at AHRI and IRESSEF, while DGH had no formal guidance to data sharing. The project team then developed a consolidated consortium-wide data sharing agreement with the guidance of legal representatives from all partners and signed by all consortium members. The agreement is currently implemented to allow ethical and responsible sharing of data. Data sharing was combined with an additional data requisition requirement to allow sharing of data specified for a single analysis project – and not blanket sharing of all data in the consortium. With these checks in order, the team explored the feasibility of implementing federated analysis as an alternative solution to overcome the hurdles of sharing in Africa.

Training and Capacity Building:

To initiate the process of deploying DataSHIELD in the DSWB consortium, we started off with a comprehensive training. We designed a two-block, hands-on curriculum that coupled privacy-preserving federated analytics in DataSHIELD with standardized data management in the OMOP Common Data Model (CDM), explicitly bridged by dsOMOP. Block 1 (DataSHIELD) introduced the client-server workflow, authenticated connections to remote nodes, assignment vs. aggregate operations, and the practical application of disclosure control, governance, and auditing. Block 2 (OMOP via dsOMOP) covered connecting to local OMOP CDM databases from within DataSHIELD, exploring standardized tables/vocabularies, constructing cohorts, and running cross-site summary analyses over harmonized clinical data. The workshops were delivered in a strongly practical format: participants connected to remote nodes and executed the exercises on live servers, mirroring production usage and reinforcing operational fluency. Throughout the training, we emphasised dsOMOP as the enabling layer that automates server-side extraction and transformation of OMOP tables into DataSHIELD objects so that only non-disclosive aggregates leave the three countries [36]. To ensure alignment with community standards, the course drew on materials and practices curated by the DataSHIELD Community Education Theme, which coordinates training materials across the community [38].

Training participants

The cohorts comprised multi-disciplinary teams essential for sustaining a federated network: data scientists/statisticians (analysis and quality control), IT and systems administrators (server deployment, security, networking), database managers/data officers (OMOP databases, ETL processes), and clinical/public-health researchers (protocol design and interpretation). In line with the workshop design, all participants were provisioned with access to their institutional nodes and completed the practical blocks on remotely hosted servers, not on local setups - an approach that our materials explicitly encourage. The setup was aimed to connect to remote nodes and manage sessions, perform federated, reproducible analyses on harmonized data.

The training effectiveness was assessed formatively through embedded practical check-offs rather than solely through written tests. By the end of the training, (i) researchers successfully executed federated descriptive and modelling tasks on their own institutional data using the OMOP -> dsOMOP -> DataSHIELD pipeline; (ii) IT staff deployed and operated a functional Opal/DataSHIELD stack (including user/role management and TLS-secured access); and (iii) data managers completed or validated source-to-OMOP mappings and exercised dsOMOP-based resource registration and filtered table assignment. These outcomes reflect the intended competencies of the two-block syllabus and the dsOMOP design - namely, standardized semantics from OMOP combined with server-side, disclosure-controlled computation in DataSHIELD.

Technical Implementation of the Federated Network:

Each participating site operated an Opal application server with a local R execution layer (Rock/Rserve) and linked it to the site's OMOP Common Data Model (CDM) database via dsOMOP R package. In this architecture, dsOMOP registers the OMOP database as an Opal "resource" and automates the server-side transformation of OMOP tables into DataSHIELD-compatible objects; analysts then run privacy-preserving R code from the client, and only non-disclosive aggregate results leave the site. This design preserves data locality (the OMOP database remains under institutional control) while avoiding duplication of large tables inside Opal and enabling harmonised multi-centre analyses. Opal's resources provide first-class connectors to external data/computation back ends so that data can be analysed in situ rather than imported into Opal's internal store. Under the hood, this mechanism relies on the resourcer R package, which interprets a resource description (URL, driver, credentials, format) and instantiates the appropriate connection (e.g., to files, SQL engines, or other services). dsOMOP leverages this resource layer to expose OMOP CDM databases to DataSHIELD sessions in a standardized way – crucial for scaling across institutions and keeping OMOP as the single source of truth [33]. Across nodes, Opal and the R servers ran on Linux (Ubuntu LTS) on server-class hosts or dedicated VMs with multi-core CPUs, sufficient RAM for concurrent R sessions, and SSD storage to minimise I/O latency. Opal connected to one or more R server profiles ("clusters") so administrators could pin package versions and scale workers consistently across sites; DataSHIELD profiles were derived from these R profiles to

enumerate permitted functions under disclosure control. The network specifications varied across centers; APHRC and IRESSEF hosted both the Opal and DB in AWS, while DGH deployed it on the local machine that met the relevant specifications. Opal's UI/API was published HTTPS-only with TLS certificates managed by the institution. Connectivity from Opal/R to the OMOP database was limited to site-controlled private routes -either an institutional LAN (on-prem) or private networking within a cloud VPC- so routine analysis traffic did not rely on public database endpoints. Role-based access governed analyst permissions, and administrative access followed institutional security policy.

Software and System Setup

We followed a standardized, automation-first installation procedure that minimizes site-to-site variability and maximizes reproducibility: (i) provision the Opal stack with the easy-opal CLI, (ii) register a Rock R-server profile that includes dsOMOP, (iii) declare an Opal resource pointing to that database so that DataSHIELD can query it in situ via the resourcer framework.

Step 1: Provision Opal with easy-opal. Administrators installed and configured Opal 4.x using the easy-opal CLI, which sets up the Opal container, reverse-proxy/TLS, baseline security, and service dependencies in one run (interactive or scripted). This provides a reproducible, cross-platform bootstrap aligned with Opal's administrator guidance, Fig. 2.

Step 2: Add a dsOMOP-enabled Rock profile with easy-opal. In order to expose a server-side R execution environment to Opal, we registered a Rock R-server profile using the easy-opal CLI and pointed it to a dsOMOP-enabled Rock image. This yielded a named R profile/cluster that Opal discovers and that analysts can target consistently across nodes.

Step 3: Declare the site's OMOP database as an Opal resource and test access. Within each Opal project, we created a resource that points to the site's OMOP CDM endpoint -- supplying the standard parameters (driver, host, port, db, credentials). Opal's resource mechanism (backed by the resourcer package) allows DataSHIELD to operate in situ on the database; dsOMOP then automates the server-side extraction/transformations into DataSHIELD objects. This pattern works identically whether the database is in a cloud VPC or on-prem and is exactly the workflow described in the resources/resourcer architecture and the dsOMOP reference.

Data Preparation and Harmonization

The team conducted a comprehensive data mapping of all available datasets in the consortium and identified 37 datasets. Of these, three were cross-cutting and these included datasets on HIV, TB, and Malaria [37]. To support DataSHIELD deployment at all three sites, we selected one use case dataset from each site. At APHRC, we used datasets specific to health diseases, at DGH, a data set on stroke was used, while IRESSEF used a data set specific to COVID-19. The different datasets were harmonized to the OMOP Common Data Model (CDM) using a conventional OHDSI ETL pipeline and tools. The deployment

using different datasets justified the need for harmonized datasets and the goal was to achieve both semantic interoperability (shared vocabularies and table conventions) and analytic interoperability (uniform, privacy-preserving queries across nodes), if one dataset was to be analysed using DataSHIELD.

Our Experience

Overall, we successfully deployed DataSHIELD to address challenges related to data sharing within the DSWB consortium and, more broadly, across African research settings. We deployed DataSHIELD on both AWS and local servers or computers of the participating institutions. This flexibility enabled the three sites, each at a different level of data ecosystem maturity, to operate within their existing infrastructures. The installation was tested using the site-specific datasets selected from the consortium, allowing the technical teams to experience and understand the critical steps involved in data access, assignment of administrative rights, and procedures for sharing aggregated results once federated analyses are deployed. Demand for federated analysis has since increased across partner sites, and the first federated analysis use case will involve cross-cutting HIV datasets from the three partner institutions.

Training and Skill Development Outcomes

Before the training, all participating data personnel were subjected to a pre-training survey to ascertain knowledge and key skills. We observed the majority of participants in need for skills in use of data harmonization tools such as the OMOP Common Data Model (53.8%), use of Docker and containerization (84.6%), and use of Linux Operating System for the installation of DataSHIELD (53.8%), Fig. 3.

This underscores the need for training before and during deployment of federated tools and the usability of joining Community groups – including the DataSHIELD community. The majority of the participants had no prior knowledge and experience using privacy enhancing technology and DataSHIELD, and showed interest in learning new skills, specific to federation, Fig. 4.

In response to the identified capacity gaps, and to promote sustainability and reuse in line with Open Science and FAIR principles, we developed and published an openly accessible, web-based training course that consolidates the DataSHIELD and OMOP-based federated analysis curriculum and supports continued, reusable skills development (https://isglobal-brge.github.io/workshop_DSWB/). The training programme was delivered in complementary blocks addressing analytical, data management, and infrastructure competencies required to sustain a federated analysis network. As summarized in Table 2, Block 1 focused on DataSHIELD concepts, including the client–server workflow, disclosure control, and governance, enabling data scientists and IT staff to successfully execute federated analyses that returned only non-disclosive outputs. Block 2 emphasized data harmonization through the OMOP Common Data Model and dsOMOP integration, equipping data managers and analysts to construct cohorts and query harmonized OMOP data via DataSHIELD. In parallel, infrastructure-focused sessions

supported IT administrators in deploying and maintaining secure OPAL and Rock environments with TLS and role-based access controls.

Table 1
Training Block Core Topics Participants Demonstrated Outcomes

Block	Topic	Participants	Outcome
Block 1: DataSHIELD	Client–server workflow, disclosure control, statistical analyses	Data scientists, IT staff	Executed federated analyses with non-disclosive outputs
Block 2: OMOP & dsOMOP	OMOP CDM, cohort construction, dsOMOP integration	Data managers, analysts	Successfully queried harmonized OMOP data via DataSHIELD
Infrastructure	Opal, Rock/R, TLS, role-based access, governance	IT administrators	Deployed and maintained secure federated nodes

Challenges and Solutions

The deployment of a federated DataSHIELD infrastructure across multiple African research institutions presented several technical and operational challenges related to connectivity, system configuration, and software heterogeneity (Table 2). Network latency and unstable internet connectivity were common across sites and posed potential barriers to interactive analysis. These challenges were mitigated by leveraging DataSHIELD's server-side, in-situ computation model, which limits data transfer to non-disclosive aggregate outputs, thereby reducing sensitivity to network latency. In addition, lightweight analytical queries and persistent client–server sessions were adopted to tolerate intermittent connectivity and minimize disruption during analysis workflows.

Server configuration and institutional security requirements also represented a key challenge, particularly with respect to firewall restrictions and approved network ports. To address this, deployments were conducted in close collaboration with institutional IT officers, and all installations adhered strictly to existing security protocols to ensure that DataSHIELD implementation did not compromise institutional infrastructure or data security.

Software compatibility issues arose from the use of heterogeneous versions of R and Python across participating sites. This was addressed by defining and enforcing standardized software versions prior to training and deployment. To further reduce variability across local environments, Docker containerization was employed to encapsulate system dependencies and enable reproducible, localized deployments of OPAL, Rock, and dsOMOP services.

Table 2
Technical and Logistical Challenges and Solutions during DataSHIELD Deployment under the DSWB Consortium

Challenge Category	Specific Challenge Encountered	Solution Implemented
Network/Connectivity	Latency issues	Server-side DataSHIELD computation limited data transfer to aggregate results, reducing sensitivity to network latency.
	Unstable internet	Lightweight queries and persistent sessions were used to tolerate intermittent connectivity.
Server Configuration	Firewall and security issues	Referred to the institution IT officer and followed the security protocols to avoid comprising institutional safety during DataSHIELD deployment
Software compatibility	Different versions of R software at the sites	Ensured participants installed standardized software prior to the training
		Used Docker containers to allow localized deployment

Reflections and Discussion

The deployment of DataSHIELD through the DSWB consortium represents a landmark achievement for data science in Africa. Across the continent, implementations of open-source software for federated analysis and learning remain rare. According to recent updates from the DataSHIELD community, only one or two use cases have been documented in Africa, underscoring a substantial gap in methodological advances and comparative work between Africa and the rest of the global community. For multi-country data-driven projects where data sharing and trust are critical for successful collaboration, the feasibility of data federation is essential. This need is particularly pronounced in Africa, where weak policy and legal frameworks, the absence of a continental data-sharing framework, limited institutional data policies, and poor data practices collectively hinder trust in data sharing [5,6]. Consequently, demand is growing for solutions that allow data to remain at the point of collection while still enabling robust federated analyses.

Broader Implications for Research

The DataSHIELD framework offers substantial opportunities to advance health research in Africa. This privacy-preserving analytical platform can be strategically applied to infectious disease surveillance, non-communicable disease research, and other emerging public health challenges. For infectious diseases, DataSHIELD enables near real-time, multi-country analyses of outbreaks without centralizing sensitive patient data, facilitating rapid epidemic response while maintaining data sovereignty. For non-communicable diseases such as diabetes, hypertension, cardiovascular disease, and cancer, the

framework supports harmonization of heterogeneous datasets from different health systems, enabling large-scale comparative studies to identify risk factors, treatment outcomes, and health disparities across diverse African populations. A continental data network underpinned by DataSHIELD principles could address key barriers to pan-African health research, including data protection concerns, limited technical infrastructure, and the need for local control of sensitive health information. Such a network would align with the African Union's health agenda, strengthen health systems research capacity, support evidence-based policy, and position African institutions as equal partners in global health research. By fostering a federated data ecosystem that respects national sovereignty while enabling large-scale analyses, DataSHIELD can support rigorous, ethically robust collaborative research and contribute to improved health outcomes across Africa.

Comparison of DataSHIELD to Other Models of Federated Analysis

While several federated data analysis frameworks exist, DataSHIELD and Vantage6 exemplify two distinct approaches to privacy-preserving collaborative research. DataSHIELD, implemented in R, uses a disclosure-control framework that limits analyses to non-disclosive functions, ensuring that individual-level data remain at their source while still supporting advanced methods such as generalized linear models, survival analysis, and machine learning. This provides strong protections against data disclosure and integrates well with established R-based workflows, making it particularly suitable for epidemiologists and biostatisticians. In contrast, Vantage6 offers a more flexible, containerized infrastructure that supports multiple programming languages (e.g., Python, R) and enables deployment of custom algorithms via Docker containers, offering greater algorithmic freedom and adaptability. Both retain data within institutional firewalls but differ in governance: DataSHIELD's function-based restrictions embed privacy by design, whereas Vantage6's container-based model requires explicit review and approval of analytic algorithms by data custodians. In the African context, DataSHIELD's lower technical barrier, proven use in multi-country health research, and alignment with existing R-based capacity make it highly attractive, while Vantage6 may be preferable for projects needing specialized computation or Python-based machine learning. Both frameworks support the overarching goal of enabling collaborative research without compromising data sovereignty, a central concern for African institutions seeking equitable participation in global health research while retaining control over sensitive population data.

Conclusion

The successful deployment of DataSHIELD under the Data Science Without Borders in Africa project demonstrates the feasibility of robust, real-world federated analysis in low- and middle-income settings. By demonstrating both a clear demand for and a practical pathway to the secure deployment of DataSHIELD and other federated tools, this work shows that there is potential for collaborative data analysis with a common research question to support pooled meta-analysis, which can be achieved

while fully respecting data sovereignty and privacy. This initiative has laid a crucial foundation for a new era of data-driven research and evidence-informed policy-making on the African continent, and it highlights the substantial potential of federated approaches to strengthen health research, foster regional collaboration, and ultimately improve population outcomes.

Declarations

Competing Interests

The authors declare no competing interests.

Funding

The study was funded through a grant from the Wellcome, Grant number 228139, awarded to Agnes Kiragga at the African Population and Health Research Center, Kenya.

Author Contribution

AK and JG designed and conceptualized the study. AK, JG and DS wrote the original draft of the manuscript. Visualizations were done by MO, SC and PA. RW, JG and DS produced the Software. All authors reviewed and edited the manuscript and approved the final version for submission.

Acknowledgement

We acknowledge support from the Barcelona Institute of Global Health and the London School of Hygiene and Tropical Medicine for the support towards training and deployment of DataSHIELD.

Data Availability

The datasets used in this study are available from the corresponding author upon request.

References

1. Naghavi, M. et al. Global burden of 292 causes of death in 204 countries and territories and 660 subnational locations, 1990–2023: a systematic analysis for the Global Burden of Disease Study 2023. *Lancet* **406**, 1811–1872 (2025).
2. Ndembi, N. et al. Evolving Epidemiology of Mpox in Africa in 2024. *N Engl. J. Med.* **392**, 666–676 (2025).

3. Brand, D., Singh, J. A., McKay, A. G. N., Cengiz, N. & Moodley, K. Data sharing governance in sub-Saharan Africa during public health emergencies: Gaps and guidance. *South Afr. J. Sci* **118**, (2023).
4. Badriyya, Y. *Harmonization of Data Governance Frameworks in Africa*. (2024).
5. Kabanda, S. M. et al. Data sharing and data governance in sub-Saharan Africa: Perspectives from researchers and scientists engaged in data-intensive research. *South Afr. J. Sci* **119**, (2023).
6. ICT Works. 14 Barriers to Using Open Data for Better Development Decisions. *14 Barriers to Using Open Data for Better Development Decisions* (2024). https://www.ictworks.org/open-data-development-decisions/
7. European Union Parliament. REGULATION (EU) 2022/868 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act). (2022).
8. Orji, U. J. The African Union Convention on Cybersecurity: A Regional Response Towards Cyber Stability? *Masaryk Univ. J. Law Technol.* **12**, 91–130 (2018).
9. Collaborators Observational Health Data Sciences and Informatics (OHDSI). (2026).
10. Human-Centered Artificial Intelligence. *Artificial Intelligence Index Report 2024*. (2024). https://hai-production.s3.amazonaws.com/files/hai_ai-index-report-2024-smaller2.pdf
11. Gaye, A. et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int. J. Epidemiol.* **43**, 1929–1944 (2014).
12. Wilson, R. C. et al. DataSHIELD – New Directions and Dimensions. *Data Sci. J.* **16**, 21 (2017).
13. Avraam, D. et al. DataSHIELD: mitigating disclosure risk in a multi-site federated analysis platform. *Bioinforma Adv.* **5**, vbaf046 (2025).
14. Budin-Ljøsne, I. et al. DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis. *Public. Health Genomics*. **18**, 87–96 (2015).
15. Murtagh, M. J. et al. Securing the Data Economy: Translating Privacy and Enacting Security in the Development of DataSHIELD. *Public. Health Genomics*. **15**, 243–253 (2012).
16. Murtagh, M. J. et al. The ECOUTER methodology for stakeholder engagement in translational research. *BMC Med. Ethics*. **18**, 24 (2017).
17. Wallace, S. E., Gaye, A., Shoush, O. & Burton, P. R. Protecting Personal Data in Epidemiological Research: DataSHIELD and UK Law. *Public. Health Genomics*. **17**, 149–157 (2014).
18. Ritchie, F. Five Safes: designing data access for research. Preprint at https://doi.org/ (2016). 10.13140/RG.2.1.3661.1604
19. Standard Architecture for Trusted Research Environments. The SATRE Specification. (2026). https://satre-specification.readthedocs.io/en/stable/specification.html
20. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*. **3**, 160018 (2016).
21. Jones, E. M. et al. DataSHIELD – shared individual-level analysis without sharing the data: a biostatistical perspective. *Nor Epidemiol* **21**, (2012).

22. Jones, E. M., Sheehan, N. A., Gaye, A., Laflamme, P. & Burton, P. Combined analysis of correlated data when data cannot be pooled. *Stat* **2**, 72–85 (2013).
23. Wikipedia The DataSHIELD Community wiki. (2026). https://wiki.datashield.org/en/home
24. Goldacre, B. & Morley, J. *Better, Broader, Safer: Using Health Data for Research and Analysis*. 130 (2022). https://assets.publishing.service.gov.uk/media/624ea0ade90e072a014d508a/goldacre-review-using-health-data-for-research-and-analysis.pdf
25. Butters, O. W. et al. The Biomedical Research Infrastructure Software as a Service Kit (BRISSKit): technical description. *F1000Research* 5, 1905 (2016).
26. NFDI4Health. *NFDI4Health* (2026). https://www.nfdi4health.de/en/
27. Van Vliet-Ostaptchouk, J. V. et al. The prevalence of metabolic syndrome and metabolically healthy obesity in Europe: a collaborative analysis of ten large cohort studies. *BMC Endocr. Disord.* **14**, 9 (2014).
28. Vinther, J. L. et al. Gestational age at birth and body size from infancy through adolescence: An individual participant data meta-analysis on 253,810 singletons in 16 birth cohort studies. *PLOS Med.* **20**, e1004036 (2023).
29. Vrijheid, M. et al. Advancing tools for human early lifecourse exposome research and translation (ATHLETE): Project overview. *Environ. Epidemiol.* **5**, e166 (2021).
30. Zijlema, W. et al. Road traffic noise, blood pressure and heart rate: Pooled analyses of harmonized data from 88,336 participants. *Environ. Res.* **151**, 804–813 (2016).
31. Delfin, C. et al. A Federated Database for Obesity Research: An IMI-SOPHIA Study. *Life* **14**, 262 (2024).
32. Puskaric, M. et al. Privacy-Preserving Workflow for the Cross-Border Federated Analysis of Clinical Data. in Studies in Health Technology and Informatics (ed (ed Mantas, J.) (IOS, doi:10.3233/SHTI240737. (2024).
33. Marcon, Y. et al. Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD. *PLoS Comput. Biol.* **17**, e1008880 (2021).
34. Escriba-Montagut, X. et al. Federated privacy-protected meta- and mega-omics data analysis in multi-center studies with a fully open-source analytic platform. *PLoS Comput. Biol.* **20**, e1012626 (2024).
35. Cao, H. et al. dsMTL: a computational framework for privacy-preserving, distributed multi-task machine learning. *Bioinformatics* **38**, 4919–4926 (2022).
36. Sarrat-González, D., Escribà-Montagut, X., Houghtaling, J. & González, J. R. dsOMOP: bridging OMOP CDM and DataSHIELD for secure federated analysis of standardized clinical data. *Bioinformatics* **41**, btaf286 (2025).
37. Kiragga, A. N. et al. Data science without borders: bridging the divide in data science capacity across African health institutions. *Front. Public. Health.* **13**, 1695907 (2025).

38. Gonzalez, J. & Sarrat-González, D. *DSWB Training Materials*. (2025). https://isglobal-brge.github.io/workshop_DSWB/

Figures

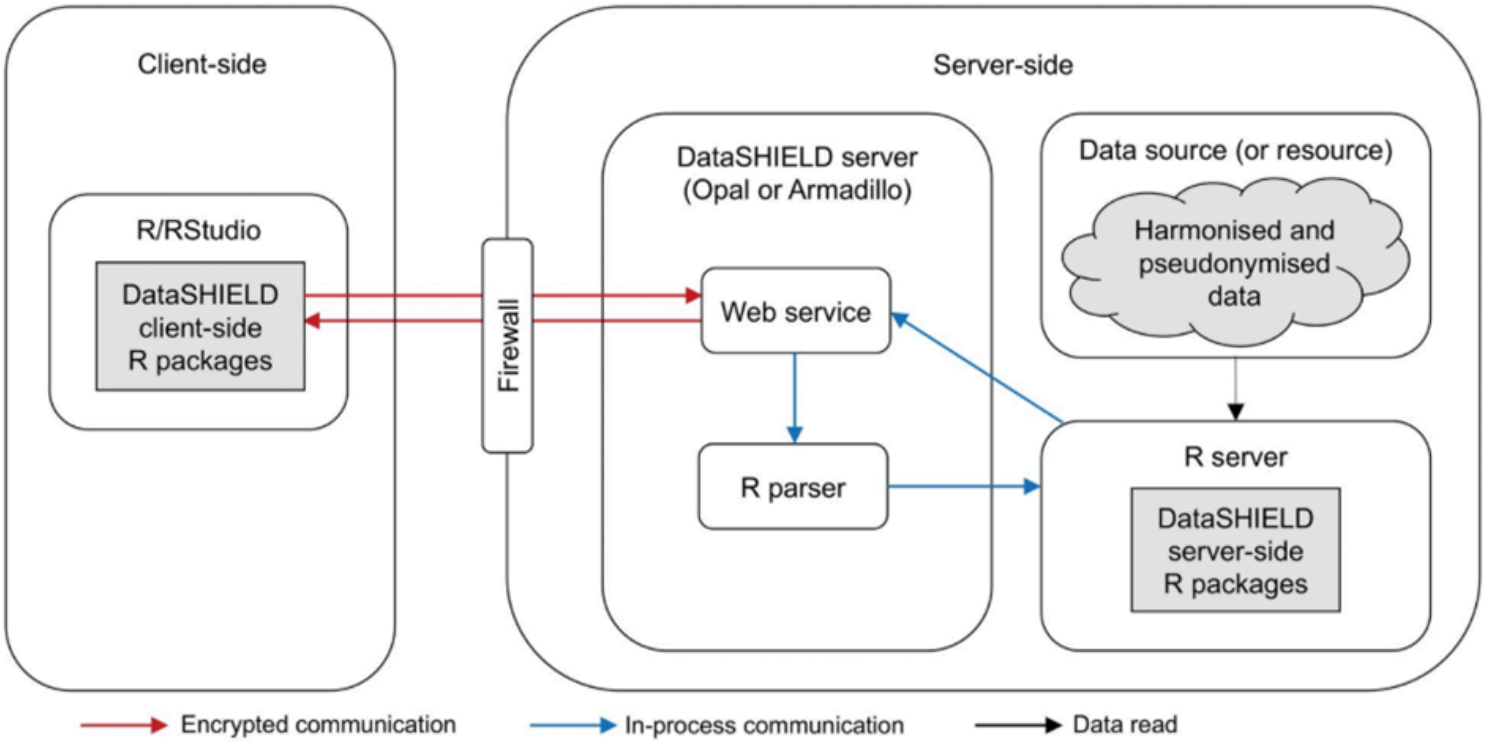

Figure 1

Overview of DataSHIELD computing architecture (source [13])

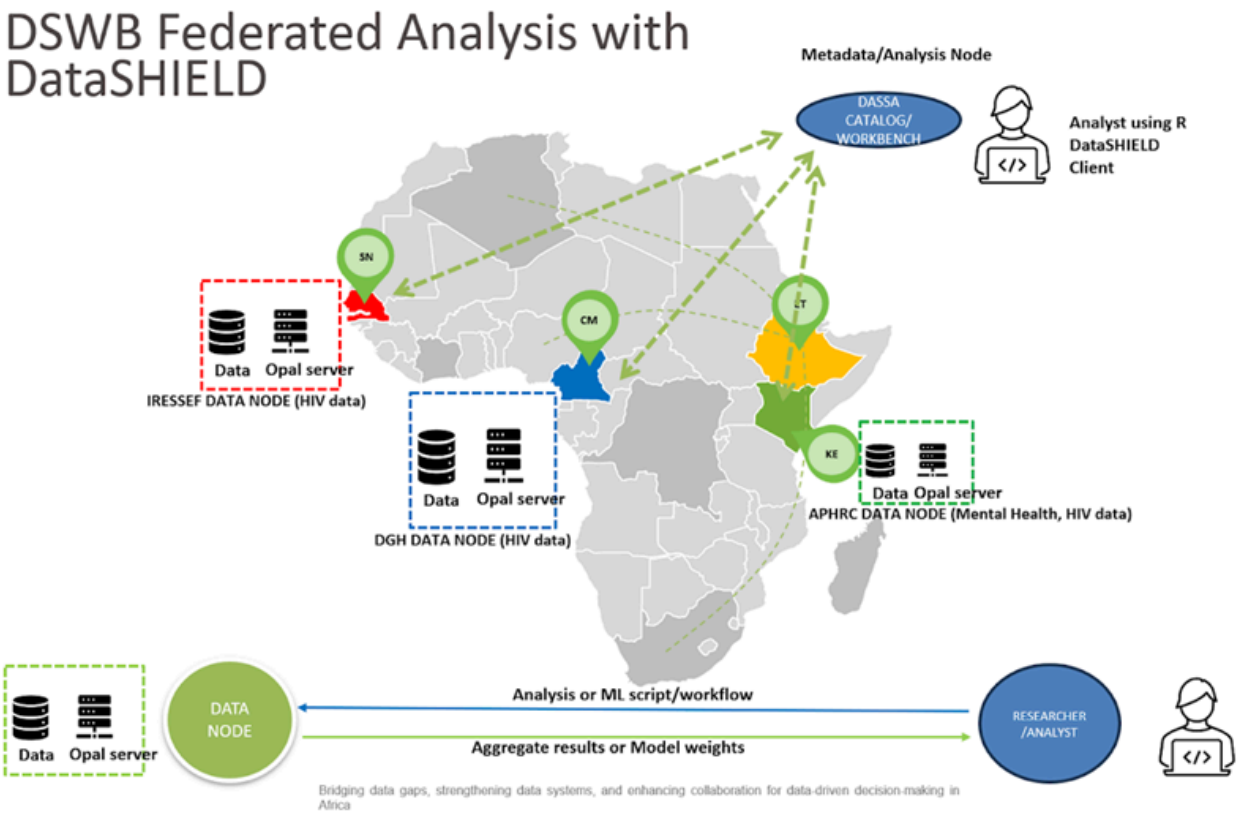

Figure 2

Deployment of DataSHIELD under the Data Science Without Borders Consortium

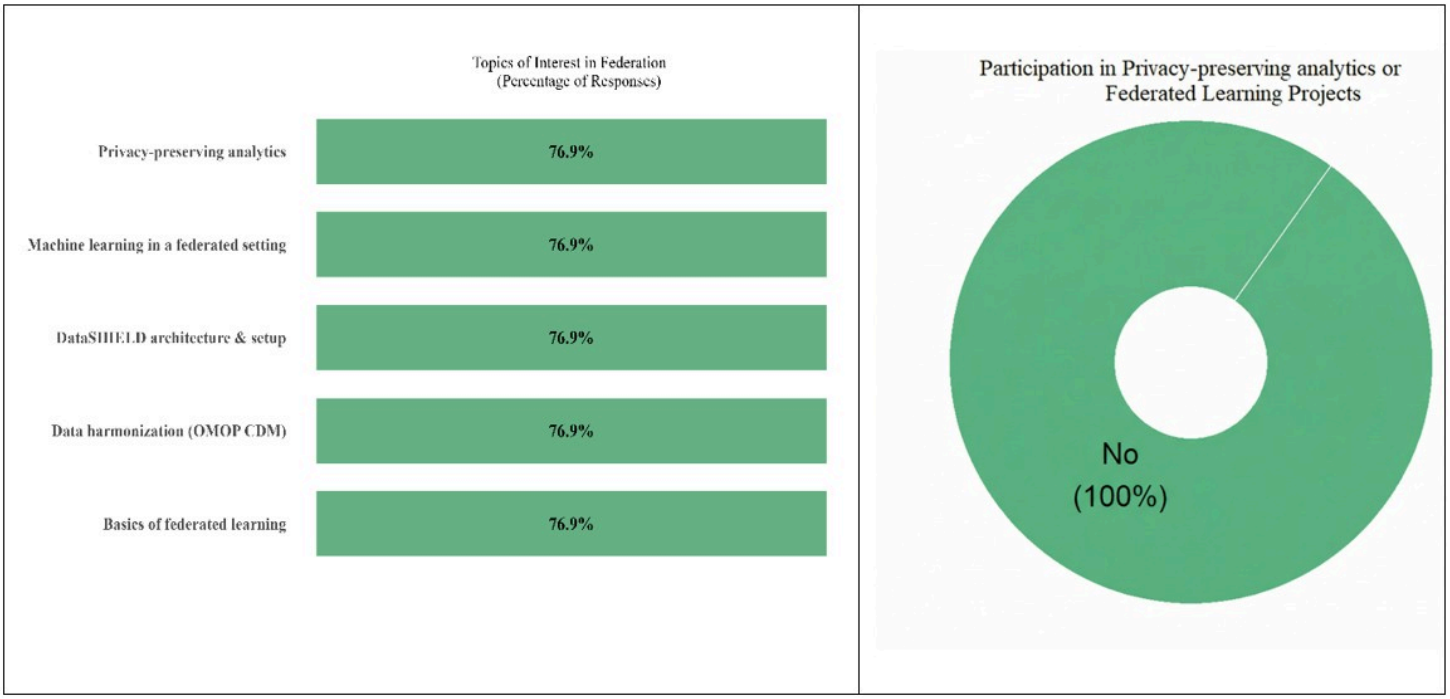

Figure 3

Proficiency in technical skills among DSWB consortium members in Cameroon, Senegal and Kenya

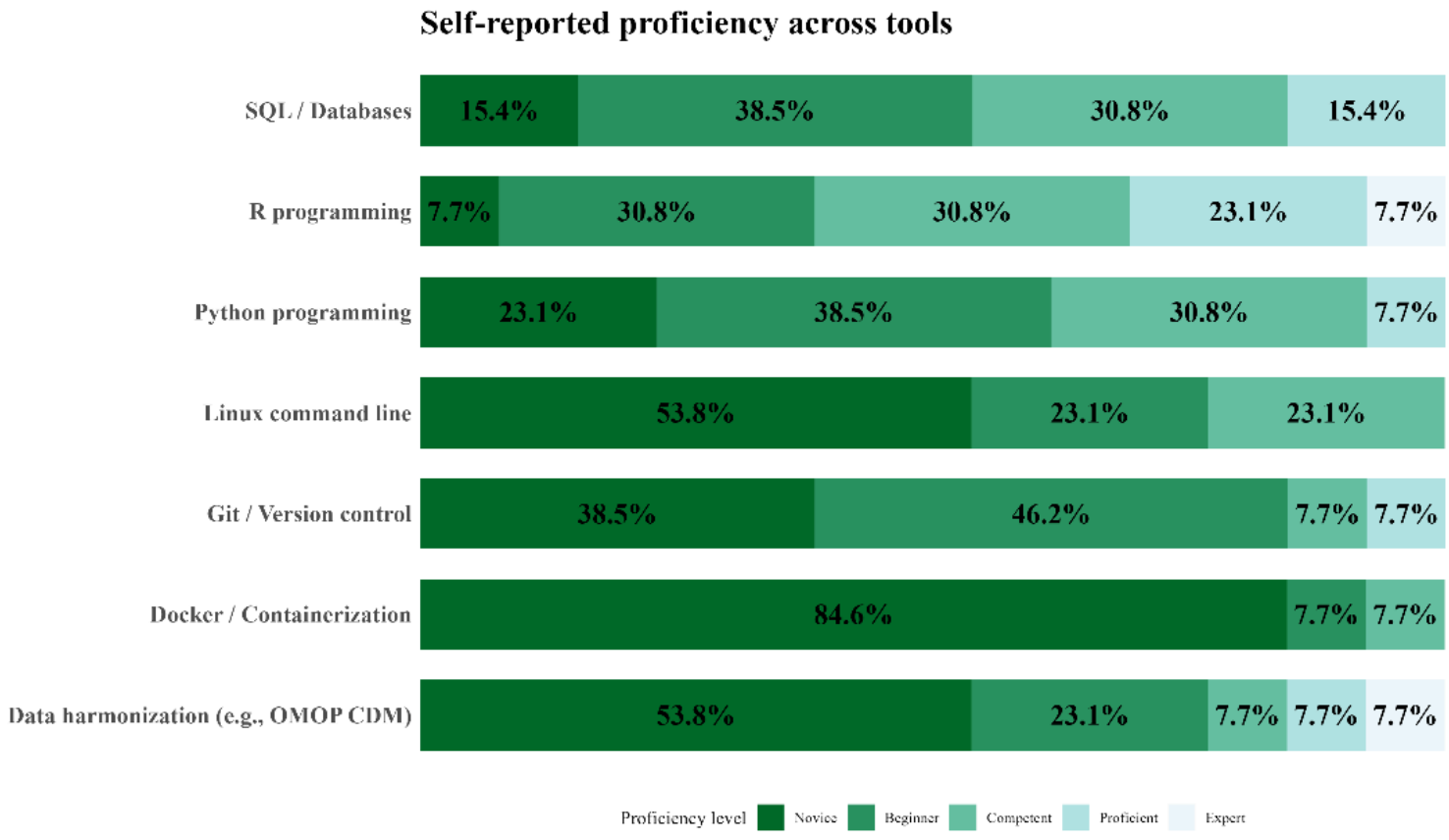

Figure 4

Most desired training for effective use of DataSHIELD among data personnel in the DSWB consortium